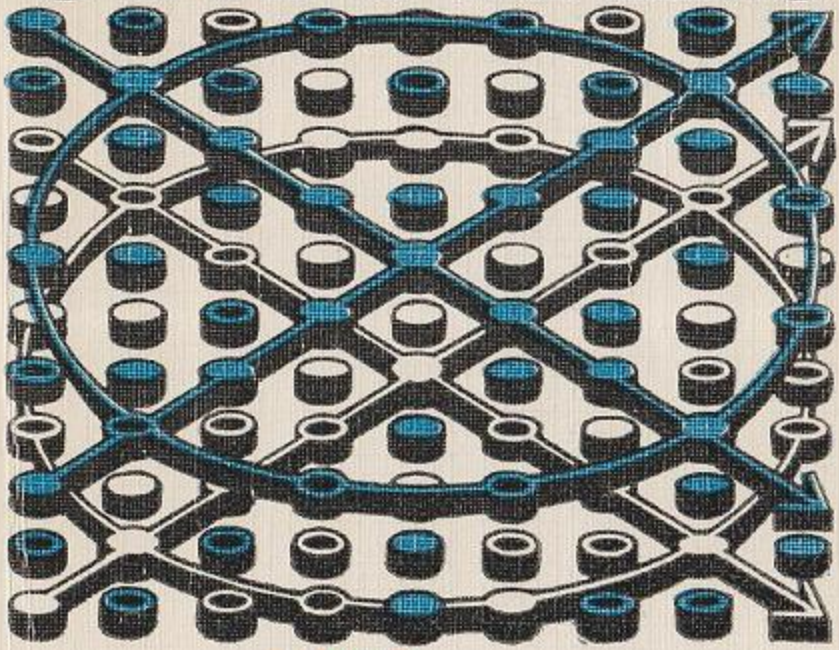
SCIENCE FOR EVERYONE

Th. WOLKENSTEIN

ELECTRONS AND CRYSTALS



MIR

SCIENCE FOR EVERYONE

The increasingly important field of solid-state physics concerns the behaviour of electrons in various crystals. Problems of solid-state physics, which include specific differences between metals and dielectrics and the remarkable properties of semiconductors, are particularly topical in today's 'electronic' society.

Electrons and Crystals by Dr. Theodore Wolkenstein covers the fundamentals of solid-state physics in an engaging way. Written in an easy, readable style, the book is intended as a supplement to textbooks in secondary-school physics courses, and the approach to certain topics in the volume is, therefore, unique. The material is presented in terms of models and requires no special additional knowledge.

Suitable for the general reader with a good command of elementary physics and mathematics, this book can also serve as a useful study guide for high-school students.

ИЗДАТЕЛЬСТВО МИР

Ф. Ф. Волькенштейн

Электроны и кристаллы

Издательство «Наука», Москва

Th. Wolkenstein

Electrons and Crystals

Translated from the Russian
by Michael Burov

English Translation Edited
by R. N. Hainsworth

Mir
Publishers
Moscow

First published 1985
Revised from the 1983 Russian edition

На английском языке

Instead of Preface: On the Laws of Popular Science

Any scientist or scholar knows that writing a scientific monograph is much easier than writing a popular science book on the same topic. It is easier to present a paper in front of a group of colleagues than to give a popular science lecture to the public.

This is because scientists speak their own language, in fact each branch of knowledge has its own. Hence even experts in other fields of science will not understand them, the more so people without a scientific education.

A popular science writer must translate from the scientific language into everyday language. The popularization of science like the translation of verse into another language is an art in itself requiring special skills.

I would like to propose some basic rules (a 'legal code') any popular science writer should adhere to.

First, a popular science writer has to be absolutely accurate in his presentation.

The second rule is that each new term or concept should be explained immediately, otherwise the text will become unclear for the reader, and the most terrible thing for a writer will happen: the reader will put his book aside.

An author (be he a professional writer or a scientist) should always 'see' his reader as no one can write into a void. The reader is always in front of him, and an author should be aware

of what the reader knows and what he does not. This defines the language the author uses to communicate with the reader, and will define the boundary between what is clear and what is not.

The whole book (except perhaps certain points) has to be written at the same level. This is the third (quite obvious) rule of popular science writing.

Many science writers think it necessary to insert some fictional digressions. Metaphors, analogies, associations are warranted and even desirable, but fictional excursions that are not related to the subject-matter do not make a text clear, they only dilute it, like a glass of water poured into a bowl of soup.

There is no harm in repeating the material several times in different parts of the book. Of course, if the light is thrown differently on the material, like theatre floodlights can colour the same stage differently, it is more deeply ingrained on the mind of the reader.

When a reader opens a book, he puts his hand in the hand of the writer, and the writer leads him through the thick forest called 'science'. So as not to lose his way, the reader always has to remember, and the author always has to remind him, where he is going and the destination of his journey. In other words, a reader has to keep before him the plan of the book, like a tourist keeps a map of his route in his hands. When he starts reading a book, he begins with its 'Contents' which provides information about the topics and the structure of the book. The author should lead his readers to their destination by

the shortest route possible and without zigzags.

Let us mention one more rule that is possibly the most essential in this 'legal code' for popular science. The book has to be interesting. What is 'interesting' and 'uninteresting' is subjective. Everything depends on the way you look at things. A writer may make 'interesting' things seem boring, or 'uninteresting' things fascinating. The art of the popular science writer is to turn the 'uninteresting' into the 'interesting'.

The aim is not so much to supply the reader with knowledge about a subject, but to kindle his interest in it. If a reader, having read the book, reaches for another book on the same or a similar topic, then the popular science writer has fulfilled his task.

This book tackles some of the problems of solid-state physics. We are going to discuss the behaviour of electrons in metals, semiconductors, and dielectrics (insulators), and some of the properties of solids affected by this behaviour. We do not pretend this is an exhaustive review of the latest achievements in solid-state physics, but we hope we have described some of the fundamental concepts in the physics of metals, semiconductors, and dielectrics. The book is a supplement to appropriate parts of a physics course in secondary schools and is intended for someone who is finishing secondary school or who has finished it and remembers his school physics.

We do not expect a reader to know anything beyond secondary school physics. Moreover, sometimes a topic is tackled somewhat differently than it might be treated in school. The mathematics used in the book does not go beyond ele-

mentary algebra and the mathematical analysis that should be familiar to anyone finishing secondary school.

We have attempted to present the material in terms of models and have not departed, as far as is possible, from classical concepts. However, modern solid-state physics is based on quantum mechanics, which is what governs the world of atoms and electrons. This makes it hard for us since we cannot count on the reader knowing very much about quantum mechanics. Therefore sometimes we have to be content with presenting the results of solid-state theory (taking care to make these results understandable) without showing how they were obtained. The reader will just have to believe them, but this is, alas, unavoidable.

We do not undertake to judge how far this book meets our popular science 'code'. At any rate, it was a guide to us when we wrote this book.

A well-known French physicist, and one of the founders of quantum mechanics, Louis de Broglie, once said that 'science is the daughter of astonishment and curiosity'. The mission of a popular science writer is to reawaken in the reader his feelings of astonishment and curiosity.

Contents

Introduction

1.1. Which Electrons?

Any matter, as is generally known, can occur as a solid, a liquid, or a gas. Modern physics has added another to these three states, which may be referred to as the 'classical' states, and that is the plasma state. A material is a plazma when all its atoms are ionized, i.e. all the electron shells have been partially or completely stripped away from the atoms. We are not going to discuss gases, liquids or plasma. We shall focus entirely on solids.

Modern physics has two branches which are the physics of atomic nuclei and elementary particles on the one hand and solid-state physics on the other. At present these two disciplines are independent and almost do not meet. Sometimes a physicist studying atomic nuclei knows very little about solid-state physics, and vice versa.

Speaking of solids, we have to distinguish between two kinds, i.e. *crystalline* and *amorphous* solids. An example of a crystalline solid is the common (or rock) salt we see every day at the dinner table. A typical example of an amorphous solid is glass. Perhaps one of the basic differences between an amorphous solid and a crystalline one is the absence of a distinct melting point in amorphous solids. Instead, there is a *softening range* in which the material gradually turns

from solid to liquid. We are not going to discuss amorphous solids in this book and shall only consider crystalline solids.

Crystalline solids are much more common in nature than it seems at first sight. A crystalline solid may have a regular geometric shape with some symmetry. Such materials are called *monocrystals*. At the same time there are what we call *polycrystals* which are aggregates of vast numbers of tiny monocrystals sticking together. A grain of common salt is an example of a monocrystal, while a piece of copper wire is an example of a polycrystalline solid. As is often true, the tiny monocrystals of copper cannot be seen with the naked eye.

In physics, the word 'crystal' has a wider meaning than we usually give it. Commonly, what we call 'crystals' a physicist calls 'monocrystals'. This book will basically deal with monocrystals, but when we cover polycrystals, we shall see that all the properties of monocrystals are retained. However, some new properties appear that are caused by the joints between separate tiny monocrystals. The joints can, for instance, increase the resistance to an electric current flowing through the polycrystal.

From now on we shall use the word 'crystal' instead of 'monocrystal'.

Any crystal consists of individual particles (they may all be the same or there may be several different kinds) arranged in a regularly repeating pattern. A crystal is like a honeycomb of cells in an indefinite sequence. Each of these *unit* (or *structural*) *cells* has the same configuration of the particles the material consists of. The assembly

of unit cells is what is called a *crystal lattice*. Each unit cell has *faces* which intersect to form *edges*. The intersections of the edges are called lattice *points*. Consequently, a crystal is like a house built of uniform bricks. This strict spatial regularity of a crystal's structure is its characteristic feature, and can moreover be considered to be a definition of a crystal. A crystal is a symbol of order, while its opposite, a symbol of chaos, is a gas for its particles rush about, run into each other, and change direction after each collision.

Crystals are classified as *molecular*, *atomic*, or *ionic* ones, according to the particles comprising them. In molecular crystals the structural elements of the crystal are individual molecules, each of which is a group of closely located atoms that are strongly bound to each other and define an entity. Atomic crystals are composed of atoms, each of which consists of a positively charged nucleus containing almost all the mass of the atom, and a collection of electrons whose total negative charge matches the positive charge of the nucleus. Atomic crystals may be made up of either one kind of atoms or atoms of several different kinds each possessing a place of its own in the crystal lattice. Ionic crystals are composed of atoms that may have either too many or too few electrons, i.e. atoms from whose electron shells one or more electrons have been withdrawn or, vice versa, into whose electron shells one or more electrons have been introduced (as everyone knows electrically charged atoms are called ions). An ionic crystal must clearly contain at least two kinds of oppositely charged ions.

Further on we shall only discuss atomic and ionic crystals. Molecular crystals have much in common with them, but at the same time they have some features of their own that we shall not touch upon.

The atoms or ions are located at particular points in the crystal lattice and oscillate around the lattice points. The amplitude of the oscillations increases with temperature.

Consequently, a crystal contains electrons (a few from each lattice point) that are initially bound to their atoms or ions, but which can be fairly easily separated from them and start travelling around the crystal. This concerns mainly the electrons belonging to the outer electron shells. Sometimes the electrons are released easier when the crystal is affected in some way, for instance, heated or irradiated with light of a certain wavelength (i.e. light of a certain colour). The behaviour of these electrons is governed by certain laws, and their behaviour, in its turn, gives rise to many properties of the crystal. For instance, the electrical and thermal conductivity, heat capacity, optical absorption, and many others which we are going to discuss below are all properties dependent on the free electrons. The electrons behave differently in metals, semiconductors, and dielectrics (we shall show what these terms mean exactly in the next section). It is more accurate to say that a crystal is a metal, a semiconductor, or a dielectric depending on the behaviour of its electrons.

The electrons populating a crystal are the main dramatis personae in this book. We shall also deal with the behaviour of these electrons

and discuss how the properties of crystals are determined by their behaviour.

I.2. Which Crystals?

One of the basic properties of a crystal is its electrical conduction, i.e. its ability to carry an electric current. Some crystals have very high conduction rate and some crystals have negligible conductivities.

It is a very important fact that in nature there are solids with high conductivities (conductors) and solids with conductivities that are practically nil (insulators). Our whole civilization is actually based on this fact. Indeed, every electrical device is a combination of conductors and insulators. What would happen if one fine day every insulator became a conductor? A catastrophe would happen: every electric lamp would go out, every trolley and motor car would stop, and telephones would not ring. Electric current would go into the earth, which is a bottomless pit for electric charge. Now let us imagine that, conversely, every conductor became an insulator. The current would 'die' in the wires. Electric charge would freeze, having lost its ability to run along a wire. The same catastrophe would occur and the civilized world would be plunged into darkness and become paralyzed.

How large is the conductivity of a conductor (a typical example being a metal), and how large is the conductivity of an insulator (a dielectric; an example is a crystal of rock salt)? How many times is the conductivity of an insulator less than the conductivity of a conductor? In order

to answer these questions we should remember the units conductivity is measured in.

Let us consider Ohm's law. Suppose we have a piece of wire with a length L and a cross-sectional area S and suppose a potential difference V is applied across the ends of the wire (V is also called a voltage). Let us designate the current that flows through the wire I and the resistance of the wire R. According to Ohm's law we have

$$I = \frac{V}{R}, \text{ whence } R = \frac{V}{I}. \tag{I.1}$$

If V is measured in volts and I in amperes, then R is measured in ohms. The reciprocal of R (i.e. $1/R$) is called the *conductance*. It is measured, therefore, in reciprocal ohms or mhos (Ω^{-1}). We know from experience that a wire's conductance is lower, the longer it is and the smaller its cross-sectional area, i.e.

$$\frac{1}{R} = \sigma \frac{S}{L}. \tag{I.2}$$

The factor σ here depends on what the wire is made of and is called the *electrical conductivity*, or *specific conductance* of the material. It follows from (I.2) that

$$\sigma = \frac{L}{SR}. \tag{I.3}$$

If L is measured in centimeters, S in square centimeters, and R in ohms, then we obtain for the

'dimension' of σ:

$$[\sigma] = \frac{\text{cm}}{\text{cm}^2 \cdot \Omega} = \Omega^{-1} \cdot \text{cm}^{-1}.$$

The International System of Units (SI) is used by scientists at present, and its unit of conductance is the siemens (S), 1 S = 1 Ω^{-1}. Therefore the SI unit for conductivity σ is $\text{S} \cdot \text{m}^{-1}$.

Note, by the way, that Ohm's law can be viewed differently on the basis of (I.2). Substituting (I.2) into (I.1), we get

$$I = \sigma S \frac{V}{L}, \quad \text{whence } \frac{I}{S} = \sigma \frac{V}{L}. \tag{I.4}$$

If we introduce the following notation:

$$i = \frac{I}{S} \text{ and } E = \frac{V}{L},$$

we get instead of (I.4)

$$i = \sigma E. \tag{I.5}$$

This is nothing more than Ohm's law written differently. Here i is *current density* (the current per unit cross-sectional area of the conductor), while E is the *strength of electric field* (the change of potential per unit length). It follows from Ohm's law given in form (I.5) that the current density is directly proportional to the electric field strength (and this assertion is Ohm's law) with the conductivity σ being the proportionality constant. It is different for different crystals.

For metals, σ is about $10^4\ \Omega^{-1}\cdot cm^{-1} = 10^6 S \times \times m^{-1}$. For dielectrics, σ is about $10^{-15}\ \Omega^{-1} \times \times cm^{-1} = 10^{-13}\ S\cdot m^{-1}$. Thus the conductivity of a dielectric is 10^{19} times less than that of a metal. This is an enormous number (a one followed by nineteen zeros).

We shall pause here to make an important comment. There is a vast group of solids whose conductivities are essentially less than those of metals, but at the same time considerably greater than those of dielectrics. These solids have became objects of study because of some of the remarkable properties they have. These solids are called *semiconductors* (they could as well have

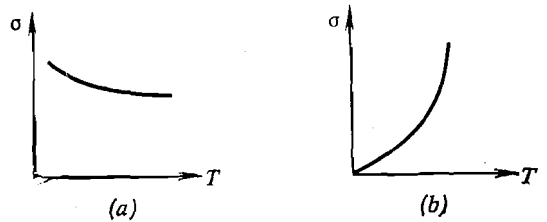


Fig. 1. onductivity as a function of temperature: (*a*) in metals, (*b*) in dielectrics and some semiconductors

been called semidielectrics). The conductivities of semiconductors vary widely from $\sigma = 10^2$ to $\sigma = 10^{-10}\ \Omega^{-1}\cdot cm^{-1}$ or, what is the same, from $\sigma = 10^4$ to $\sigma = 10^{-8}\ S\cdot m^{-1}$. However, the difference between metals, semiconductors, and dielectrics is not just that they have different conductivities, but mainly that the conductivities

of the materials in these groups respond differently to the same external factors.

The simplest factor is heat. When a metal is heated, its conductivity decreases slowly, as shown in Fig. 1*a*. A dielectric behaves the other way round, with its conductivity increasing rapidly as the temperature is increased. This is presented diagrammatically in Fig. 1*b*. As far as semiconductors are concerned, some behave like metals, while others, conversely, behave like dielectrics. A rise in the temperature of some (very many) semiconductors can increase their conductivities by several million times.

Another factor to which the conductivity of a crystal is usually sensitive is the introduction of an impurity. An impurity may be a foreign

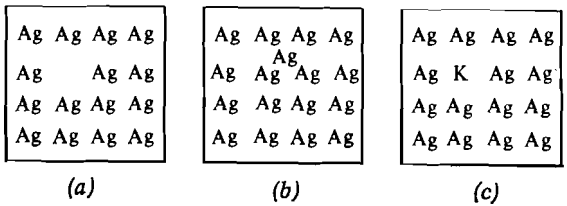


Fig. 2. Structural defects in a crystal of silver: (*a*) vacancy, (*b*) interstitial atom, (*c*) foreign atom substituting a proper atom of the lattice

atom (or ion) which is introduced into an interstitial space in the crystal lattice, or which substitutes an atom (or ion) in the lattice itself. Typically, a real crystal contains a certain concentration of impurities if the crystal

has not been subjected to special treatment. An atom (ion) of the lattice itself that is not located where it should be, i.e. not at a lattice point but in the interstice, and a vacant point in the lattice, i.e. a point from which an atom (or ion) has been removed (it is called a *vacancy*) also fall into the category of impurities. In the physics of crystals, the notion of 'impurity' is wider than what we attribute to it in our daily life. An impurity is a *defect* that spoils the regular pattern of a crystal lattice (Fig. 2).

The conductivity of a metal decreases somewhat when an impurity of any kind is introduced, and decreases further the more impurity there is. As a rule, the conductivity of a dielectric is, on the contrary, somewhat increased by an impurity. Some semiconductors behave like metals, but in most cases semiconductors are extremely sensitive to impurities and the introduction of even tiny amounts of impurity results in large gain in their conductivities. Thus, one foreign atom per thousand proper atoms in a lattice can change its conductivity by many hundreds of times. This is indeed a 'homeopathic' effect.

Let us note one more factor on which the conductivity of a crystal can depend. This is the application of an electric field to the crystal. All metals, however strong the field may be, obey Ohm's law (remember (I.5)). It states that the current in the metal is directly proportional to the strength E of the field, i.e. σ stays constant and does not depend on E. In nonmetal crystals, Ohm's law is only observed for small values of E. If E increases further Ohm's law is broken and

the current either starts increasing faster than the strength of the field or sometimes more slowly than is required by Ohm's law (Fig. 3), i.e. σ stops being constant and changes with the increase of E. At large enough E the crystal is destroyed or *breakdown* occurs.

When we speak of electric current, we mean the transfer of an electric charge in one direction. The charges in crystals are mainly the electrons that leave the atoms or ions they belong to and start wandering round the crystal. This kind of current is called *electron current*. In ionic crystals the charge carriers are the ions themselves as they can leave their lattice points for the interstitial space and thus they acquire the ability to travel through the crystal. In this case we say there is an *ion current* in the crystal. Electron current occurs in metals and semiconductors,

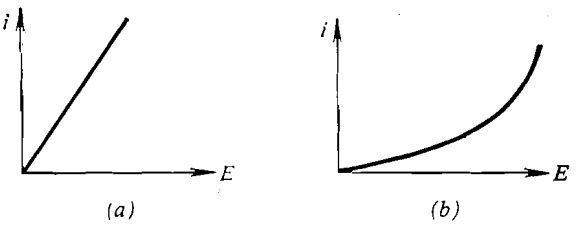


Fig. 3. Current density i as a function of strength E of the field: (*a*) in metals, (*b*) in nonmetallic crystals

while in dielectrics ion current (at moderate E) is the rule which gives way to electron current at large E. Later we shall describe in more detail

how current flows in the crystal, and the various types of electron and ion currents.

In this section we described some of the features and properties of metals, semiconductors, and dielectrics which occur because of the behaviour of the electrons within the crystals. Electrons govern the properties of crystals. We shall try now to explain how they do it.

Chapter 1

Electrons in Metals

1.1. 'Free' and 'Bound' Electrons

In this chapter we shall discuss the electrons populating metals, their behaviour and the laws they obey. Our aim here is to explain some of the properties of metals.

Let us consider a metal crystal made of atoms of one kind. For simplicity's sake, suppose they are monovalent atoms (e.g. Ag or Na). A discussion of metals with higher valences (e.g. bivalent ones like Zn or Ni) will add nothing fundamentally new to our picture.

Let us take only one atom in the lattice, imagine all the others around it are at infinity and see what happens when we move them back into place to form a crystal lattice.

Our atom consists of a positively charged atomic core and a valence electron at some distance from it. The atomic core can be approximated as a point charge (an analogue of a hydrogen atom). Let us assume that it is positioned at a *reference point (origin) O*. Thus, the valence electron travels in the Coulomb field of a point charge. The potential energy U of the electron depends directly on $|x|$, the distance of the electron from the reference point (the U versus x curve is given in Fig. 4). We shall discuss a unidimensional model, i.e. we consider the crystal to be

a chain of atoms, therefore

$$U = -\frac{e^2}{|x|}, \tag{1.1}$$

where e is the charge of the electron. Expressions for potential energy always require an extra additive constant which fixes the origin of the energy scale. In (1.1) this constant has been chosen so that the potential energy U is zero at $x = \infty$, i.e. when the electron is at infinity. If

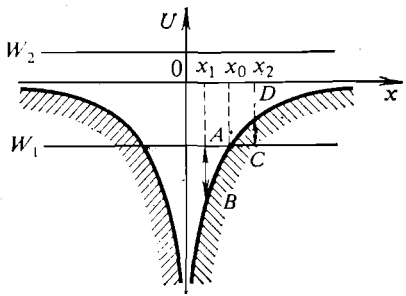


Fig. 4. Potential crater of an isolated atom

the origin is chosen this way the potential energy U is negative for all finite x.

Although the potential energy U vs. x is shown in Fig. 4, we can also show there the total energy of the electron. Let us designate it W. Obviously W does not depend on x, i.e. it is shown by a horizontal straight line. It can have a specified value, but it should be constant regardless of the motion of the electron. The horizontal lines W_1 and W_2 are two arbitrary values of W.

The kinetic energy K of the electron can be shown in the same figure. Clearly

$$W = K + U \text{ and therefore } K = W - U. \tag{1.2}$$

If the electron has a total energy W_1 and is x_1 away from the origin, then its kinetic energy is given in Fig. 4 (see (1.2)) as AB. But if the electron is at a distance x_2, then its kinetic energy is given as CD. In this case $W < U$ and hence $K < 0$, i.e. the kinetic energy turns to be negative which is impossible. Indeed, in classical mechanics

$$K = \frac{1}{2} mv^2,$$

where m is the mass and v is the velocity of the electron. Evidently, the kinetic energy K can only be negative if the mass of the electron is negative, which is absurd, or if the velocity v of the electron is an imaginary number, which is no less absurd. It follows that an electron with a total energy of W_1 cannot be farther from the origin than x_0. The area that is cross-hatched in Fig. 4 is forbidden for the electron from the viewpoint of classical mechanics. Obviously, the higher the total energy W of the electron, i.e. the higher the line W_1 in Fig. 4 is, the broader is the area allowed for the electron. The inference is that this is a *bound* electron, i.e. an electron attached to an atom and incapable of leaving it.

Conversely, if the total energy of the electron is given in Fig. 4 by the horizontal line W_2, which is above U everywhere, then $K > 0$ for any x, and the electron can go anywhere. This implies

that this electron is *free* and its bond with the atom has broken; it has the right to travel as far from the atom as need be.

Consequently, a bound electron has a negative total energy W ($W < 0$, see the horizontal line $W = W_1$ in Fig. 4, which is *below* the abscissa),

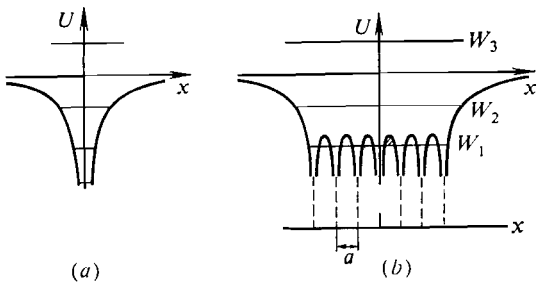


Fig. 5. Potential energy diagram for an electron: (*a*) in an isolated atom, (*b*) in a crystal

and a free electron has a total energy W that is positive ($W > 0$, see the horizontal line $W = W_2$, which is *over* the abscissa). An electron in a bound state can be transferred into the free state if it is supplied with the energy needed for the transition from level W_1 to level W_2. This act of release of a bound electron is called the *ionization of the atom*.

So far we have discussed an isolated atom. Now let us consider a chain of atoms spaced out regularly with an interval a between them (see Fig. 5*b*), i.e. the unidimensional model of a crystal. This is a system of atomic cores and a system of electrons. Each electron interacts with its

own atomic core, with all the other cores and with all the other electrons. We are going to neglect the last category, although there are no logical grounds to do so. Indeed, the interactions between the electrons are of the same order of magnitude as their interactions with the atomic cores. There is no real reason for neglecting the electron-electron interactions while taking into account the electron-core ones, but we shall do so, and shall consider each electron to be independent of the others, as it were. This simplifies the problem appreciably and then we can return to account for the electron-electron interactions.

Let us consider a model in which each electron moves in the common field of all the atomic cores. The electron's potential energy in such a force field can be given by a curve that is the sum of several curves like the one shown in Fig. 4, but shifted over a distance a. The resultant curve (the potential energy U of the electron as a function of x) is plotted in Fig. 5*b*. This is a system of potential craters alternating with potential barriers and fringed on both the right and the left by potential thresholds. The thresholds correspond to the end atoms (the first and the last) of the chain. The potential crater of a single atom is shown once again on the left (Fig. 5*a*). The essential point about Fig. 5*b* is that the potential energy U of the electron within the crystal lattice is periodic with a period equal to the lattice constant a.

Let us ask ourselves what the behaviour of our electrons is now. Have they acquired the ability to travel around the crystal, or will they remain attached to their respective atoms? The

answer depends on the total energy W of the electron and there are three cases:

(1) If the electron is on level W_1, i.e. if it has a total energy shown by a horizontal line that cuts the tops of the barriers, then according to classical mechanics the electron is locked within its potential crater, i.e. it cannot get into the adjacent potential crater. In fact for an electron at W_1, to get into the next crater, it must pierce the potential barrier, i.e. the region in which its total energy is more than its potential energy and which therefore is forbidden to it. Consequently, the crystal turns to be, as it were, divided by walls that are impenetrable for the electron.

(2) If the electron is on level W_2, which passes over the tops of the potential barriers (as shown in Fig. 5*b*), then it can travel freely within the crystal, passing from atom to atom. However it cannot go beyond the crystal as the space in which the electron is permitted to travel is bounded by the potential barriers on the left and on the right.

(3) And lastly, if the total energy of electron in Fig. 5*b* is given by W_3, then the electron can not only travel without hindrance within the crystal, but it can also go beyond the crystal, however far.

So far we have tacitly assumed that the electron, whose behaviour we discussed, is a classical particle, i.e. it obeys the laws of Newton's classical mechanics. But in reality electrons are quantum particles and obey quantum mechanics, which dominates the microcosm. How is our picture corrected by quantum mechanics?

From the viewpoint of quantum mechanics, there is a nonzero probability that an electron

may exist where it cannot occur from the purely classical viewpoint. In other words, a quantum particle can penetrate the area forbidden to a classical particle. Accordingly, there is a nonzero probability that an electron with a total energy W_1 (Fig. 5*b*) can penetrate the area $W_1 < U$, i.e. inside the potential barrier, and therefore escape through it. The probability of such an escape increases the narrower and lower the potential barrier becomes, i.e. as the cross-hatched area in Fig. 5*b* decreases. An electron on one side of the barrier can occur on the other side owing to a 'tunnel' transition through the barrier even though it possesses a total energy W_1. This 'tunnel' transition (tunnelling effect) is characteristic for modern solid-state quantum theory and we shall discuss it in this book more than once.

Thus, quantum mechanics results in an 'absolutely' bound electron becoming less bound. The use of quantum mechanics in solid-state theory thus produces a certain 'liberation' of bound electrons.

Let us make one more remark in conclusion. Perhaps you remember that at the beginning of this section we assumed the simplifying but groundless approximation that there were no electron-electron interactions. This type of interaction can be taken into account by introducing what is called a *self-consistent* field. This field is the one produced by the averaged charge of all the electrons, and each electron is considered to move both in the field of the atomic cores and in the self-consistent field of the electrons. The field makes us include another summand into

the expression for the potential energy. However, this does not interfere with the potential energy's basic property, viz. its periodic nature, and we may still use Fig. 5*b*. Accounting for the interactions between electrons is an involved problem and we shall not discuss it anymore. Let us only note that in modern solid-state quantum theory a system of interacting electrons is considered to be a system of *noninteracting* '*quasiparticles*' (or *elementary excitations*).

1.2. 'Electron Gas' in Metals

Let us return to Fig. 5*b*, which is a diagram of the potential energy of an electron in a metal. The diagram can be substantially simplified if the periodic function of the potential energy is replaced by a value averaged over the entire crystal, and if the potential thresholds at the boundaries of the crystal are assumed to be vertical. Fig. 5*b* thus turns into Fig. 6. There is now no difficulty in going from our unidimensional crystal model (a chain of atoms) to a three-dimensional model, in which the crystal is represented as a cube (Fig. 7) with an edge L and containing N atoms, so that

$$L = aN, \tag{1.3}$$

where a is the distance between adjacent atoms (the lattice constant). The x-axis in Fig. 6 can be replaced by the y- or z-axes.

This simplified model of the crystal, which is a potential box with a flat bottom and vertical walls, is the basis of Drude's classical theory of metals and Sommerfeld's quantum theory of

metals. All the electrons in the system 'live' in the potential box. To take a definite example for the metals we shall still consider them to consist of monovalent atoms. Thus the number of

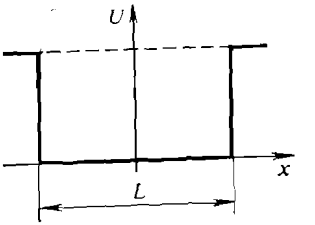


Fig. 6. Potential box

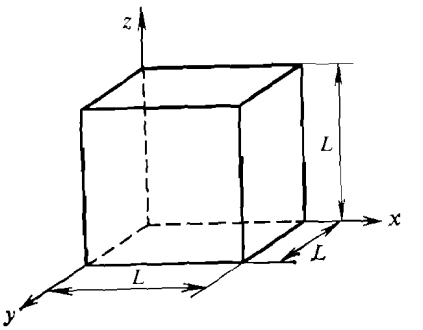


Fig. 7. Cubic crystal

electrons in the system is equal to N^3, and the concentration of electrons N_0, will according to (1.3) be

$$N_0 = \frac{N^3}{L^3} = \frac{1}{a^3}. \tag{1.4}$$

This is a very large number. There are about 10^{23} electrons in a cubic centimeter of a metal crystal (a one followed by 23 zeros!). All of these electrons can be considered to be free since they move in the absence of any force field (the potential energy of an electron within a crystal is constant), i.e. they behave like the molecules of an ideal gas. A metal in this model is a framework of atomic cores immersed in a gas of electrons.

Each electron at a given moment in time occupies a definite position in space or, in other words, it has three definite coordinates x, y, z, and a finite velocity $\mathbf{v}$. Thus it can also be described by the three components of its velocity v_x, v_y, v_z, or it occupies a definite position in velocity space, i.e. in the space where the coordinates are not x, y, z, but the velocities v_x, v_y, v_z.

Now let us recall that our electrons are quantum particles. Each velocity component proves to be *quantized*, i.e. it can not have just any value, it must have one of a set of distinct values, i.e.

$$v_x = \frac{h}{maN} n_x, \quad v_y = \frac{h}{maN} n_y, \quad v_z = \frac{h}{maN} n_z. \tag{1.5}$$

Here h is *Planck's constant* (it is always present in quantum-mechanical formulas) and n_x, n_y, n_z are arbitrary, but always integer-valued numbers (including zero):

$$n_x, n_y, n_z = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{1.6}$$

The numbers n_x, n_y, n_z are called *quantum numbers*. We see that the velocities v_x, v_y, v_z are not continuous, but are a discrete sequence of values.

It follows that velocity space consists of uniform cells, $\Delta\gamma$ in volume (Fig. 8), i.e.

$$\Delta\gamma = \left(\frac{h}{maN}\right)^3.$$

In agreement with (1.5) and (1.6), an electron may not therefore exist anywhere in velocity

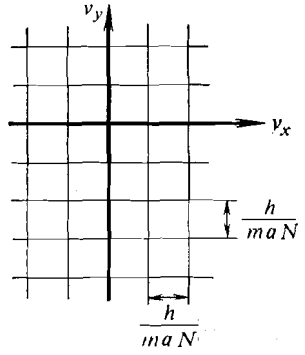


Fig. 8. Quantized space of velocities

space but may only exist at the points of the lattice shown in Fig. 8.

According to (1.5), the energy W of the electron is quantized as well. Indeed, the potential energy U of the electron in our model is constant throughout the crystal and hence it can be assumed to be zero without loss of generality. The total energy of an electron is therefore equal to its kinetic energy:

$$W = \frac{1}{2} mv^2 = \frac{1}{2} m (v_x^2 + v_y^2 + v_z^2). \tag{1.7}$$

Substituting (1.5) into this, we get

$$W = \frac{h^2}{2ma^2N^2} (n_x^2 + n_y^2 + n_z^2). \tag{1.8}$$

We see that the total energy W, like the three velocity components v_x, v_y, v_z can only assume a discrete sequence of values. Both W and v_x, v_y, v_z are shown in Fig. 9 as functions of the quantum numbers n_x, n_y, n_z. In Figs. 9*a* and 9*b* the values of n_x (or n_y or n_z) are plotted along the

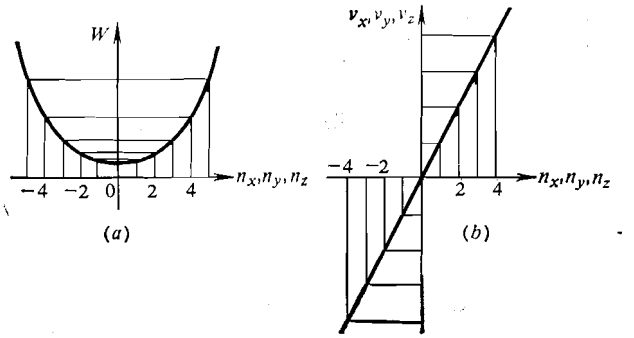


Fig. 9. Quantization of energy (*a*), quantization of the three components of velocity (*b*)

abscissa, and the values of W in Fig. 9*a* and the values of v_x (or v_y or v_z respectively) in Fig. 9*b* along the ordinate.

Now we have to take an important step in our presentation and introduce what is called the *Pauli principle* which a pool of electrons obey. This principle states that only one electron can possess a given set of three quantum numbers. In other words, each point in velocity space can only be occupied by one electron. The Pauli principle is the principle of impenetrability carried over from usual coordinate space into

velocity space. The impenetrability principle forbids two or more electrons to occupy the same point in space at the same time. This is why it is often called the Pauli *exclusion principle*. There are some essential consequences of the Pauli principle that we shall come across more than once.

It is necessary to make a correction here. The three quantum numbers are not an exhaustive description of the state of an electron. There is a fourth parameter called *spin*. The presence of spin is just as essential and inalienable a property of an electron as its mass or electric charge.

We can imagine electron spin best of all if the electron is imagined to be a small ball spinning around its axis. It can spin around an arbitrary axis both clockwise and anticlockwise. That is why two values of spin are possible. According to the Pauli principle not one, but two electrons can possess a given set of three quantum numbers if the electrons possess opposite spins. In other words, if electron spin is taken into account, each point in velocity space can accomodate two electrons.

Obviously, all electrons of our system should somehow be distributed in velocity space. The total summed energy of our system (i.e. the sum of the energies of all the electrons) depends on this distribution. Let us consider which distribution satisfies the requirement of minimum energy, or, what is the same, the condition of absolute zero temperature. (Recall that absolute zero is the temperature at which the energy of a system is at the minimum.)

But for the Pauli principle, every electron would occupy the point $v_x = v_y = v_z = 0$, i.e.

the origin of the velocity space. Every electron would thus be at rest and the energy of the system would be zero. However, this is forbidden by the Pauli principle. The energy of the system is at the minimum, albeit nonzero, if the electrons

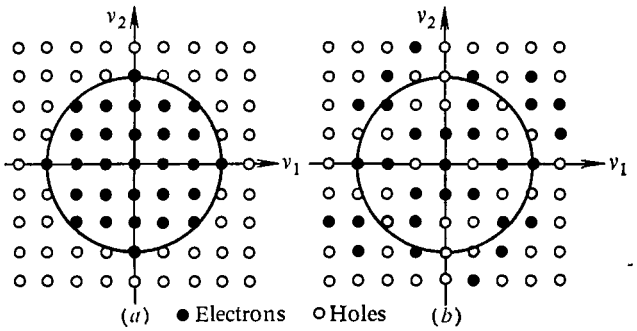


Fig. 10. Distribution of electrons in velocity space: (*a*) for absolute zero, (*b*) for a temperature above absolute zero

are distributed as close to the origin as possible while obeying the Pauli principle, i.e. if they fill a sphere containing half as many cells as there are electrons (each cell contains two electrons). The surface of this sphere is called the Fermi surface and the sphere itself is called the *Fermi sphere*. This distribution is shown in Fig. 10*a*, where v_1 and v_2 are any two numbers out of the set of three numbers v_x, v_y, v_z. It is clear because of this that the motion of electrons cannot stop even at absolute zero. This follows from the application of the Pauli exclusion principle to the pool of electrons.

When the temperature rises, some of the electrons in velocity space escape the sphere shown in Fig. 10*a* leaving vacant places. For that reason we go from Fig. 10*a* to Fig. 10*b*. We can say that heating loosens the dense packing of the electrons in velocity space. The higher the temperature, the looser the packing gets, and the summed energy of the electrons rises.

It is essential that there is a symmetry in the distribution of the electrons in velocity space. In other words, each electron with a velocity v has a counterpart with a velocity —v, so that the average vector velocity of all electrons is still zero at any temperature. This means that there is no electric current in the crystal.

That is the way it is until an extraneous force is applied to the electrons. If the crystal is placed in an external electric field by applying a difference of potentials to it, the symmetric distribution of electrons is upset. A predominant direction then appears in which the number of occupied places is greater than in the opposite direction. The average velocity of the electrons then becomes nonzero. This means that there is an electric current in the crystal.

We have already noted that the degree of loosening of the electron packing in velocity space increases with a rise of temperature. Commonly this degree of loosening is characterized by a function $F(W)$ that we shall call the *energy distribution function*. To define it consider a sphere in velocity space with a layer of a radius v and a thickness dv. It follows from (1.7) that

$$v = \sqrt{2\frac{W}{m}} \qquad (1.9)$$

and

$$dv = \frac{1}{\sqrt{2mW}}\, dW. \qquad (1.10)$$

The number of electrons in this layer, i.e. the number possessing energies from W to $W + dW$, is designated $X(W)\,dW$. The number of points available for the electrons (twice the number of cells in the velocity space) that the layer contains is designated $Y(W)\,dW$. The *distribution function* is the ratio of these two numbers:

$$F(W) = \frac{X(W)\,dW}{Y(W)\,dW}. \qquad (1.11)$$

This function has for our electron gas the form

$$F(W) = \frac{1}{1+\exp\dfrac{W-W_F}{kT}}. \qquad (1.12)$$

Here T is the absolute temperature, k is Boltzmann's constant, which should be familiar to the reader from the kinetic theory of gases, and W_F is the energy corresponding to the Fermi surface. The energy W_F is called the *Fermi energy* or *Fermi level*, and the function given in (1.12) is the *Fermi distribution function*. Particles described by the Fermi function are said to satisfy the Fermi-Dirac statistics.

The function in (1.12) is illustrated in Fig. 11. The step shown by the thin line corresponds to the function's shape at absolute zero. We see that this is the case when all the cells in velocity space that are within the Fermi sphere are filled by electrons ($F = 1$), while all cells that are beyond this sphere are vacant ($F = 0$). The smooth

bold curve in Fig. 11 corresponds to a temperature above absolute zero. This curve $F = F(W)$ passes through point A (see Fig. 11) for which $W = W_F$, $F = 1/2$ at any temperature.

To conclude this section, let us return to what we noted in the beginning. Sommerfeld's quantum theory of metals deals with the same model

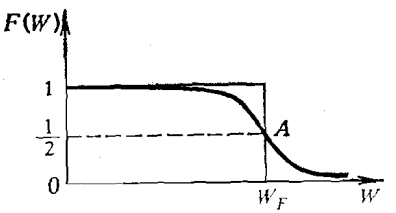


Fig. 11. Fermi distribution function

as Drude's classical theory: that is an electron gas locked in a potential box with a flat bottom. But the electron gas Sommerfeld considered differs from the electron gas of the classical theory in two respects.

(a) Each electron in Drude's theory obeys classical mechanics, while in Sommerfeld's theory it is considered to be a quantum particle governed by the laws of quantum mechanics.

(b) When describing the behaviour of a pool of electrons, Drude's theory relied on the Maxwell-Boltzmann classical statistics, i.e. the statistical laws that describe an ordinary gas consisting of independent molecules. Sommerfeld's theory applies the Fermi-Dirac quantum statistics to the electron gas, i.e. the statistical

laws that govern particles which are, first, identical and, second, controlled by the Pauli principle. However, it is essential to point out and emphasize that all the formulas of the quantum statistics are transformed under certain conditions (which are going to be discussed below) into the classical formulas. The electron gas would then be called *nondegenerate.*

1.3. The Successes and Failures of the Classical Theory of Metals

Metals, unlike nonmetallic crystals, have very high electrical and thermal conductivities. We can use our earlier notation, i.e. electrical conductivity σ, and thermal conductivity $\varkappa$, and further designate the current density in metals i. This is directly proportional to the potential *gradient* dV/dx. Let us designate the thermal current density j. It is directly proportional to the temperature gradient dT/dx. (The derivatives dV/dx and dT/dx show how fast the potential and temperature, respectively, change along the x-axis.) We have

$$i = -\sigma \frac{dV}{dx}, \tag{1.13}$$

$$j = -\varkappa \frac{dT}{dx}. \tag{1.14}$$

Equation (1.14) holds true if there is no external electric field. The minus sign on the right-hand side of equation (1.14) means that the thermal flux (i.e. the kinetic energy carried by electrons) is in the opposite direction to the temperature gradient. In other words, heat is transferred

from where the crystal is hotter to where the temperature is lower.

Equation (1.13), on the contrary, refers to the case in which the metal is in an electric field, but the temperature is constant throughout the crystal, i.e. the temperature gradient is everywhere zero ($dT/dx = 0$). Recall that

$$-\frac{dV}{dx} = E, \tag{1.15}$$

where E is electric field strength, and substitute (1.15) into (1.13) to obtain equation (I.5):

$i = \sigma E$

which is Ohm's law (for σ independent of E).

The constants σ and $\varkappa$ in (1.13) and (1.14) can be calculated. The results of these calculations we shall cover below. Let us note that the classical and quantum theories bring about different results even though both theories proceed from the same model.

According to both the classical and quantum theories the electrons in metals are in constant chaotic thermal motion colliding with the atoms (with the atomic cores, to be more exact) of the lattice. The atoms (atomic cores) are arranged in a regular pattern and oscillate about their points of equilibrium. The amplitude of these oscillations depends on the temperature; the higher the temperature the larger the oscillations. According to the quantum theory, these oscillations do not stop even at absolute zero.

When there is no electric field, the electron's motion between two collisions is linear and uniform. But when a field is present, each electron moves between the collisions with an acceleration

proportional to the field strength E and accumulates a kinetic energy that is then transferred to the lattice via collisions. This is the source of the resistance that the lattice offers to current, and the source of the heat that is evolved when a current flows.

However, there is an essential difference between the classical and quantum theories here. In the classical theory, the resistance is caused by the very presence of the crystal lattice, which resists the freely moving electrons. According to the quantum theory, the lattice does not itself resist the flow and is transparent to the electrons, they do not notice it. The resistance is due to the oscillations of the lattice. The greater the temperature, i.e. the greater the amplitude of the oscillations, the greater the resistance. It is the atoms, deviating from their equilibrium positions, that scatter the free electrons. The foreign atoms of an impurity, which are outside or inside the regular pattern of the lattice, are additional centers of scattering and offer additional resistance.

The mean distance an electron travels between two successive collisions with the lattice atoms is called the *mean free path*. We shall designate this value by the letter l. It decreases as the temperature increases.

The classical and the quantum theories produce the following expressions for σ and $\varkappa$. We have in the classical theory

$$\sigma = \frac{4}{3} \frac{N_0 l e^2}{\sqrt{2\pi m k T}}, \tag{1.16a}$$

$$\varkappa = \frac{4}{3} N_0 l \sqrt{\frac{2k^3 T}{\pi m}}, \tag{1.17a}$$

where N_0 is the concentration of free electrons as before, e is the electron charge, and m is the mass of an electron. The quantum theory produces

$$\sigma = \frac{2\pi}{3} \frac{e^2 l}{h} \left(\frac{3N_0}{\pi}\right)^{2/3}, \tag{1.16b}$$

$$\varkappa = \frac{2\pi^3}{9} \frac{lk^2 T}{h} \left(\frac{3N_0}{\pi}\right)^{2/3}. \tag{1.17b}$$

Apart from the universal constants h, e, and k, these expressions include N_0 and l, which are different for different metals and must be determined.

When dealing with electrical and thermal conductivities we cannot avoid the Wiedemann-Franz law. This law was discovered experimentally and it states that the ratio of the thermal conductivity to the electrical conductivity of a metal is a constant independent of the metal and is directly proportional to the absolute temperature. Both the quantum and the classical theories of metals explain this law. Dividing (1.17a) into (1.16a) we obtain for the classical theory

$$\frac{\varkappa}{\sigma} = 2 \left(\frac{k}{e}\right)^2 T. \tag{1.18a}$$

And dividing (1.17b) into (1.16b) we get for the quantum theory

$$\frac{\varkappa}{\sigma} = \frac{\pi^2}{3} \left(\frac{k}{e}\right)^2 T. \tag{1.18b}$$

Evidently, this is a case when both the classical and quantum theories bring about almost the

same results. The only difference is in the factor of $\pi^2/3 \simeq 3.3$ in the quantum theory instead of 2 in the classical one.

The deduction of the Wiedemann-Franz law is a great achievement of the classical theory of metals. We could also cite some other inferences from the classical theory that are in good agreement with experiment. However, the development of the classical theory came to a dead end. This was the problem of the heat capacity. The classical theory gave results that were in stark contrast to experimental results. Let us discuss this problem in some detail.

The heat capacity of a metal is the sum of two components: the heat capacity that is due to the thermal oscillations of the lattice atoms, and the heat capacity of the electron gas. Let us designate the electron heat capacity per unit volume c and designate the total energy of the electron gas per unit volume $\overline{W}$. $\overline{W}$ is the sum of the kinetic energies of all the electrons in a cubic centimeter of a metal. By definition

$$c = \frac{d\overline{W}}{dT}. \tag{1.19}$$

For that reason the specific heat capacity of the electron gas indicates the rate at which the total energy of the electron gas rises with temperature.

According to the classical theory, which attributes to the electron gas all the properties of an ordinary ideal gas, we have

$$\overline{W} = \frac{3}{2} N_0 kT. \tag{1.20}$$

and therefore, according to (1.19)

$$c = \frac{d}{dT}\left(\frac{3}{2} N_0 kT\right) = \frac{3}{2} N_0 k. \tag{1.21}$$

However, the experimental data show that the heat capacity of the lattice is almost equal to the total heat capacity of the metal, so that almost nothing is left for the electron gas. This

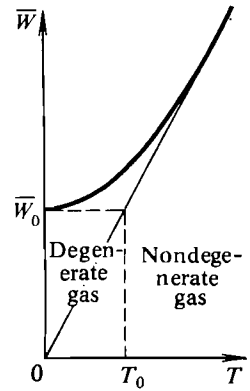


Fig. 12. Total energy $\overline{W}$ of electron gas as a function of temperature T

means that N_0 in (1.21) must be small but then the conductivity σ, the expression for which also contains N_0, according to (1.16a) and (1.16b), turns to be too small. The way out of this difficulty was the use of the quantum theory instead of the classical one. We now obtain another expression for $\overline{W}$ instead of (1.20).

We are not going to give this expression here and naturally, we are not going to deduce it. Instead, we shall present Fig. 12 in which the total energy $\overline{W}$ of the electron gas is shown as a function of temperature T. The bold curve com-

plies with the quantum theory of metals, while the thin straight line follows from the classical theory. The steepness of the curve at each given point characterizes the heat capacity of the electron gas at the respective temperature. We see that according to the classical theory the heat capacity does not depend on temperature. But the heat capacity in the quantum theory increases as the temperature rises: it grows from zero (at absolute zero) to its classical values (1.21) (at high temperatures). The temperature T_0 is called the *degeneration temperature* and an electron gas behaves like a classical one at temperatures that are substantially greater ($T \gg T_0$). Such a gas is called *nondegenerate* while the electron gas is said to be *degenerate* at temperatures less than the temperature of degeneration. The heat capacity of a degenerate gas is lower than its classical value and tends to zero as the temperature falls. The degeneration temperature T_0 can be determined thus

$$\frac{3}{2} N_0 k T_0 = \overline{W}_0,$$

whence

$$T_0 = \frac{2}{3} \frac{\overline{W}_0}{N_0 k}, \qquad (1.22)$$

where $\overline{W}_0$ corresponds to the Fermi energy (see section 1.2). By deducing $\overline{W}_0$ and substituting it into (1.22), it can be shown that the lower N_0 the lower T_0. For metals, in which, as we have already noted, $N_0 = 10^{23}$ cm^{-3}, the degeneration temperature proves to be higher than the melting point. That is why the electron gas in metals is

in practice degenerate at all temperatures. This accounts for the failures of Drude's classical theory because it assumed a nondegenerate gas.

1.4. Electron Emission from Metals

As we mentioned before the potential in a metal can be represented as a potential box or potential well containing electrons. It is thus natural to ask whether there is any way of extracting the electrons from a metal, which is a practically inexhaustible source of them.

We shall consider three of such methods in this section. A metal can be made to emit electrons as the result of:

(1) heating it to a high temperature (to induce the *thermionic emission*);

(2) applying a strong electric field (to induce the *field* or *cold emission*);

(3) radiating it with high frequency light (to induce the *photoelectric effect*).

We shall discuss each of these methods.

1. Only those electrons filling the metal that are capable of overcoming the potential threshold W_m shown in Fig. 13*a* can leave the metal, i.e. go beyond its surface. This figure is very much like Fig. 6 (only W is the ordinate instead of U; Fig. 13*a* shows the energy levels occupied by electrons at absolute zero). Obviously, only those electrons that possess sufficient velocity in the direction at right angles to the surface of metal can make it. If we choose a coordinate system such that the x-axis is perpendicular to the surface of metal, then there will be a lower limit v_{x0} for the velocity component v_x such that only elec-

trons for which $v_x > v_{x0}$ will be able to overcome the potential threshold.

The velocity v_{x0} can be determined from

$$\frac{1}{2} m (v_{x0})^2 = W_F$$

(see Fig. 13*a*). It is clear from Fig. 13*b*, which shows the Fermi distribution function (as in

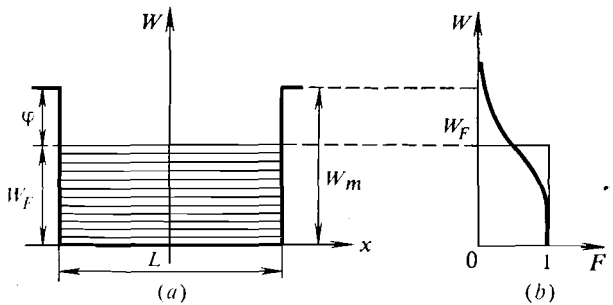


Fig. 13. Metal as a potential well (*a*), Fermi distribution function (*b*)

Fig. 11), that the electrons under discussion belong to the 'tail' of the Fermi distribution function.

Calculations of the thermionic current density i, i.e. the flux of electrons liberated from a unit surface, produce

$$i = A (kT)^2 \exp\left(-\frac{\varphi}{kT}\right), \qquad (1.23)$$

which is in good agreement with experimental data. The quantity φ (see Fig. 13*a* as well) is called the *work function*. It is a constant that is different for different metals and is one of their characteristics. The work function is the smallest amount of energy an electron must spend to leave the metal at absolute zero.

As can be seen from (1.23), the way the thermionic current i depends on temperature T is determined both by the exponent and the pre-exponent factor. They both act in the same direction hence i increases with the rise of T. However, the pre-exponent factor changes with the temperature very slowly (compared with the rapid change due to the exponent), and therefore it can be assumed to be constant at a first approximation. Thus

$$i = C \exp\left(-\frac{\varphi}{kT}\right),$$

or

$$\ln i = \ln C - \frac{\varphi}{kT}. \qquad (1.24)$$

If we plot ln i and $\frac{1}{kT}$ along two coordinate axes then equation (1.24) will be a straight line whose slope α gives the work function φ (see Fig. 14). Commonly this value is several electron volts*.

The phenomenon of thermionic emission, the outflow of electrons from a heated metal, is also known as the *Richardson effect*.

2. A metal can be made to emit electrons without heating it. All that is necessary is to put the metal into a strong external electric field. Suppose there are two flat electrodes facing each other,

* An electron volt is a unit of energy and equals the energy acquired by an electron when it passes through a potential difference of one volt.

and let us consider the metal is the cathode. The electric field with a strength E is at right angles to the metal surface. The bold curve in Fig. 15

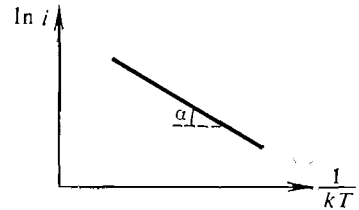


Fig. 14. Thermionic current as a function of temperature ($\tan \alpha = \varphi$)

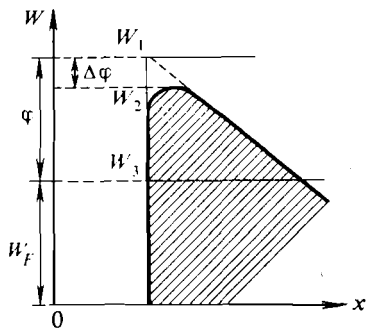


Fig. 15. Potential energy diagram of an electron in an external electric field

shows the potential energy of an electron as a function of x (bear in mind that the x-axis is perpendicular to the surface of the metal). The thin curve in Fig. 15 shows the potential energy in the absence of the field ($E = 0$). We see that the field transforms the potential threshold on

the metal surface into a potential barrier. There are two possible mechanisms for the appearance of the electron emission current in this case.

First of all, the external field decreases the work function φ by $\Delta\varphi$. This facilitates thermionic emission, which is extremely sensitive to changes in φ (according to (1.23) φ is in the exponent). Thus when the external electric field is applied, thermionic emission becomes noticeable at lower temperatures than in the absence of the field. When there is no field, only the electrons for which $W > W_1$ could be emitted (see Fig. 15, W, as before, is the total energy of an electron). When the field is present, electrons for which $W_2 < W < W_1$ can also be emitted.

Another mechanism for electron emission is possible, in which slower electrons can take part. We refer to electrons whose energy W is shown in Fig. 15 by a line below the top of the barrier ($W < W_2$). These electrons (recall that electrons are quantum particles) can penetrate inside the barrier with a nonzero probability. The probability for a classical particle to do this is zero. We have already come across the tunnel effect (see Section 1.1). The probability an electron has of being able to penetrate through the barrier depends on the energy level W_3 at which the electron is residing. The probability that an electron within a metal occurs to the right of the barrier in Fig. 15, i.e. leaves the metal, depends on the height and width of the barrier, the narrower and the lower it is, the greater the probability, or to put it in another way, the larger the cross-hatched region in Fig. 15, the smaller the probability of a tunnel effect.

Consequently, an electron can leave a metal, to which an external electric field is applied either by passing *over* the potential barrier (thermionic emission) or by passing *through* the barrier (cold emission). The first mechanism prevails at moderate fields, while the second at strong fields.

A mathematical treatment has shown that in cold emission the flux of electrons i (i.e. emission current density) depends on the strength of the applied field E in the following way:

$$i = aE \exp(-b/E). \qquad (1.25)$$

Both a and b include the universal constants h, e, and m, as well as φ. Therefore i does not depend on temperature T, but depends on φ, i.e on the nature of the metal. The current density i at a given E is the same at different temperatures, but it is different for different metals.

Formula (1.25) is similar in form to formula (1.23) where temperature T is replaced by field strength E. However, the mechanism of cold emission is not at all similar to the mechanism of thermionic emission. But the current density in cold emission is as sensitive to the external field E as it is sensitive to temperature T in thermionic emission.

Calculations show that according to formula (1.25) cold emission must become noticeable at field strengths of a thousand million volts per meter (10^9 V/m), but in reality it can be observed in weaker fields. This is because of inhomogeneities on the surface which cause very strong local fields in some points.

3. Let us discuss one more method of liberating electrons from metals. Suppose a metal

is exposed to light of a certain frequency. This means that the metal surface is bombarded by light quanta (photons) carrying a certain amount of energy. Penetrating into the metal, the quanta pass their energy to electrons which can then use it to overcome the potential threshold and leave the metal. This phenomenon is called the *external photoelectric effect* or just the *photoeffect*. The electrons emitted from a metal due to the action of light are called *photoelectrons*.

The photoelectrons leave the metal with a broad spectrum of velocities, i.e. there are electrons with several different velocities in the photoelectron stream at a given frequency ν of the incident light. The velocity of a photoelectron at a given ν depends on the energy level from which the electron was displaced by the light.

The highest energy level an electron can occupy at absolute zero is the Fermi level W_F (Fig. 16).

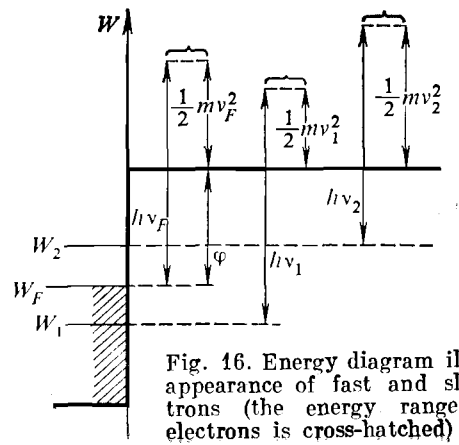


Fig. 16. Energy diagram illustrating the appearance of fast and slow photoelectrons (the energy range occupied by electrons is cross-hatched)

Every level higher than W_F is vacant at $T = 0$, and every level below W_F is occupied. Let us designate the velocity of an electron when it has been removed from level W_F by the notation v_F. Clearly (see Fig. 16) we get

$$\frac{1}{2} mv_F^2 = h\nu - \varphi, \tag{1.26}$$

where m is the electron's mass, $h\nu$ is the energy of the incident quantum, and φ is the work function we discussed while speaking of thermionic emission. Equation (1.26) is an expression of the law of the conservation of energy. The frequency ν_{cr} for which the energy $h\nu_{cr}$ of an incident quantum is equal to the work function φ is called the *red limit* of the photoeffect:

$$h\nu_{cr} = \varphi. \tag{1.27}$$

The velocity v_F is the fastest at which an electron can travel at $T = 0$. Lower velocity photoelectrons are removed from deeper levels (e.g. from level W_1 in Fig. 16).

At $T > 0$ electrons with the energy greater than W_F (e.g. with energy W_2, see Fig. 16) appear. Energies less than the work function φ are required to remove them from the metal's surface. We can write $v > v_F$ for such electrons. Although the red limit at $T = 0$ is sharp, the red limit at $T > 0$ turns to be somewhat blurred (depending on the temperature).

Note that if the metal is put into an external electric field, the red limit for the photoeffect becomes shifted in the red direction, i.e. in the direction of lower frequencies. This is because the electric field decreases the potential threshold

at the metal surface, as can be seen in Fig. 15, and therefore the work function φ, and also (as can be seen from (1.27)) the red limit ν_{cr} are lowered.

It should be emphasized in conclusion that the optics of metals cannot be based on the model of free electrons that we used so far. It can be shown that free electrons, i.e. electrons filling a potential box with a flat bottom (Fig. 6), are unable to absorb photons. However, electrons in a periodic field (Fig. 5*b*) can do so. Nonetheless, our argument remains valid since the electrons in a periodic field possess, as we shall see in the next section, a practically continuous energy spectrum, which is like free electrons do. It consists of very many closely arranged, or almost overlapping energy levels.

1.5. Electrons in the Periodic Field. Conductors and Insulators

To conclude this chapter dealing with metals, it is natural to ask why some crystals are conductors and others insulators. If an external electric field is applied, an electric current appears in some crystals while no current appears in other crystals (more exactly, the current that does appear is many thousands of millions of times less than a current in a conductor). We shall try to explain why in this section. We shall, however, have to put aside the simple model of the crystal we have used so far. Instead of considering electrons to be locked in a potential box with a *flat* bottom, we must consider the *periodic nature* of an electron's potential energy within the crys-

tal. In other words, we shall have to return from Fig. 6 to Fig. 5.

Let us proceed from a single isolated atom, and build up a crystal lattice from similar atoms

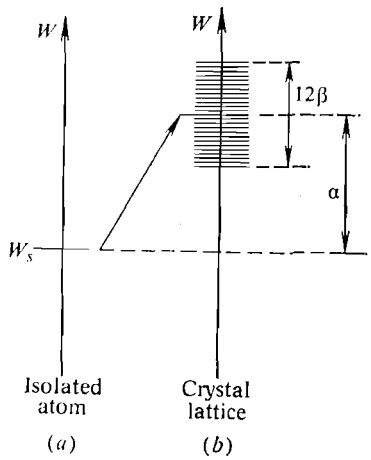


Fig. 17. Origination of an energy band: (*a*) isolated atom, (*b*) crystal lattice

as we did it in section 1.1. An atom of a chemical element differs from an atom of another element by its energy spectrum, i.e. by the distribution of energy levels on the energy scale. The energy spectrum of an atom is the set of all the values of energy that the atom may possess. The energy spectrum of an atom is a system of discrete energy levels that become more concentrated together at higher energies. In case of a monovalent atom the energy spectrum of an atom is the energy spectrum

of its valence electron. The electron can reside on any of the energy levels. If it is on the lowest level, the energy of the atom is at the minimum and the atom is said to be in its ground state. The transfer of an electron to any of the levels above this one means that the atom has been *excited.*

Now let us go on from an isolated atom to a crystal lattice. Every level W_s in an atomic energy spectrum can shift and split into n closely positioned levels to form an *energy band*; this is shown schematically in Fig. 17. Here $n = N^3$ is the number of atoms in the crystal. Like we did before (see Section 1.2), we shall imagine for the sake of certainty that the crystal is a cube, where N is the number of atoms along the edge of the cube. Consequently, the electron's energy spectrum in the crystal becomes a system of energy bands that are separated by forbidden bands or overlapping.

As in Sommerfeld's theory (the theory of free electrons, see Section 1.2) the state of each electron is characterized by a set of three quantum numbers n_x, n_y, n_z that can assume the values

$$n_x,\ n_y,\ n_z = 0,\ \pm 1,\ \pm 2,\ \ldots,\ \pm \frac{N}{2}.$$

Instead of (1.5), we shall have the following for an electron's velocity

$$\begin{aligned} v_x &= \frac{4\pi a}{h}\,\beta \sin \frac{2\pi}{N}\, n_x, \\ v_y &= \frac{4\pi a}{h}\,\beta \sin \frac{2\pi}{N}\, n_y, \\ v_z &= \frac{4\pi a}{h}\,\beta \sin \frac{2\pi}{N}\, n_z. \end{aligned} \qquad (1.28)$$

Instead of (1.6), we shall obtain the following for an electron's energy W at the lowest energy band

$$W = W_s + \alpha - 2\beta \left(\cos \frac{2\pi}{N} n_x + \cos \frac{2\pi}{N} n_y + \cos \frac{2\pi}{N} n_z \right). \quad (1.29)$$

Here W_s is the lowest energy level of an isolated atom, α and β are parameters ($\alpha > 0$, $\beta > 0$), and a, as before, is the lattice constant (the distance between adjacent atoms). The upper limit W_{max} of the band occurs for values of the quantum numbers n_x, n_y, n_z at which the cosines on the right-hand side of (1.29) are equal to -1. When each of the cosines equals $+1$, we get the lower limit W_{min} of the band. Therefore, according to (1.29)

the upper limit of the band is

$$W_{max} = W_s + \alpha + 6\beta,$$

the lower limit of the band is

$$W_{min} = W_s + \alpha - 6\beta,$$

the width of the band is

$$W_{max} - W_{min} = 12\beta.$$

The only thing left is to distribute all electrons in the collection including the valence electrons of all atoms, among the levels of the band keeping the Pauli principle in mind. This principle requires that a given set of three quantum numbers can only belong to two electrons with opposite spins. That is why the band has a certain 'capacity' with respect to electrons, i.e.

it can only accomodate a limited number of electrons.

Now we can come back to the question we raised at the beginning of the section, as to why some crystals are conductors and some insulators. From the viewpoint of the classical theory, the difference between a conductor and an insulator reduces to the presence or absence in the crystals of free electrons, i.e. electrons capable of travelling throughout the crystal without hindrance. From the viewpoint of the quantum theory, every electron in the crystal (see Section 1.1) is free in this sense. Although it gives electrons a certain freedom of motion, the quantum theory imposes other constraints on their movement. The difference between a conductor and an insulator is related to these new constraints.

We should note that the ability of electrons to travel is a necessary, but not a sufficient condition for the lattice to be a conductor. Indeed, if the electrons travel randomly in all directions, there is no electric current. For conduction to appear, it is necessary that the external electric field regulates the random motion to a certain degree, causing the requisite asymmetry in the velocity distribution of the electrons. This second condition is a necessary complement to the first one. In the classical theory, the first of these two conditions was the prerequisite of conduction, while in the quantum theory the second condition is prerequisite. In fact, the laws governing the motion of electrons are such that the symmetry of the velocity distribution of the electrons cannot be disturbed for every crystal, i.e. can be disturbed for some crystals but not for others. For

this reason not every crystal is a conductor, though every crystal contains electrons that travel freely over the lattice. Let us discuss the origin of the new restrictive law that is inherent in the quantum theory.

Let us return to the energy spectrum of the crystal and consider the lowest energy band. At absolute zero the electrons are distributed among the levels of this band as densely as possible, starting with the lowest levels of the band, which correspond to the lowest total energy of the system. But the band may be either partially or completely filled with electrons. These two cases are different. Let us consider each case separately.

(1) Imagine that the total number of electrons is less than the number of places within the band. At absolute zero they will thus be distributed at the bottom of the band with the upper part of the band unoccupied. This is shown in Fig. 18*a* where the energy W and one of the velocity components, v_x, are shown as functions of the respective quantum number n_x. The range of n_x that is filled with electrons is double-hatched. We see that the v_x velocity distribution of the electrons corresponding to the densest packing is symmetrical. Thus the current in the crystal is zero because for each electron with a velocity v_x there is an electron with a velocity $-v_x$. Note that we can put v_y or v_z into Fig. 18 instead of v_x, and n_y and n_z, respectively, instead of n_x.

When an electric field is applied, the electrons are redistributed among the levels so that the number of electrons with positive n_x exceeds the number of electrons with negative n_x (or vice versa). This implies that there are more electrons

travelling in one direction than in the opposite one. We shall have to replace Fig. 18*a* with Fig. 18*b*, which shows the distribution of the electrons in the presence of an electric field. The role of the field is to liberate some of the occupied levels

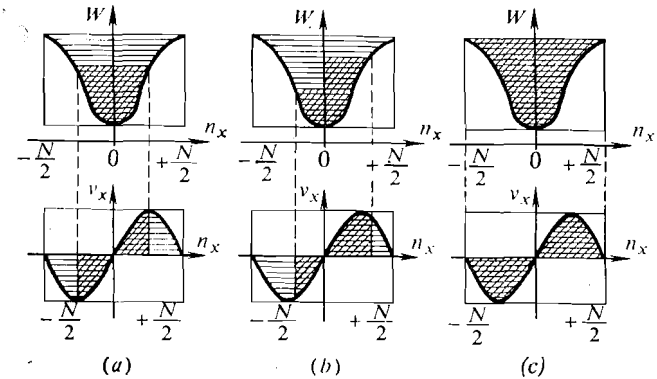


Fig. 18. Distribution of electrons by states n_x: (*a*) band partially filled with electrons (in the absence of an external field), (*b*) band partially filled with electrons (in the presence of a field), (*c*) completely filled band

and to fill some of the vacant ones thereby introducing some asymmetry into the velocity distribution of electrons, as it is seen by comparing Figs. 18*a* and 18*b*.

This shift of electrons between levels is only possible, however, when there are vacant levels in the vicinity of the occupied ones. This is possible in this case since the band is only partially full of electrons.

(2) Now imagine that the number of electrons in the lattice is equal to the number of levels in

the band. This is the case when the densest packing of the electrons results in every level in the band being occupied. This case is illustrated in Fig. 18*c* where the whole range of n_x is double-hatched. It is also essential in the case under consideration to know whether the filled band is adjacent to the vacant band lying above it or separated from it by a forbidden band.

(a) If the band occupied by electrons is in contact with or overlaps the vacant band, then the vacant band can be considered to be an extension of the filled one. This is the case when the velocity redistribution of the electrons can be carried out at the expense of this overlying band. The two contacting or overlapping bands, the lower of which being occupied by the electrons and the upper being vacant, can in reality be considered to be a single partially occupied band.

(b) If the band occupied by the electrons is separated from the vacant band above it by a forbidden zone, then a velocity redistribution of the electrons within the occupied band becomes impossible. The only thing possible is for the electrons to exchange places, but this will not bring about the desired asymmetry in the distribution of the electrons (see Fig. 18*c*). An external field will then be incapable of producing any effect on the motion of electrons within the occupied band, and therefore the lattice will be insensitive to the influence of the external field.

Consequently, we come to the following criteria for distinguishing between a conductor and an insulator (see Fig. 18):

(1) A material is a conductor if the band containing the electrons has a number of unoccupied levels, or if it is in contact with or overlaps the vacant band above it.

(2) A material is an insulator if the band is completely filled with the electrons and is separated from the vacant band above it by a forbidden zone.

The forbidden zone between the bands can be overcome as the result of external effects, and a certain number of electrons can be removed from the underlying occupied band to appear in the vacant band. This results in an insulator acquiring the ability to conduct electricity and the topic will be considered in more detail in later chapters that deal with nonmetallic crystals.

Chapter 2

Electrons in Semiconductors

2.1. 'Order' and 'Disorder' in Crystals

In this chapter we shall discuss the solids that are called *semiconductors*. As we already know (see 'Introduction', section I.2), they are a remarkable group of solids that are intermediate in their properties between metals and dielectrics. Metals can be regarded as an extreme case of semiconductors with dielectrics at the other extreme.

Note that all the macroscopic (i.e. directly observable) properties of crystals (including semiconductors) can be put into two classes. The first contains all the properties that are determined by the periodic structure of the crystal lattice and for which deviations from the regularity of the structure (called *defects*) which are inevitably present in any real lattice, do not play a significant role. These properties are called *structure-stable*. The other class includes all the properties that are determined by the local irregularities in the periodic structure of the lattice. The defects in this case are very important. The properties belonging to this second class are commonly called *structure-sensitive*.

Theoretical interpretations of structure-stable and structure-sensitive properties require different approaches. In the first case we can start from the theory of ideal crystal lattice, while

in the second case we have to deal with concepts of real crystals. The theory of ideal crystals can not handle any structure-sensitive property. The theory of semiconductors does not deal with ideal crystals, but with real ones.

A real crystal differs from an ideal one in that it has defects, i.e. local irregularities in the periodic structure of the lattice. We should distinguish between the macroscopic and microscopic defects which are inherent in any real lattice. A *macroscopic* defect is a disturbance in the periodic structure embracing a region whose dimensions are substantially greater than the lattice constant. These include fractures, pores, and macroscopic inclusions. We are not going to discuss defects of this kind here. A *microscopic* defect is an irregularity whose dimensions are of the same order as the dimensions of an individual crystallographic cell. Let us list the main microscopic defects:

(1) an empty lattice point (a *vacancy*), the result of removing an atom or ion from an ideal lattice (Fig. 19*a*);

(2) an atom or ion of the lattice in an interstice (Fig. 19*b*);

(3) a foreign atom in an interstice (Fig. 19*c*);

(4) a foreign atom at a lattice point (Fig. 19*d*).

The last two types of defects can be called *chemical* defects. This is an 'impurity' in the narrow sense of this word, an irregularity in the chemical composition of the crystal. Chemical defects come into a crystal from without: i.e. they appear due to its processing.

The first two types of defects can be called *structural* defects. In case of a monatomic lattice

(i.e. a lattice composed of atoms of only one kind), such defects do not distort the chemical composition of the crystal. If the lattice is multicomponent, they can distort the crystal's stoichiometry, i.e. change the relative quantities

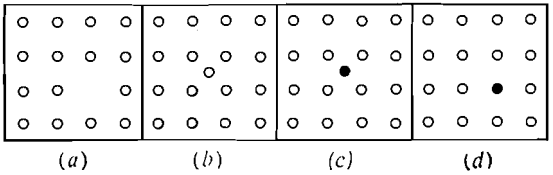


Fig. 19. Microdefects in a semiconductor crystal (a blank circle is a proper atom of the lattice, a filled circle is a foreign atom): (*a*) vacancy, (*b*) proper atom of the lattice in an interstice, (*c*) foreign atom in an interstice, (*d*) foreign atom substituting a proper atom of the lattice

of the component atoms that is a characteristic of an ideal lattice*. Structural defects can be introduced from without and can appear when the lattice is heated.

Note that each microdefect deforms somewhat the lattice around it. Strictly speaking, the whole region within which the lattice is deformed

* For instance, if the crystal is of cuprous oxide Cu_2O, then there should be 200 atoms of copper for every 100 oxygen atoms, in which case the stoichiometric relationship is said to be correct. But if there are 205 or 195 copper atoms for every 100 oxygen atoms, then the stoichiometry of the crystal is said to be distorted, in the former case towards an excess of copper (or deficiency of oxygen), in the latter case towards a hyperstoichiometric prevalence of oxygen (or deficiency of copper).

should be considered to be a defect. Irrespective of their actual nature, microdefects have a number of common properties.

1. Microdefects have a certain mobility that increases with temperature (i.e. they have an activation energy). In fact, as a defect moves in any direction the system's energy changes periodically, i.e. the energy passes through alternate minima and maxima. Defect movement, therefore, is connected with overcoming energy barriers whose levels are determined by the nature of the defect, the lattice structure, and direction of travel. Defect movement requires, therefore, activation energy that is generally different for different crystallographic directions. Defects can only be assumed to be fixed within a crystal at moderate temperatures.

2. Another common feature of microdefects is that they interact when they approach each other. The system's energy generally depends on the relative distribution of the defects, which is the evidence of the interaction between them. Defects can attract or repel each other. For instance, in an ionic lattice MR, where M is a metal and R is a nonmetal, and which consists of the ions M^+ and R^-, nonmetallic vacancies repel each other, but attract metal vacancies.

3. Having met, defects can produce compounds by forming groups that should be considered new defects and which generally possess other features. For instance, when a nonmetallic and a metallic vacancy join in an MR lattice, they produce a different sort of formation that has properties which are different from those its components have when considered separately.

The 'reactions' between defects, can, like ordinary reactions, be both exothermic and endothermic, and can occur both with and without activation, depending on the nature of the reacting defects.

4. Defects of each kind, while 'reacting' with other defects, can thereby appear and disappear. A mean lifetime can thus be attributed to them in equilibrium conditions. They can in addition be absorbed or generated by the lattice itself. An example of this is movement of an atom (or ion) from a lattice point into an interstice which results in the appearance of two kinds of defects: interstitial atoms (or ions) and vacancies. Another example is, vice versa, the recombination of an interstitial atom (or ion) and a vacancy to result in the disappearance of two defects which, as it were, 'absorb' each other.

The presence of defects in a crystal testifies that the order characteristic of an ideal crystal is disturbed in a real crystal. Individual crystal cells in such a crystal are 'spoiled'. The ratio of the number of spoiled cells to the overall number of cells is what can be called the *degree of disorder* in the crystal.

The degree of disorder is determined primarily by the 'biography' (history) of the crystal, i.e. it depends on how the crystal was produced and what treatment it was subjected to. The degree of disorder also generally depends on temperature. Heating actually tends to remove lattice atoms (ions) from their lattice points, pushing them into interstices, and therefore results in more disorder in the lattice. Thus,

disorder in the lattice can be either 'biographical' or thermal in origin. Biographical disorder is irreversible and remains at absolute zero, i.e. it is, as it were, a congenital disorder of the lattice. Disorder thermal in origin is superimposed on biographical disorder.

In some cases thermal disorder prevails over biographical disorder, it then being possible to neglect biographical defects compared with defects of thermal origin. But in other cases it is possible to neglect thermal disorder compared with biographical disorder and this happens to crystals that have many defects and at moderate temperatures. This is the case when we deal with semiconductors.

Consequently, a crystal lattice with its defects is an indivisible system whose properties are governed by two competing factors: the factor of order and the factor of disorder. The order controls all the stable properties, and the disorder all the structure-sensitive properties of the crystal.

2.2. Free Electrons and Free Holes in Semiconductors

What are free electrons and free holes? How do they behave within semiconductors? These are the questions we shall try to answer in this section.

Imagine the crystal lattice of a semiconductor. Suppose that an extra electron, in addition to the normal set, is assigned to an atom (or ion) of the crystal lattice. This redundant electron is called a *free* electron since it is capable of

travelling from one atom (ion) to a neighbouring atom (ion) and therefore of wandering throughout the lattice. However, it only needs a potential difference to be applied to the crystal for a predominant direction in the chaotic movement of the

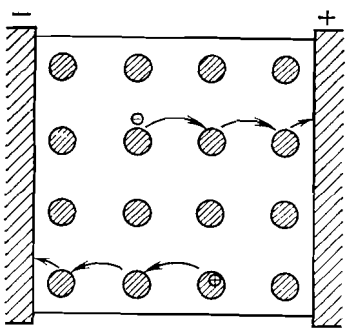


Fig. 20. Electronic and hole currents (arrows show the direction of the motion of an electron and a hole)

electrons to become apparent. On the average, they all begin travelling in one direction (from the negative to the positive electrode), and this is illustrated diagrammatically in Fig. 20. Current appears in the semiconductor. We say that there is electronic conduction in the semiconductor since the current is due to electrons travelling. A semiconductor containing a concentration of free electrons i.e. possessing electronic conduction, is called an *electronic semiconductor* or *n*-type semiconductor (*n* is for 'negative' because negative charges are the current carriers).

Now suppose we do not add an extra electron to an atom (ion) of the lattice, but instead remove an electron from the normal set. We obtain an atom (or ion) that differs from all other atoms (ions) by its lack of an electron. This is called a *hole*. We call it free because it can travel around the crystal without hindrance, wandering from atom (ion) to atom (ion) of the same element. When an external electric field is applied, a *hole current* appears in the semiconductor (see Fig. 20). We say that there is *hole conduction* in the semiconductor, and the semiconductor itself is called a *hole*, or *p*-type, *semiconductor* (*p* is for 'positive' since, as it were, positive charges are the current carriers).

Take for example zinc oxide ZnO whose lattice we shall consider to be an ionic lattice built up of the Zn^{++} and O^{--} ions. A free electron corresponds to the Zn^{+} state, while a free hole corresponds to the O^{-} state. As another example, consider the lattice of cuprous oxide Cu_2O, which is composed of the Cu^{+} and O^{--} ions. A free electron corresponds to the Cu state, while a free hole corresponds to the Cu^{++} state, both wandering among the regular Cu^{+} ions. In a monatomic lattice, that of germanium for instance, which is built up of the neutral Ge atoms, a free electron or a free hole implies the presence of a Ge^{-} ion or a Ge^{+} ion, respectively, among the neutral Ge atoms.

The difference between electronic and hole currents is illustrated by the following analogy. Suppose we have a school classroom with only one schoolboy in the first row, and all the places in this row are vacant. The boy changes the

seats, say, from left to right and therefore moves within the row. This is an analogy of an electronic current flowing through a semiconductor. Now suppose all but one of the places in the first row are occupied by schoolboys, and the boy on the left of the unoccupied seat moves to take it. The vacant place thus moves along the row from right to left. This is an analogy of a hole current.

Note that a free electron (like a free hole) produces an electric field around it. The field becomes less intense with the distance from the source. If the semiconductor's lattice is ionic, then the ions surrounding the free electron (or hole) are within the field, and it tends to displace them from their equilibrium positions. Ions of one sign are attracted to the electron (or hole) while those with other sign are repelled from the electron (or hole). The crystal lattice around the electron (or hole) is thus somewhat deformed, the deformation however decreases rapidly with the distance. Thus it only covers the immediate vicinity of the electron (or hole). This formation (the electron or hole and the surrounding deformation of the lattice) is called a *polaron*. A polaron can travel across the lattice. A moving polaron is a moving electron (or hole) pulling behind it the deformation of the lattice it produces. Naturally, it is more difficult for a polaron to move across a crystal than it is for just an electron (or hole) if we neglect the distortion.

A semiconductor can under certain conditions have at the same time both free electrons and free holes in any concentrations. A semiconductor is said to be *intrinsic* if the concentrations

are the same and both electronic and hole currents appear if an external field is applied. We then say that there is *mixed conduction* in the semiconductor.

What is the source of these free electrons and holes? This is where the structural defects or impurity atoms in a crystal come into play.

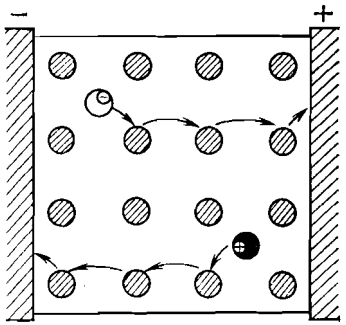


Fig. 21. Origination of free electrons and holes (arrows show the motion of an electron and a hole). A cross-hatched circle is a proper atom of the lattice, a blank circle is a donor impurity atom, a filled circle is an acceptor impurity atom

They can be divided into two groups according to the role they play in the electronic set-up of the crystal and we distinguish between *donor* and *acceptor* defects.

Donor defects are the suppliers of free electrons. An electron escaping from such a defect starts travelling around the lattice. It leaves a hole that remains localized (Fig. 21). Thus donor defects are the reservoir from which the semiconductor draws free electrons. There are

as many free electrons in the semiconductor as there are ionized donor defects.

Clearly, the concentration of free electrons, and thus the electronic conduction of the semiconductor, increases when the concentration of a donor impurity (or donor defects) is increased. The concentration of free electrons also increases with the rise of temperature, because more impurity atoms (defects) are ionized at higher temperatures.

Donor defects play the opposite role with respect to free holes. They do not supply free holes but rather trap them. An electron from a donor defect can recombine with an approaching free hole, as a result of which the free hole disappears and a localized hole appears in the defect. The larger the concentration of donor defects the less the concentration of free holes becomes. That is why adding donor defects to hole semiconductors reduces rather than raises their conductivity.

Let us consider now acceptor defects. They play the role that is opposite to that of donor defects. Acceptor defects are traps for free electrons and suppliers of free holes. They catch electrons, and therefore the introduction of acceptor defects brings about a decrease in electronic conduction and an increase in hole conduction.

Hyperstoichiometric atoms of zinc in a crystal of zinc oxide ZnO, which is a typical *n*-type semiconductor, are an example of donor defects. An example of acceptor defects are vacancies of copper atoms in a crystal of cuprous oxide Cu_2O, which is a *p*-type semiconductor.

A crystal of germanium can be made either to be an *n*-type or a *p*-type semiconductor depending on whether acceptor or donor impurities are added. Let us have a look at Mendeleev's periodic table of elements. Germanium (Ge) is in group 4. This means that a germanium atom is tetravalent, i.e. it has four valence electrons. Suppose we change the crystal lattice of germanium by adding as an impurity some atoms of a fifth-group element, for instance, arsenic (As) or antimony (Sb), so that they substitute germanium atoms. Unlike germanium atoms, these atoms possess five valence electrons not four, i.e. one electron more than germanium. These impurities in a germanium crystal will act as donors, so a germanium crystal with an arsenic or antimony impurity is an *n*-type semiconductor.

By contrast, if we include third-group atoms in a germanium crystal (for instance, gallium (Ga) or indium (In) atoms which possess three valence electrons, i.e. one electron less than germanium atoms) then we have added an acceptor impurity. A crystal of germanium with an impurity of gallium or indium is a *p*-type semiconductor.

Note that a semiconductor can draw free electrons and holes from the lattice itself at high temperatures or low defect concentrations. By moving an electron from an atom (or ion) of the lattice to another atom (or ion) of the lattice, we get an electron-hole pair. Having pulled the electron and the hole far enough apart so that any electrostatic interaction between them can be neglected, we get a free electron and a free hole, independent of each other.

2.3. 'Energy Bands' and 'Localized Levels'

Until now we have described things in terms of models, but it is more convenient to describe the electronic processes in semiconductors in terms of energy.

The energy spectrum of an electron in an ideal crystal lattice, as we have already mentioned, is a system of energy bands that either are separated by forbidden zones or overlap. A finite number of electrostatic states corresponds to each band, so that each band (recall the Pauli principle) has a certain capacity with respect to the electrons, i.e. it can accommodate only a finite number of them.

At absolute zero, when the system energy is at the minimum, all the lower bands are filled with electrons to capacity, while all the bands above them are totally vacant. We shall discuss the uppermost of the occupied bands, which we shall call the *valence band*, and the lowermost of the unoccupied bands, which we shall call the *conduction band*. When the width of the forbidden zone between these two bands is zero, it is a metal. If the width is not zero, it is a nonmetallic crystal (a semiconductor or insulator). The appearance of an electron in the conduction band means a free electron has emerged. The appearance of a hole in the valence band means a free hole has emerged.

Let us consider an electron in the conduction band. Suppose its state is characterized by a set of three integer quantum numbers n_x, n_y, n_z and a spin that can have either of two values. The energy W of the electron (as in the theory of

metals, Chapter 1) can be presented in the simplest case as (1.29). Let us write this formula once again:

$$W = W_s + \alpha - 2\beta \left(\cos \frac{2\pi}{N} n_x + \cos \frac{2\pi}{N} n_y + \cos \frac{2\pi}{N} n_z\right), \quad (2.1)$$

where N is the number of atoms along the edge of the cube (we assume, as in the preceding chapter, that the crystal is cubic). Near the bottom of the conduction band, where the absolute values of the numbers n_x, n_y, n_z are small, each of the cosines on the right-hand side of (2.1) is close to unity, and so we can use approximate formulas that are the first expansion terms:

$$\begin{aligned} \cos \frac{2\pi}{N} n_x &= 1 - \frac{1}{2}\left(\frac{2\pi}{N} n_x\right)^2, \\ \cos \frac{2\pi}{N} n_y &= 1 - \frac{1}{2}\left(\frac{2\pi}{N} n_y\right)^2, \\ \cos \frac{2\pi}{N} n_z &= 1 - \frac{1}{2}\left(\frac{2\pi}{N} n_z\right)^2. \end{aligned} \quad (2.2)$$

Substituting (2.2) into (2.1), we obtain

$$W = \frac{4\pi^2}{N^2} \beta (n_x^2 + n_y^2 + n_z^2) + W_0, \quad (2.3)$$

where W_0 is a constant, $W_0 = W_s + \alpha - 6\beta$. We come to the familiar formula (1.8) from Section 1.2 for a free electron gas

$$W = \frac{h^2}{2ma^2N^2} (n_x^2 + n_y^2 + n_z^2). \quad (2.4)$$

Compare (2.3) with (2.4) and note that these formulas are the same correct to the additive

constant W_0 if m in (2.4) is replaced with

$$m^* = \frac{h^2}{8a^2\pi^2\beta},$$

i.e. if the electron is ascribed an *effective mass* m^* instead of its true mass m. It is clear that our electron in the semiconductor conduction band behaves like a free electron of the Sommerfeld gas, if an effective mass m^* is ascribed to it instead of its true mass m. As for the additive constant W_0 that distinguishes (2.4) from (2.3), its presence is inessential, because it only means a change in the reference point for energy.

At temperatures above absolute zero some of the electrons are moved from the valence band into the conduction band. Let us call the concentrations of free electrons and free holes n and p respectively. We shall determine n and p for the case when the equilibrium in the semiconductor is established. The probability for an electron to be driven from the valence band into the conduction band is proportional to

$$\exp\left(-\frac{W_C - W_V}{kT}\right), \tag{2.5}$$

where $W_C - W_V$ is the width of the forbidden zone between the bands (Fig. 22*a*). The probability that a free electron travelling in the conduction band will recombine with a free hole travelling in the valence band is proportional to np. Let us note that in our case

$$p = n,$$

i.e. the number of electrons moved to the conduction band equals the number of holes left in the valence band. Thus, under equilibrium conditions

$$n = C \exp\left(-\frac{W_C - W_V}{2kT}\right), \tag{2.6}$$

where C is a proportionality factor whose expression is inessential here. Formula (2.6) refers to an ideal lattice, i.e. a semiconductor in which

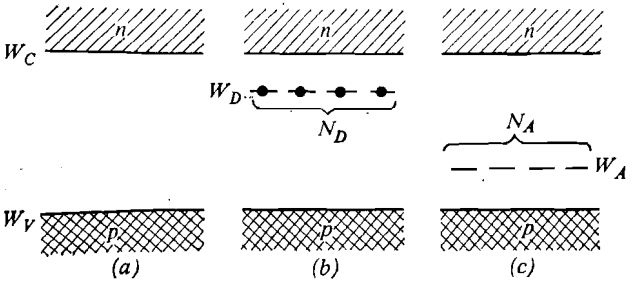


Fig. 22. Energy diagrams for semiconductors. W_C is the bottom of the conduction band, W_V is the ceiling of the valence band, W_D is a donor localized level, W_A is an acceptor localized level

impurities and structural defects can be neglected. This is called an *intrinsic semiconductor* (*i*-type semiconductor).

Now let us deal with a real lattice, i.e. a lattice with either donor or acceptor defects, or both at the same time. The presence of defects influences the energy spectrum, and *localized levels* appear in the energy spectrum along with the energy bands. The energy bands can be considered to be spread over the whole crystal in the sense that an electron belonging to a band can occur with equal probability anywhere

within the crystal. An electron belonging to a localized level is localized to a fairly limited space around a defect.

We should distinguish between two classes of localized levels: *donor* and *acceptor* levels. A donor level is a level occupied by an electron at absolute zero. By definition, an acceptor level is a level vacant at absolute zero. With a rise in temperature, acceptor levels become partially filled with electrons while donor levels lose some of their electrons. Donor levels can release their electrons to the conduction band, thus enriching the pool of free electrons. On the other hand, an electron from a donor level can recombine with a free hole, thus reducing the pool of free holes. At the same time, an acceptor localized level can accommodate an electron from the conduction band, thus reducing the pool of free electrons in the crystal. On the other hand, an electron from the valence band can be drawn to a vacant acceptor level leaving a hole instead. The crystal can be enriched with holes in this manner. Clearly, the roles of acceptors and donors are quite different and in a sense oppose each other.

We shall determine now the concentration n of free electrons in an electronic semiconductor with donor defects of only one kind (see Fig. 22*b*). Suppose the localized levels of these defects are not too far below (at a distance $W_C - W_D$) the conduction band and suppose N_D is the concentration of defects. At a steady electronic equilibrium, the number of electrons jumping due to thermal motion from the localized levels to the conduction band per unit time per unit volume should be equal to the number of reverse transitions of electrons from the conduction band to the localized levels, i.e. to the number of recombinations between free electrons and the bound holes that stay at the localized levels when their electrons leave them. Thus we get

$$np = C_D N_D \exp\left(-\frac{W_C - W_D}{kT}\right), \qquad (2.7)$$

where $W_C - W_D$ is the depth of the donor levels below the conduction band, C_D is a proportionality factor, and p is the concentration of vacant localized levels (holes). Since $p = n$, we get

$$n = \sqrt{C_D N_D} \exp\left(-\frac{W_C - W_D}{2kT}\right). \qquad (2.8)$$

Obviously, the concentration n of free electrons increases as more donor impurities are added to the semiconductor (N_D increases), and as the temperature T rises.

Now let us tackle a p-type semiconductor (Fig. 22*c*). Let us determine concentration p of free holes in the valence band at electronic equilibrium. Just as in the determination of the concentration n of free electrons in an n-type semiconductor, we shall get

$$p = \sqrt{C_A N_A} \exp\left(-\frac{W_A - W_V}{2kT}\right), \qquad (2.9)$$

where N_A is the concentration of acceptors in the crystal, $W_A - W_V$ is the distance from acceptor localized levels to the ceiling of the valence band, and C_A is a proportionality factor.

Formulas (2.8) and (2.9) correspond to the simplest case when an n-type semiconductor

contains only one kind of donors, and a p-type semiconductor contains only one kind of acceptors. In reality the same crystal can possess several kinds of donors and several kinds of acceptors, and then the expressions for n and p are more complicated.

2.4. Semiconductor Conductivity

Electrical conductivity or specific conductance is one of the most interesting and important characteristics of a semiconductor. As before, we are going to designate it by the letter σ. Conductivity can be expressed as

$$\sigma = eun \tag{2.10}$$

where e is the absolute value of the electron charge; u is the *mobility* of the current carriers (electrons or holes), i.e. the average drift velocity of the current carriers in the direction of electric field $\mathbf{E}$ (at $E = 1$ V/cm); and n is the concentration of carriers in the semiconductor. Formula (2.10) is quite general in nature and the conductivity of a metal can also be expressed this way.

A semiconductor differs from a metal above all by the fact that the concentration of carriers in a semiconductor is several hundred thousand times less than that in a metal. Metals contain 10^{23} (or more) carriers per cubic centimeter. Besides, n in semiconductors is, as a rule, very sensitive to temperature, increasing hundreds of thousands of times as the temperature rises moderately. And, finally, n in semiconductors is usually very sensitive to impurities (to their

nature and concentration), while n in metals is a constant that depends solely on the nature of the metal and does not depend at all on the temperature or the impurity concentration.

The mobility u that appears in (2.10) poorly depends on temperature. This poor dependence of u on T, compared with the strong dependence of n on T, allows us in many cases to assume that u in formula (2.10) is constant and does not depend on T at all.

Let us see how σ in an electronic semiconductor depends on T and N_D. We substitute (2.8) into (2.10) to get

$$\sigma = \sigma_0 \exp\left(-\frac{W_C - W_D}{2kT}\right), \tag{2.11}$$

where

$$\sigma_0 = eu\sqrt{C_D N_D}. \tag{2.12}$$

Taking logarithms of both sides of (2.11) we obtain

$$\ln\sigma = \ln\sigma_0 - \frac{W_C - W_D}{2kT}. \tag{2.13}$$

By plotting the logarithm of conductivity along the ordinate, and the reciprocal of the absolute temperature along the abscissa, equation (2.13) is shown by a straight line (Fig. 23). The angle α between this line and the abscissa rises as the localized donor level that provides the conductivity sinks below the conduction band:

$$\tan\alpha = \frac{W_C - W_D}{kT}.$$

Now let us take a hole semiconductor. Assuming $p = n$ in (2.10) and substituting (2.9)

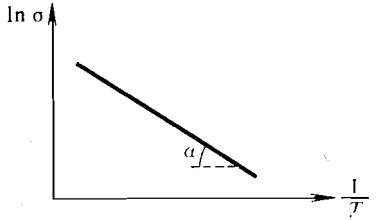


Fig. 23. Semiconductor conductivity as a function of temperature

into (2.10), we obtain an expression similar to (2.13), i.e.

$$\ln \sigma = \ln \sigma_0 - \frac{W_A - W_V}{2kT}, \qquad (2.14)$$

where

$$\sigma_0 = eu\sqrt{C_A N_A}.$$

In this case the straight line in Fig. 23 gets steeper the higher above the valence band the localized acceptor level W_A is, i.e.

$$\tan \alpha = \frac{W_A - W_V}{kT}.$$

Now let us take, at last, an intrinsic semiconductor (i-type semiconductor). This type of material conducts in both ways and so instead of (2.10) we have

$$\sigma = e\,(u_n n + u_p p), \qquad (2.15)$$

where u_n and u_p are the mobilities of electrons and holes respectively, n and p are the concentrations of free electrons and free holes in the conduction band and the valence band respectively. If all the electrons in the conduction band are borrowed from the valence band, formula (2.15) transforms into (2.10), where u is the

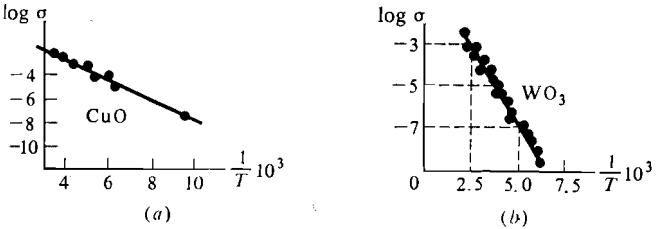


Fig. 24. Conductivity as a function of temperature: (a) for a crystal of CuO, (b) for a crystal of WO_3

sum of mobilities for carriers of two types:

$$u = u_n + u_p.$$

In this case the slope of the line in Fig. 23 is the more, the wider the forbidden zone between the bands:

$$\tan \alpha = \frac{W_C - W_V}{kT}.$$

Some typical experimental curves are shown in Fig. 24. Obviously, they are in good agreement with theory and conductivity increases with the rise of temperature very sharply. In Fig. 24b, a doubling of temperature from $1/T = 5 \times 10^{-3}$ to $1/T = 2.5 \times 10^{-3}$ raises the conductivity about 10,000 times (from $\sigma = 10^{-7}$ to $\sigma = 10^{-3}\ \Omega^{-1}\cdot\mathrm{cm}^{-1} = 10^{-1}\ \mathrm{S}\cdot\mathrm{m}^{-1}$).

The straight line in Fig. 23 corresponds to

the simplest cases. This line sometimes has a bend, like in Figs. 25*a*, 25*b*. These cases also comply with the simple theory if the concentrations of acceptor and donor impurities, as well

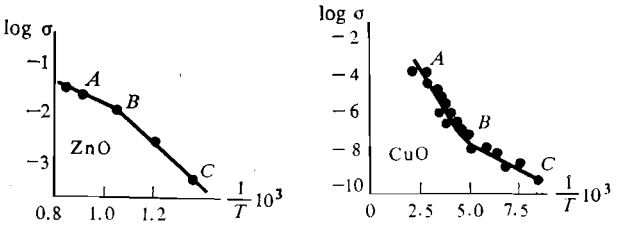


Fig. 25. Bends of the conductivity versus temperature diagrams: (*a*) for a crystal of ZnO, (*b*) for a crystal of CuO

as the respective localized levels in the forbidden zone between the energy bands, are selected properly.

Let us consider Fig. 26. This shows a family of curves for the same semiconductor but with different concentrations of an introduced impurity. The impurity concentration corresponding to the A_3B_3C curve is greater than that corresponding to the A_2B_2C curve, which in turn is greater than that corresponding to the A_1B_1C curve. Obviously, the conductivity of a crystal at moderate temperatures is the greater, the greater the concentration of impurity in it. Conduction in this case is due to electron transitions between localized levels and a band. This can be either electronic or hole conduction. At high temperatures the conductivity is the same for all crystals (i.e. it does not depend on

impurity concentration) since the conduction is due to electron transitions from band to band.

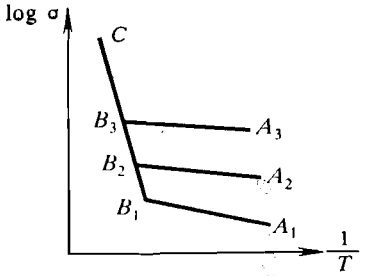
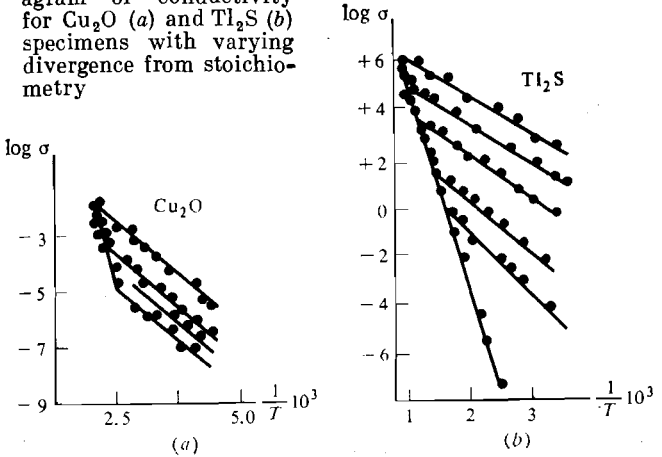


Fig. 26. Temperature diagram of conductivity for specimens with different impurity concentrations

This is a case of mixed conduction. To illustrate the point, Fig. 27 shows experimental curves

Fig. 27. Temperature diagram of conductivity for Cu_2O (*a*) and Tl_2S (*b*) specimens with varying divergence from stoichiometry



for cuprous oxide Cu_2O and thallium sulphide Tl_2S. Fig. 27*a* gives curves for crystals of Cu_2O (a hole semiconductor) with different concentrations of hyperstoichiometric oxygen, while the

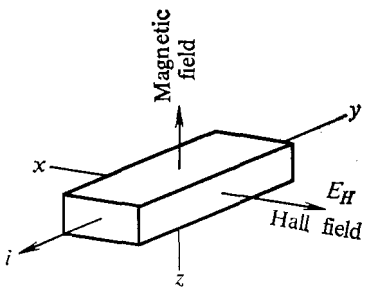


Fig. 28. Transverse field E_H caused by the Hall effect

curves in Fig. 27*b* are for Tl_2S (an electronic semiconductor) with different concentrations of hyperstoichiometric thallium.

A temperature diagram of conductivity does not allow us to infer the nature of conduction, i.e. whether there is electronic or hole conduction. To do this we need some additional measurements, and those pertaining to the *Hall effect* are suitable.

To understand the Hall effect suppose there is an electric current of density i flowing along the y-axis of a specimen (Fig. 28). Then suppose that the specimen is placed in a magnetic field with a strength H directed along the z-axis, i.e. perpendicular to the current flow. A potential difference arises between the sides of the specimen, i.e. an electric field is set up directed

along the x-axis and we shall designate its strength E_H. This is called the *Hall field*. It is perpendicular both to the direction of the current and to the direction of the magnetic field. The appearance of an electric field due to a magnetic field is the Hall effect.

E_H, H, and i are related thus:

$$E_H = RiH, \tag{2.16}$$

where R is a proportionality constant called the *Hall coefficient* or *Hall constant*. A mathematical treatment produces the following expression for R:

$$R = \alpha \frac{pu_p^2 - nu_n^2}{(nu_n + pu_p)^2}. \tag{2.17}$$

Here α is a scalar constant that is of no interest for us. The rest of the notation is the same as before.

We have $p = 0$ for electronic semiconductors, and hence (2.17) becomes

$$R = -\frac{\alpha}{n}.$$

Meanwhile $n = 0$ for hole semiconductors, and hence (2.17) becomes

$$R = \frac{\alpha}{p}.$$

Therefore

$R > 0$ for p-type semiconductors, and

$R < 0$ for n-type semiconductors.

In the case of intrinsic semiconductors, when $n = p$, we obtain from (2.17)

$$R = \frac{\alpha}{n_i} \frac{u_p^2 - u_n^2}{(u_p + u_n)^2}, \qquad (2.18)$$

where we introduced the notation $n_i = n = p$. Since

$$u_p^2 - u_n^2 = (u_p + u_n)(u_p - u_n),$$

we can rewrite (2.18) as

$$R = \frac{\alpha}{n_i} \frac{u_p - u_n}{u_p + u_n}. \qquad (2.19)$$

If the electrons and holes are equally mobile, i.e. $u_p = u_n$, then it follows from (2.19) that the Hall constant becomes zero ($R = 0$).

Note in conclusion that along with the determination of the sign of the Hall constant there is another way of establishing the sign of the charges that carry the current in semiconductors. It consists in the study of the *thermal electromotive force* (*thermo-emf*) that appears in any semiconductor if there is a drop of temperature within it. Suppose a semiconductor is a rod one end of which is heated and the other end of which is cooled. Since the concentration of carriers in a semiconductor is usually very sensitive to temperature, the concentration of carriers at the hot end of the specimen will be greater than that at the cool end. There will thus be diffusion, i.e. carriers from the hot end will travel to the cool one to try to equalize the concentrations. If the current carriers are electrons, then the cool end will receive an excess of electrons, and so will become negatively

charged, while the hot end will be left without some of its electrons and so will become positively charged. But if the current carriers are holes, then by contrast the cool end will become positively charged and the hot end negatively charged. The data we obtain from the sign in the Hall effect usually coincide with the data from the sign obtained using the thermo-emf technique.

2.5. Electrons and Quanta

What will happen in a semiconductor if it is exposed to visible or ultraviolet light or X-rays? We shall try to answer this question in this section.

When speaking of the exposure of a semiconductor, we mean the bombardment of a specimen with light quanta with a given frequency ν or, in other words, by quanta with a mass m, because ν and m are related, i.e.

$$m = \frac{h\nu}{c^2}, \qquad (2.20)$$

where h and c are universal constants, h being Planck's constant ($h = 6.62 \times 10^{-27}$ erg·s), and c being the speed of light ($c = 3 \times 10^8$ m/s). Therefore the smaller ν is, i.e. the 'redder' the light, the lower m is, i.e. the 'lighter' the quantum.

Not every incident quantum is absorbed by a crystal. The absorption spectrum is a system of bands that more or less overlap. Some bands are structure-stable, i.e. they are not sensitive to the way a specimen has been processed (*in-*

trinsic absorption), other bands are structure-sensitive, i.e. they depend on the concentration and the nature of impurities the crystal contains (*impurity absorption*). The absorption of a quantum by a crystal is the evidence that the system has become excited, i.e. its energy has been increased. Usually, this excitation is related to

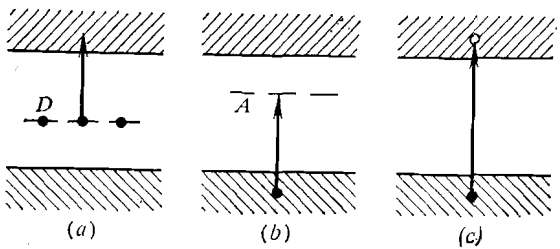


Fig. 29. Origination of: (*a*) photoelectron, (*b*) photohole, (*c*) pair: photoelectron plus photohole

a transition of an electron from a lower to a higher energy level. When a quantum is absorbed, an electron-hole pair appears. There are three possible cases:

(1) one of the partners (the electron or the hole) becomes free, and the other stays bound (localized) following the absorption of the quantum (Figs. 29*a*, 29*b*);

(2) both partners (the electron and the hole) become free and independent of each other (Fig. 29*c*);

(3) both partners stay bound to each other and behave as an entity.

To begin with, let us take the first case. Suppose the crystal contains donor impurities

(or donor structural defects). The absorption of a quantum ionizes a defect. An electron leaves a donor localized level and goes to a level in the conduction band, i.e. becomes free. This electron, which we shall call a *photoelectron*, becomes able to travel over the lattice. The hole that is left remains localized at the site the electron left.

Suppose that the crystal contains acceptor localized levels. When a quantum is absorbed, an electron leaves a level in the valence band and goes to an acceptor localized level, i.e. a free hole (a *photohole*) appears in the valence band, and a localized electron appears on an acceptor localized level.

Let us remark that the donor and acceptor localized levels responsible for the appearance of photoelectrons and photoholes may be not the same donor and acceptor levels that bring about conduction in the dark.

Now let us take the second case, when the absorption of a quantum results in the appearance of two free carriers: an electron and a hole.

While impurity absorption was the point in the preceding case, now the point is intrinsic absorption, when an electron leaves a level in the valence band (the emergence of a free hole) and goes to a level in the conduction band (the emergence of a free electron).

Let us note that the appearance in a crystal of either photoelectrons or photoholes, or both is called the *internal photoelectric effect*, as opposed to the *external photoelectric effect* which is when electrons are emitted due to radiation both from semiconductor and metal crystals (see Section 1.4).

Now let us take the third case. Suppose the absorption of a quantum results in an electron going from one atom or ion of the lattice to an adjacent atom or ion. The electron and the hole that appear remain close together and interact electrostatically. This formation is called *exciton* (from the word 'excitation'), and can wander about the lattice as an entity that is not a current carrier. We shall return to the mechanisms by which excitons appear and disappear at the end of this section.

Meanwhile we shall discuss the destiny of the photoelectron and the photohole that appeared in the crystal as the result of light absorption.

The photoelectron can be moved by a light quantum high into the conduction band. It then has a reserve of energy. Any electron travelling around a lattice collides with the atoms (ions) of the lattice, thus spending its energy. While it stays in the conduction band, the electron gradually moves down from its higher to its lower levels due to the collisions. When it reaches the bottom of the conduction band, the electron joins the pool of thermal electrons, i.e. the electrons that are present in some concentration in any crystal lattice without any incident radiation. A photoelectron is identical to a thermal electron, only its origin is different. It has the same mobility as a thermal electron and at the end of its life it recombines with a hole in the crystal. The mean lifetime of a photoelectron is estimated to be 10^{-3}-10^{-7} s.

At the start of its life, when the energy of the photoelectron is quite large, it can collide with an atom (ion) of the lattice and ionize it, i.e.

bring about a secondary free electron. This secondary electron, if its energy is sufficient, can bring about a tertiary electron. In this way the electrons can multiply. We see that one absorbed quantum can give rise to several photoelectrons and not just one. The destiny of a photohole is similar.

When there is a steady incident radiation, a stationary process can be established, when the number of appearing photoelectrons (photoholes) equals the number of recombining ones for the same time period. It turns out that the concentration of carriers (electrons or holes) is generally different from that in the absence of radiation. Therefore the conductivity of a crystal is changed. The difference between the conductivities of a crystal due to irradiation and that which can appear in the dark is called *photoconductivity*.

Photoconductivity depends on the intensity of the absorbed light. We shall discuss only the simplest case when the number of thermal transitions of electrons into the conduction band is negligible compared with the number of photo-transitions. Let us call the latter α_1. Clearly, α_1 is proportional to the number of absorbed quanta, which we shall designate J, i.e.

$$\alpha_1 = \beta J, \tag{2.21}$$

where β is a proportionality factor. The number α_2 of recombining electrons is proportional to the concentration n of electrons and that of holes, also equal to n:

$$\alpha_2 = \gamma n^2, \tag{2.22}$$

where γ is another proportionality factor. From $\alpha_1 = \alpha_2$ we obtain, according to (2.21) and (2.22)

$$\beta J = \gamma n^2,$$

whence

$$n = \sqrt{\frac{\beta}{\gamma} J}. \tag{2.23}$$

We see that the concentration n of photoelectrons that bring about photoconduction is directly proportional to the square root of the intensity of absorbed light.

As a rule, photoconduction is positive, i.e. light causes an increase in the number of free carriers. However, there are known cases when light brings about not an increase, but a decrease in the number of carriers. It can be shown that this may be related to the fact that light not only changes the conditions of the generation, but also the conditions of the recombination of carriers.

In conclusion we shall return to the photoelectrically nonactive light absorption which we mentioned at the beginning of this section, that is when the absorption of a quantum does not cause either a free electron or a hole to appear. This is the exciton mechanism of light absorption. An exciton cannot live forever and sooner or later it is extinguished. There are two ways for this to happen.

First, the electron and the hole of which the exciton is composed can annihilate each other, i.e. the electron recombines with the hole. The electron is like a billiard ball that falls into a pocket leaving neither a free electron nor a free hole. This sort of disappearance is accompanied by a release of energy.

The other way is for the exciton to dissociate into a free electron and a free hole. The electron and the hole making up the exciton move far enough apart for the electrostatic attraction binding the exciton together to fall practically to zero. An exciton requires some energy to disappear in this manner.

An exciton can disappear in a collision with a donor or acceptor structural defect in the lattice. In case of a collision with a donor defect, the energy released in the annihilation is used to ionize the defect. A free electron and a hole bound to the defect appear. In case of collision with an acceptor defect, a free hole and a localized electron appear. In both cases the exciton disappears and a free electron or a free hole appears instead. In these cases the absorption of a quantum results, in the long run, in the appearance of a photoelectron or a photohole, but these are the secondary not the primary consequence. The exciton is an intermediate stage.

It is very interesting to look into a possible mechanism for the collision between two excitons, when both excitons destroy each other. One of the excitons is annihilated, and its annihilation energy is used to dissociate the other exciton. Thus instead of two excitons, we get a free electron and a free hole.

The exciton we have discussed so far, i.e. an electron and a hole close together and bound to each other electrostatically, is called a *Wannier-Mott exciton*. It is possible (in an atomic

crystal) for an electron and a hole to coexist on the same atom. This should be imagined as follows. Suppose, for the sake of simplicity, there is a monovalent atom in which the valence electron is not in a normal internal orbit (which is occupied by a hole in this case), but is in one of the outer orbits. The atom is in an excited state and as such is conventionally called a *Frenkel exciton*. A Frenkel exciton, like a Wannier-Mott exciton, can travel anywhere within the crystal without carrying a charge.

So far we have discussed the excitation of a crystal due to the absorption of a light quantum and resulting in the appearance of a free electron, hole, or exciton. Another kind of excitation which results in the crystal glowing is possible. This means that a crystal, having absorbed a quantum of a given frequency, can emit quanta of the same or another frequency. The crystal glow caused by light is called *photoluminescence*. We shall return to this phenomenon in Chapter 4 when we shall discuss the luminescence of nonmetallic crystals (semiconductors included).

If a semiconductor that can luminesce is put into an external electric field, i.e. we put a potential difference across it, the semiconductor may either glow less or, by contrast, glow more, depending on the direction of the field. This is a recently discovered effect called *field luminescence*.

Chapter 3

Electrons on a Semiconductor Surface

3.1. Semiconductor Surface Phenomena

We have operated so far with an infinitely extensive crystal that exists only in the imagination of the theorist. In reality, a crystal is bounded by a surface, and this is going to be at the focus of our discussion in this chapter. As in the preceding chapters, the dramatis personae in what takes place on the surface of the semiconductors, are the electrons and holes of the crystal lattice.

Semiconductor surfaces are becoming the object of the research of scientists from many different disciplines. First and foremost, the physicists and engineers engaged in the physics and production of semiconductors are most interested because modern semiconductor-device production rests upon the problem of surface. The quality of a semiconductor device crucially depends on the properties of its surface. Any instability in these properties, such as uncontrollable changes with temperature or under the influence of the environment, disrupts the operation of a semiconductor device. Hence, the high percentage of production rejects. Learning to manage the surface properties is one of the top priorities of the semiconductor industry.

Semiconductor surfaces are interesting for other sorts of researchers such as physical chem-

ists and chemists studying adsorption and catalysis. Semiconductor surfaces are where the adsorption and catalytic processes take place. Most semiconductors can catalyse chemical reactions. Semiconductor surfaces appear in catalysis studies more often than it may seem at first sight. The point is that many metals often have a semiconductor sheath (oxide film), so that the processes apparently occurring on a metal's surface are in reality occurring on the surface of a semiconductor. Industrial chemistry is faced with the problem of producing catalysts that are active for particular reactions. Learning to manage the adsorption and catalytic properties of a surface is a top priority in the chemical industry.

Consequently, semiconductor surfaces are interesting from the viewpoint of semiconductor physics and semiconductor-device production, and from the viewpoint of the studies in adsorption and catalysis and the chemical industry employing catalytic processes.

A semiconductor surface is the boundary between two phases (two media). It is the frontier where our dramatis personae meet. Some of them come to the surface from the gas phase: these are the gas molecules. Others come from the depth of the semiconductor: these are free electrons and holes. The semiconductor surface is the boundary of both phases and interacts with one on one side and with the other on the other side. This is why surface effects are not two-dimensional, as it may seem to be at the first glance, they are three-dimensional.

A surface, like any boundary, can be tackled

from two sides. Semiconductor physicists deal with the surface from the side of the solid. They come, we might say, to the surface from the bulk of the semiconductor and study surface-bulk effects ignoring the interaction between the surface and the environment. Catalytic chemists study the surface in another way, from the side of the gas and study the interaction between the semiconductor surface and the environment, though often forgetting the surface-bulk interactions.

Basically, these two competing approaches do not deal with the surface as it is, that is its behaviour and features escape our minds. In order to understand a surface, we have to consider it together with both phases between which it is the boundary.

A semiconductor's surface is also at a boundary between two sciences: physics and chemistry. As Lomonosov* would say, 'physics and chemistry are so intertwined that one cannot exist without the other'. Modern technology often focuses our attention on problems at the junction of two sciences. These are the most promising topics both in their scientific and applied aspects. Semiconductor surfaces being a stumbling-block for semiconductor-device production on the one hand, and useful for catalytic chemistry on the other hand, is an example of such a problem. A surface is a two-faced Janus looking in different directions: chemistry and physics.

We have so far spoken about a semiconductor in a gaseous environment. We called the semi-

* M. V. Lomonosov, an outstanding Russian scientist and scholar (1711-1765).

conductor surface the interface between the solid and gas phases, but it can be used in a wider way. A surface does not necessarily imply the interface between the semiconductor and a gas, it can be also the interface between the semiconductor and a liquid (an electrolyte), or between the semiconductor and a metal, or between two semiconductors with differing chemical compositions. In every case, the surface brings into the system its own features.

We shall return to the semiconductor contacting with the gas phase below. Here we shall discuss briefly a semiconductor film in contact with a metal and the processes that occur on the surface of a film on the metal. Then we shall consider two semiconductors in contact, one being an *n*-type semiconductor, the other a *p*-type semiconductor, and consider some of the important effects that occur at such an interface.

Suppose that a metal is covered with a film of its oxide (a semiconductor), so that the internal surface of the semiconductor touches the metal and the external surface is in contact with a gas phase containing oxygen. Molecules of oxygen settling (adsorbing) on the surface of the semiconductor film dissociate into atoms. These oxygen atoms have a great affinity for electrons and are typical acceptors capturing and retaining lattice electrons. The result is that the external surface of the semiconductor becomes negatively charged and attracts the positive metal ions both from the film and from the metal substrate. The external surface then has some atoms of oxygen, borrowed from the gas phase, and metal atoms supplied by the metal

substrate. The oxide film (the semiconductor) becomes thicker, and the metal phase less compact. Thus the metal is destroyed due to the growth of the semiconductor film on its surface. This phenomenon is called *gas corrosion*. It is a deadly disease that affects thousands of tons of metal every year and has an army of scientists fighting it. We see that surface effects attract the attention of not only semiconductor physicists and catalysis chemists, but also corrosion technologists.

Let us in conclusion consider the case of two semiconductors in contact, one of which is an electronic semiconductor and the other a hole semiconductor. Bear in mind that the same semiconductor can become either an electronic or a hole one, depending on the nature of the impurity it contains. Thus, for instance, we should distinguish between *n*-germanium and *p*-germanium. Suppose that a *p*-type semiconductor is to the left of the interface, and an *n*-type semiconductor is to the right, as shown in Fig. 30. This interface is called a *p-n junction*. Now suppose that a potential difference is applied across our semiconductors. It can be applied so that the electric field is either directed from left to right (as shown in Fig. 30*a*) or directed from right to left (as shown in Fig. 30*b*). In the first case the holes and the electrons travel towards each other and recombine when meeting at the interface. An electric current thus appears between the electrodes. If the cathode and the anode are exchanged, i.e. the direction of the electric field is reversed and we go from Fig. 30*a* to Fig. 30*b*, then the holes and the electrons

appear and travel in the opposite direction without providing any electric current.

This is the phenomenon of *unipolar conduction*, i.e. the current is only carried in one direction and is absent in the opposite one. When the

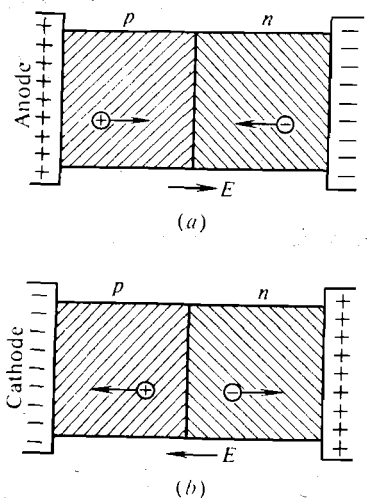


Fig. 30. Rectifying effect of a contact between p-type and n-type semiconductors

alternating current is passed, a contact that allows the current to flow in one direction only, plays the role of a rectifier.

3.2. Adsorption on a Semiconductor Surface

Suppose a semiconductor is placed in a gaseous environment and the gas molecules strike the surface of the solid and adhere to it. This is the

process of *adsorption*. The molecules pass some time on the surface and sooner or later are drawn away from the surface and return to the gas. This is called *desorption*. Consequently, a gas molecule passes a period of time in an adsorbed state on the surface of the solid. The solid, therefore, captures the gas molecules and retains some of them on its surface. The solid on which adsorption occurs is called the *adsorbent*, while the substance that is adsorbed is called the *adsorbate*. In this section and in the following one we are going to discuss adsorption and the role played by the electrons and holes of a semiconductor.

Depending on the nature of the forces retaining an adsorbed molecule on the surface of the adsorbent, we distinguish between *physical adsorption* and *chemical adsorption* (*chemisorption*). In physical adsorption, the forces are the same in nature as those between the molecules of a gas (*Van der Waals forces*). In chemisorption, the forces are chemical in nature between the atoms in the molecule (*exchange forces*).

Chemisorption differs from physical adsorption in a number of ways. Firstly, the distance between an adsorbed particle and the surface in physical adsorption is greater than in chemisorption, while a chemisorbed particle can be assumed to be pressed into the surface. Secondly, a particle is attached to the surface stronger in chemisorption than in physical adsorption. And lastly, in physical adsorption the adsorption rate is the slower, the higher the temperature, while in chemisorption the adsorption rate increases as the temperature rises. If we designate

the adsorption rate dN/dt where dN is the number of molecules captured per unit surface during time dt, then in chemisorption, as a rule, we have

$$\frac{dN}{dt}=\alpha \exp\left(-\frac{\varepsilon}{kT}\right), \qquad (3.1)$$

where α is a factor that slowly decreases with temperature, and ε is a constant that is called the adsorption *activation energy*. Adsorption that obeys (3.1) is called activated adsorption. As usual, in (3.1) the k is Boltzmann constant, and T is absolute temperature.

The process of adsorption is finished when the number of molecules adsorbing on a surface in a given time equals the number of molecules desorbing from the same surface. This means that a balance is established between the surface and the gas (the *adsorption equilibrium*), and this is characterized by a certain coverage of the surface with the gas molecules.

This coverage of the surface, i.e. the number of molecules retained by the surface in a steady adsorption equilibrium, depends above all on the temperature and pressure. It decreases with temperature at a given pressure and increases with pressure at a given temperature. A curve showing how the coverage depends on pressure (at a given temperature) is called an *adsorption isotherm*. Some adsorption isotherms are shown in Fig. 31. Here N is the number of molecules adsorbed per unit surface in conditions of an adsorption equilibrium, and P is the pressure in the gas phase. Curves *1* and *2* refer to two different temperatures, curve *2* corresponding to the greater temperature. We can see that the coverage of the surface first rises with pressure, but when the pressure becomes great enough, it reaches saturation, i.e. it does not depend on pressure anymore.

The most molecules that can be retained on a surface at a given temperature, i.e. the number N^* that corresponds to the horizontal portion of the isotherm (at high pressures, see Fig. 31) is what we call the *adsorptive capacity* of the surface. The adsorptive capacity of a surface is its holding capacity with respect to gas molecules. Naturally, the same surface can have different adsorptive capacities for different kinds of molecules.

The adsorptive capacity of a surface depends on the nature of the surface and how it has been processed. Adsorptive capacity can be changed by external effects. It is interesting that a surface often also responds to what is going on within the adsorbent. Thus, introduction of impurities inside the adsorbent affects the adsorptive capacity of its surface. Certain trace impurities noticeably increase adsorptive capacity, while other impurities decrease it. We shall return to this subject when we discuss the interaction between the surface and the bulk.

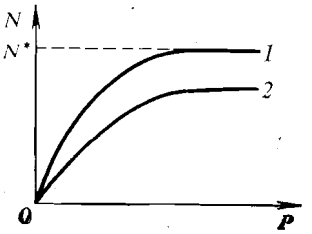


Fig. 31. Adsorption isotherms for two different temperatures (curves *1* and *2*)

In particular, sometimes the adsorptive capacity changes when the adsorbent is exposed to light. This is called the *photoadsorptive effect* and the adsorptive capacity can either increase or decrease. The *positive* and *negative* photoadsorptive effects are distinguished respectively. The sign of the effect is determined by the nature of the adsorbate molecules and by the nature of the adsorbent, and what is more it is determined by history. Some recent experiments have shown that the sign of the photoadsorptive effect can be changed if the adsorbent has been treated in a certain way.

An adsorbent is placed in a closed vessel filled with a gas at constant pressure. This means that the number of gas molecules per unit volume (i.e. the population density in the gas) is kept stable. Naturally, a number of gas molecules are attached to the adsorbent's surface.

Now if the adsorbent is exposed to photoadsorption-active light, we may observe a drop or sometimes a gain in the gas pressure. Switch off the light and the pressure returns to its initial level. This is the evidence that the exposure removes some of the adsorbed molecules from the surface and returns them into the gas phase (this is photodesorption, the negative photoadsorptive effect) or, vice versa, it makes more gas molecules stick to the surface thus increasing the surface population (this is photoadsorption, the positive photoadsorptive effect).

Sometimes the photoadsorption is irreversible. This means that the additional molecules adsorbed under the influence of the incident light are not removed from the surface when the light is

turned off. The temperature must be increased in order to remove them.

It was also found that adsorptive capacity can be changed by an external electric field. The adsorbent was used as one plate of a capacitor across which a potential difference was applied. The adsorptive capacity turned out to be different depending on whether there was an electric field or not. It was noticeably increased if the field was in a certain direction, and decreased if the field was in the opposite direction. It is curious that this effect is not symmetric with respect to the direction of the field.

Let us remark that molecules adsorbed by a solid can creep over its surface while remaining 'attached' to it. They can encounter other adsorbed molecules until they manage to break free from the surface.

When a molecule collides with the surface it can often break in the very act of adsorption, so that it is separate parts of the molecule that adhere to the surface rather than the molecule itself. Molecules or their separate constituents that exist independently on a surface can join together into new combinations to form new molecules.

Consequently, one kind of molecules may arrive at a surface from a gas and another kind leave the surface. The composition of the gas changes gradually, and its chemical transformation occurs. The surface of the solid is where the molecules are restructured.

Note that molecules which settle on the surface of a solid and are attached to it have properties different from those of the same mole-

cules in the free state. Relationships between two molecules can change, antagonism being changed for sympathy. Molecules that will not react in the free state will often react when they are prisoners on a surface.

Chemical reactions between gas molecules on semiconductor surfaces will be discussed again in Section 3.5, which deals with the catalytic effect of semiconductors.

3.3. The Role of Electrons and Holes in Adsorption

Gas molecules can adsorb on the surface of a semiconductor. Then they can desorb and return to the gas phase again. An essential role in adsorption and desorption is played by electrons and holes of the semiconductor because they control to a considerable extent the properties and behaviour of the molecules that settle on the surface.

As they wander about a semiconductor, electrons and holes can come out onto the surface. They can then encounter adsorbed molecules that arrive at the surface from the gas to pass some time on the surface, 'rest', and then return to the gas phase. The adsorbed molecules play the same role with respect to electrons and holes as foreign impurity atoms in the semiconductor; they capture free electrons and holes. (Once again: capturing a hole is the same as giving away an electron. These are just two different expressions of the same thing.)

The adsorbed molecules and atoms are surface impurities that interfere with the regular structure of the surface. Much like the usual bulk impurity, adsorbed atoms and molecules can play the role of acceptors and donors. Thus, for instance, adsorbed atoms or molecules of oxygen or chlorine are typical acceptors, while hydrocarbon molecules (e.g. C_2H_4) are an example of donors. As a rule, hydrogen atoms are also donors.

An adsorbed molecule (or atom) that managed to capture and hold an electron or a hole becomes electrically charged. The more current carriers (electrons or holes) there are in the semiconductor, the greater proportion of the adsorbed particles that are charged. The semiconductor surface thus gets charged as a result of adsorption. There are substantial consequences of this which we shall cover in the next section.

When the surface adsorbs acceptors, it gets charged negatively. Adsorption of donors, on the contrary, results in a positive charge on the surface. That is why the sign of the surface charge depends on the nature of the adsorbate particles.

We mentioned that at a given instant a certain fraction of the adsorbed particles are charged. The same can be said in other words: each adsorbed particle passes some time in a charged state during its lifetime in the adsorbed state. There is thus always a certain probability that a neutral adsorbed particle will become charged and a charge d one will become neutral. The transition of an adsorbed particle from the neutral state to the charged one or vice versa implies the localization or the delocalization of a free electron or hole on the adsorbed particle.

The fact that an electron or a hole is captured by an adsorbed particle interferes with the nature of its bond with the surface. The bond becomes stronger. In other words, an adsorbed particle in its charged state is attached more strongly to the surface than it is in the neutral state. It is clear why. It is more difficult to tear a molecule with an attached electron or hole away from the surface than it is to remove a neutral molecule. Indeed, the molecule with the extra charge needs to turn its electron or hole over to the semiconductor before it can go back into the gas. This in itself requires energy.

Consequently, we should distinguish between the two ways an adsorbed particle can be attached to the surface; these are conventionally called *weak* and *strong bonding*. The free electrons and holes of the semiconductor do not participate in a weak bond, and only the electrons belonging to the adsorbed molecule (atom) or those belonging to atoms of the semiconductor's crystal lattice participate in the bond. These electrons are drawn away from the adsorbed particle into the lattice or from the lattice to the adsorbed particle to form the bond. A strong bond is formed by the free electron or hole being captured by the adsorbed particle.

Thus, there is weak and strong adsorption. Weak adsorption is electrically neutral, strong adsorption is electrically charged. The average strength of the bond between the adsorbed particles and the surface is determined by the relative spread of the weak and strong bonds on the surface, or, in other words, by the ratio of the probabilities that the particle can be weakly or

strongly bonded. The adsorptive capacity of the surface, i.e. the possible population of adsorbed particles on the surface at given pressure and temperature, depends on this parameter. The relative spread of weak and strong bonds in turn is determined by the concentration of free electrons and holes on the semiconductor surface. This is why it is the free electrons and holes of the semiconductor that control its adsorption properties in the last analysis. They are the masters of the surface.

Obviously, the adsorptive capacity of the surface with respect to acceptor molecules is high, and with respect to donor molecules is low, if the concentration of free electrons in the plane of the surface is also high (or the concentration of free holes is low). The experimentally observed influence of incident light on the adsorptive capacity of a surface becomes clear now.

Photoelectrically active light, i.e. light that brings about photoconduction, increases the concentration of free electrons and free holes in the semiconductor. The additional electrons knocked out of the semiconductor by the light, come to the surface and partially settle there. If the surface is negatively charged, its negative charge increases, or if it is positively charged, its positive charge decreases. The extra holes that also appear in the semiconductor due to the exposure to the light come to the surface as well, but they, by contrast, remove some electrons from it. The charge of a negatively charged surface decreases, while the charge of the positively charged surface increases. The result of these two processes, that act simultaneously but

in opposite directions is that the charge of the surface is changed.

The number of adsorbed particles in a charged state on the surface changes under the influence of light. Hence the average strength of their bond to the surface also changes. It becomes on the average either harder or easier to tear a molecule away from the surface. In other words, the adsorptive capacity of the surface with respect to these molecules is affected. The result is *photoadsorption* or *photodesorption*, or positive or negative photoadsorptive effect.

When speaking of a semiconductor surface, we have so far had an idealized picture in mind. We supposed the surface was a plane with a regular structure. Any real surface differs from the ideal one in that there is a 'disorder' which disrupts the regular pattern. A real surface has a variety of steps, peaks, atoms knocked from their lattice positions to the surface, and vacancies, i.e. empty lattice points, and other macro- and microdefects of structure. These are convenient sites for the adsorption of gas molecules, and any adsorption is focused around these defects which are called *centers of adsorption*. Adsorption can occur at these centers, as it does on an ideal surface, with or without the participation of free electrons and holes ('strong' and 'weak' bonds). It is clear that the nature and concentration of the adsorption centers on the surface depends on the 'biography' of the surface, i.e. on the treatment it has been subjected to.

Let us remark in conclusion that defects of the surface, in particular surface impurities,

play in adsorption a double role. On the one hand, they regulate the concentration of free electrons and holes which, in turn, regulate the adsorptive properties of the surface. On the other hand, surface defects can themselves act as adsorption centers.

3.4. Interaction of the Surface with the Bulk

When a semiconductor is in a gaseous environment, its surface always has some gas molecules stuck to it. When gas molecules are attached to the surface (adsorbed on the surface) there is, as we now know, some charging of the surface. This extra surface charge has important consequences which we shall discuss in this section.

Let us note to begin with that the electrons populating a semiconductor are locked inside it. To knock an electron out of the semiconductor requires energy which, as is the case with metals, is called the *work function*. The work function is the bill an electron has to pay to get beyond the semiconductor. The work function characterizes how strongly the electrons of the semiconductor are attached to it. It is obvious that the work function depends on the nature of the semiconductor. We shall now see that it also depends on the nature of the gas surrounding the semiconductor.

Consequently, only those semiconductor electrons that possess sufficient energy can leave it. If we supply the electrons with energy, we can induce their emission, i.e. make the semiconductor release some of its electrons. Let us discuss

the external forces on a semiconductor that can cause it to do so.

Electron emission can be produced, first of all, by heating the semiconductor up to a high enough temperature. This is called *thermionic emission*, or *Richardson effect* in semiconductors, and is much like the same effect in metals. The main point is that the mean kinetic energy of the free electrons in the semiconductor increases with the rise of temperature. More and more high-velocity electrons appear that are 'rich' enough to 'pay' for the 'right' to be liberated from the semiconductor.

Electron emission can be also induced by exposing the semiconductor to light. The energy of the electrons in this case is increased at the expense of the absorbed light. Here the light is the source of electrons' energy. This is called *photoelectron emission*, or just *photoemission*, or the *external photoelectric effect* that we discussed in the preceding chapter.

Bombarding a semiconductor with fast particles, such as ions or electrons, can also give rise to electron emission. The striking particles transfer their energy to the electrons of the semiconductor. If this energy is sufficient for the work function, electrons are emitted from the semiconductor. This is called *secondary (electron) emission*.

And finally, electron emission can be produced by an external electric field. A strong enough field will provide a force that can extract electrons from the semiconductor. An electron, pulled by this force, can overcome the energy barrier that separates the semiconductor from the external space and go beyond the semiconductor. We speak of *cold emission* in this case.

The work function of the electron, which describes the height of the energy barrier surrounding the semiconductor, can be changed by some effects. For instance, the work function can be changed due to adsorption. The surface charge that appears due to adsorption can either enhance or hamper the emission of electrons from the semiconductor. This depends on whether the surface becomes positively or negatively charged, i.e. depends on the nature of the adsorbing gas. In case of negative charging the work function increases, in case of positive charging it decreases.

Thus, the adsorption of oxygen (an acceptor, hence the surface is negatively charged) always produces an increase in the work function, while the adsorption of hydrogen (a donor, hence the surface is positively charged) results in a decrease in the work function. In other words, the payment an electron has to make to receive the right to leave the semiconductor depends both on where it is going *from* and where it is going *to*, i.e. both on the nature of the semiconductor, and on the nature of the surrounding gas.

We can often evaluate the composition of a gas by the manner in which a work function is changed by its adsorption. At higher gas pressures, i.e. at larger gas densities, more molecules accumulate on a semiconductor's surface and hence the influence of the gas on the work function is greater.

The gas molecules that are adsorbed on the surface affect the behaviour of the electrons and holes not only when they arrive at the surface.

Electrons and holes deep within a semiconductor often feel what occurs on the surface, just as the population of a country responds to events at its frontiers.

An example is the effect adsorption can have on the conductivity of a semiconductor. Suppose a gas is introduced into a closed vessel containing a semiconductor. The gas is adsorbed and the conductivity of the semiconductor changes monotonically with adsorption until it reaches a certain constant value. Often the process of adsorption can be followed by observing the change in the conductivity.

If a semiconductor surface becomes positively charged due to adsorption, then the concentration of free electrons near the surface increases. The electronic conductivity near the surface also increases, and this may affect the conductivity of the whole specimen if it was not initially particularly high. If the surface becomes charged negatively due to adsorption, the surface layer acquires a surfeit of holes, and there is an increase in hole conductivity.

Consequently, adsorption produces a rise or a drop in the conductivity of the semiconductor depending on what *kind of gas* (acceptor or donor) is adsorbed and on the *kind of semiconductor* (electronic or hole). Thus, the adsorption of oxygen onto zinc oxide (zinc oxide is an electronic semiconductor) results in a decrease of the conductivity of zinc oxide, while the conductivity of nickel oxide (a hole semiconductor) is increased by the adsorption of oxygen. Hydrogen usually produces the opposite effect. In these examples, the gas molecules act like impurity

atoms added to a crystal in that they either take away the current carriers from the semiconductor, or supply it with them.

The appearance on a semiconductor surface of adsorbed molecules not only affects the work function and conductivity of the semiconductor, it also affects its typically bulk (and not surface) properties such as its luminescent properties. In fact, a semiconductor's luminescence can be reduced and sometimes extinguished altogether, although in other cases (depending on what kind of gas is adsorbed onto the semiconductor) the luminescence can be increased.

This all becomes clear if we recall how an external electric field affects the luminescence of a semiconductor (the luminescent field effect). The appearance of adsorbed particles on a surface means that the surface becomes charged. The electric field of the charged surface penetrates the semiconductor and plays the same role as an external field applied to the semiconductor. The field is directed according to the nature of the adsorbed gas. In the last analysis, this is a case of the luminescent field effect.

So far we have considered the influence of the surface on the bulk properties of a semiconductor. In conclusion, let us discuss the influence of the bulk on the surface. It is amazing how much the surface is sensitive to what goes on within the semiconductor. Thus, the adsorptive capacity of the surface is not only sensitive to impurities on the surface, but also to impurities and their concentrations in the semiconductor. We can change the adsorptive capacity of a semiconductor surface manyfold by introducing even tiny amounts of impurity to the bulk.

As we saw in Section 3.3, the adsorptive capacity of a surface depends, all things being equal, on the concentration of free electrons and free holes in the semiconductor (to be more exact, in the plane of the semiconductor surface). At the same time we saw that the concentration of free electrons and holes can be regulated by an impurity in the semiconductor. Therefore an impurity inside the semiconductor controls the adsorptive properties of its surface by affecting the pool of free electrons and holes populating the semiconductor. This is a long range effect. The impurity atoms are not in immediate contact with the adsorbed gas molecules. Free electrons and holes are the intermediaries between them.

3.5. Chemical Reactions on a Semiconductor Surface

We shall consider in this section what happens when a semiconductor is placed into a mixture of different gases, i.e. a mixture of molecules of different kinds. Suppose that a chemical reaction can occur in the gas resulting in the disappearance of some molecules and the appearance of new ones. Substances that participate in a chemical reaction are called *reagents*, and a substance that appears as a result of a chemical reaction is called a *product*. The quantity of the product that appears per unit time is called the *rate of the reaction*.

When a semiconductor is placed into an environment of reacting gases, gas molecules can adsorb onto its surface and react with other

adsorbed molecules or with molecules in the gas phase. In this way the reaction is transferred from the gas phase onto the solid's surface. The rate of the reaction changes, and if it is increased, the semiconductor acts as a catalyst (an accelerator) to the reaction. Sometimes a reaction can be accelerated thousandfold.

For instance, the reaction of oxidation of carbon monoxide CO proceeds according to the equation

$$\underbrace{2CO}_{\text{two molecules of carbon monoxide}} + \underbrace{O_2}_{\text{a molecule of oxygen}} \rightarrow \underbrace{2CO_2}_{\text{two molecules of carbon dioxide}}$$

in the gas phase. When there is just a mixture of carbon monoxide and oxygen 'on its own', the reaction proceeds very slowly (or practically not at all). However, if a solid, such as manganese dioxide MnO_2 or silver oxide Ag_2O (they are semiconductors), is added to the mixture the reagents react rapidly to produce carbon dioxide. Manganese dioxide or silver oxide catalyze this reaction. And if the catalyst is removed the reaction practically comes to a halt.

The relative increase in the rate of reaction brought about by a catalyst is called its *activity*. Naturally, a catalyst that is active with respect to one reaction may be inactive with respect to another. Catalytic activity depends on a number of conditions and in particular, it depends on the impurities in the catalyst or on its surface. Some impurities promote the activity, and so are called catalyst *promoters*, while some can

inhibit a reaction and are called catalyst *poisons.* The activity of a catalyst always rises with temperature.

Now let us discuss the mechanism of the catalytic activity of semiconductors and the role of the free electrons and holes in the catalytic

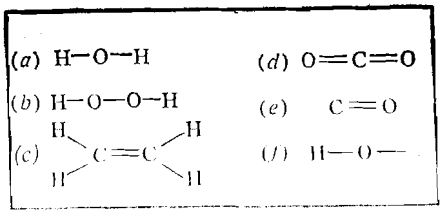


Fig. 32. Structural formulas for some simple molecules: (*a*) water H_2O, (*b*) hydrogen peroxide H_2O_2, (*c*) ethylene C_2H_4, (*d*) carbon dioxide CO_2, (*e*) carbon monoxide CO, (*f*) hydroxyl group OH

reactions that occur on a semiconductor surface. These problems are covered by the electronic theory of chemisorption and catalysis.

Recall that individual atoms and groups of atoms in a molecule are linked to each other by valence bonds. These are shown in structural chemical formulas by dashes. The structural formulas for some simple molecules are given in Fig. 32. A water molecule H_2O (see Fig. 32*a*) consists of two atoms of hydrogen each of which is linked to the oxygen atom, i.e. each hydrogen atom is monovalent and the oxygen atom is bivalent. The same valences are observed in a molecule of hydrogen peroxide H_2O_2 shown in Fig. 32*b*. In a molecule of ethylene C_2H_4

(Fig. 32*c*), each carbon atom C is linked by a single bond to two hydrogen atoms H and by a double bond to the other C atom, i.e. the C atoms in an ethylene molecule are tetravalent. The same occurs in a CO_2 molecule where the C atom is linked to two oxygen atoms O each of which is bivalent (see Fig. 32*d*). Note that some atoms can have a variable valence, i.e. when they belong to different molecules they can have different valences. Thus a C atom is tetravalent in the C_2H_4 and CO_2 molecules, while it is bivalent in CO molecules (see Fig. 32*e*).

The valence of an atom has a remarkable feature: it tends to 'saturation'. An 'unsaturated' bond always tends to become 'saturated' at the expense of another 'unsaturated' link. In other words, a valence dash tries to link an atom (or a group of atoms) with another atom (or another group of atoms). All the valence links in a stable molecule are always 'saturated', i.e. none of the valence dashes dangles as shown in Fig. 32*f*. If a 'valence-saturated' molecule is split into two parts, the result is two molecules with 'unsaturated' ('free') links. Such molecules may be called radicals. The OH or hydroxyl group which can appear if an H atom is separated from an H_2O molecule, is an example of a radical (Fig. 32*f*).

Two radicals are two molecules that hold out their hands to join each other. A molecule with one, two or more 'unsaturated' links may be conventionally called a monovalent, bivalent, or polyvalent radical, respectively. It is clear that a molecule in the radical state is always more reactive, i.e. its relative capacity to form

chemical compounds is greater than that of a valence-saturated molecule.

The secret of the semiconductor catalysts, from the point of view of the electronic theory of catalysis, is that gas molecules become radicals (at least partially) on the semiconductor's surface. This can be explained by saying that a semiconductor crystal can be considered to be a large molecule (a macromolecule) with some 'free' bonding sites or a kind of a polyradical. This brings about its catalytic properties, according to the electronic theory. It is remarkable that the role of the 'free' links of the catalyst is played by free electrons and holes of the semiconductor.

Considering free electrons and holes of a semiconductor as 'free' links, we can attribute to them the following specific properties. Generally, they are not localized and can roam freely over the semiconductor, emerge on its surface, adhere to impurity atoms and become detached again, appear and disappear. They may encounter gas molecules adsorbed on the surface, break the bonds within these molecules and become 'saturated' at the expense of these bonds, at the same time transforming the molecules into radicals, and turning radicals into 'valence-saturated' formations. This is the way the 'free' links of the catalyst come into play and carry out reactions.

As an example, let us take a catalytic reaction of oxidation of carbon monoxide CO. Suppose that an oxygen atom is adsorbed on the surface, and this atom is 'strongly' bonded to the surface, i.e. it is attached to an electron localized

around it. This state is shown in Fig. 33*a* using bond dashes. We now have a surface radical, and a CO molecule from the gas phase can now bond onto the oxygen's free electron pair. The C atom becomes tetravalent, and the surface formation shown in Fig. 33*b* appears. Now if the electron binding this formation to the surface becomes

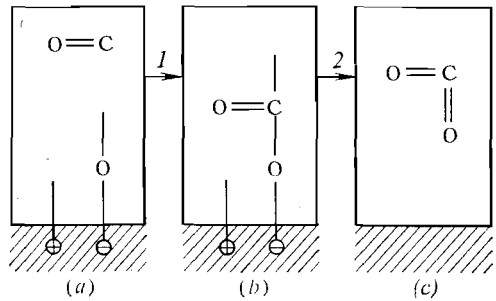


Fig. 33. Possible mechanism of carbon monoxide oxidation. Circles with 'plus' and 'minus' signs and vertical dashes are 'free' valence links of the surface (electrons, holes)

delocalized or annihilated by an approaching hole, a CO_2 molecule, the product of the reaction, is liberated into the gas phase (Fig. 33*c*).

According to Fig. 33, this reaction can be regarded as consisting of two stages, the first (arrow *1* in Fig. 33) proceeding the faster, the greater the number of oxygen atoms 'strongly' adsorbed on the surface, or, in other words, the higher the concentration of free electrons on the surface. The rate at which the second stage proceeds (arrow *2* in Fig. 33) depends on the

probability that a localized electron is annihilated by a free hole, i.e. the rate is the faster the more the concentration of free holes. We see from this example that the rate of reaction is controlled by the concentration of free electrons and holes on semiconductor surface. They are the true masters of the catalytic process (from the standpoint of the electronic theory).

Chapter 4

Electrons in Dielectrics

4.1. Dielectric Conductivity

Typical dielectrics are alkali halide crystals, i.e. crystals composed of the positive ions of an alkaline metal (for instance, Na^+, K^+, or Li^+) and the negative ions of a halogen (for instance, Cl^-, Br^-, or I^-). The conductivity σ of such crystals is extremely low. As we mentioned in Section I.2, a dielectric has a conductivity of the order of 10^{-15}-$10^{-16}\,\Omega^{-1}\cdot cm^{-1}$ (10^{-13}-$10^{-14}\,S\cdot m^{-1}$) i.e. 10^{19}-10^{20} times less than that of metals.

Dielectrics are not only unlike metals and semiconductors as regards the value of their conductivity, they are also unlike metals and semiconductors because the current in them is not electronic in nature, as it is in metals, and not electronic or hole in nature, as in semiconductors, it is rather ionic, as it is in liquid electrolytes. That is why alkali halide ionic crystals are often called solid electrolytes. Electric charge is carried by 'heavy' particles that are several thousand times heavier than electrons.

Any real ionic crystal has ions at the lattice points and at least a few in the interstices. These interstitial ions, squeezing through the regular ions of the lattice, move from one interstitial position to another. When they move in a prevailing direction, as happens when the

crystal is in an external electric field, we observe an electric current.

The electric current in an ionic crystal is not only carried by interstitial ions. Any real crystal has at least a small number of empty lattice points, i.e. lattice points from which the ions that should occupy them have been withdrawn. These are called *vacancies*. A vacancy can be filled by an ion from a neighbouring lattice point, which is tantamount to the motion of the vacancy. The motion of vacancies caused by an external electric field is also an electric current in a crystal. Vacancies resulting from removal of positive ions (for instance, due to Na^+ ions removed from a NaCl lattice) move in the direction of the anode, i.e. they behave like negative charges. Vacancies of negative ions (for instance, due to Cl^- ions removed from a NaCl lattice) move to the cathode, i.e. behave like positive charges. Accordingly, the current in an ionic crystal can be due to either interstitial ions or vacancies, or both.

If the crystal is heated, the regular ions in the lattice points oscillate about their equilibrium positions with some amplitude, which increases as the temperature rises. They can tear away from their lattice points and jump into an interstice at high temperatures. Owing to this, heating promotes the appearance of both interstitial ions and vacancies, the number of which rises with temperature. Now if the temperature is rapidly decreased, the interstitial ions do not have time enough to return into their lattice points and they get stuck in their interstices. The defects of the lattice, the concentration of which is more than the concentration that should exist at the given temperature, are said to be *frozen*.

In order to push ions from their lattice points into interstitial positions, energy is required and this is called the *dissociation energy*. As the temperature is increased the positive ions (the cations) dissociate at moderate temperatures, and then, at higher temperatures, the negative ions

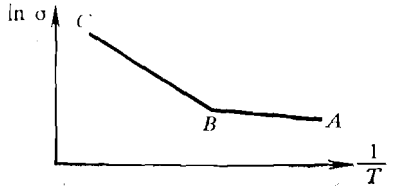


Fig. 34. Temperature diagram of ionic conductivity: AB—cationic conductivity, BC—anionic conductivity

(the anions) join the process. The anions are larger and so require greater dissociation energies. Thus, if we take crystals of common salt NaCl at room temperature, all the current is carried by sodium ions Na^+, and the migration of chlorine ions Cl^- only becomes noticeable at temperatures exceeding 600°C. The addition of impurity ions always raises the conductivity.

As the temperature rises more ions dissociate and therefore the conductivity of the crystal increases. As was the case with semiconductors, the conductivity σ of a dielectric is very sensitive to temperature. This is presented diagrammatically in Fig. 34, where AB corresponds to the cationic conductivity, and BC represents the

anionic conductivity. The slope of these straight lines yields the respective dissociation energies.

It is a feature of dielectric crystals that when a potential difference is applied across one the current appearing in the crystal is not constant, but gradually decreases with time to a constant minimum value, even if the potential difference is kept up constant. This is because as the ion current passes through the crystal, the ions get

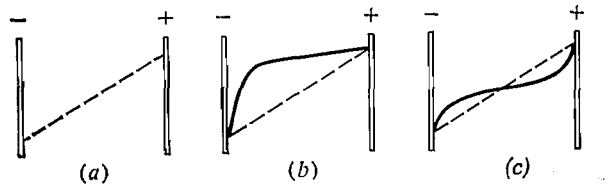


Fig. 35. Potential diagrams in dielectrics: (*a*) initially, (*b*) in a dielectric with cationic conductivity, (*c*) in a dielectric with mixed ionic conductivity

stuck near one of the electrodes (or both) producing a bulk charge. The field of this bulk charge is in the opposite direction to the external field, thus counteracting it. This phenomenon is called the *high-voltage polarization* of a dielectric. The motion of an ion through a crystal is hindered by the high-voltage polarization, and so the current drops.

The distribution of potential within a dielectric crystal is shown in Fig. 35. The dashed line represents the potential diagram at the initial instant, when there is no bulk charge and the current within the crystal obeys Ohm's law:

$$I = \frac{V}{R}, \tag{4.1}$$

where V is the potential difference, and R is the resistance of the crystal. The solid line in Figs. 35*b* and 35*c* shows the potential distribution in the presence of the bulk charge in the crystal. Fig. 35*b* refers to the case when the bulk charge is produced by positive ions condensed near the cathode. Instead of (4.1), we have in this case

$$I = \frac{V - P}{R}, \tag{4.2}$$

where P is the potential caused by the polarization. Now if we short-circuit the crystal, we shall observe a reverse current passing from the cathode to the anode without any external potential difference applied. And the amount of electricity passed in this case will be equal to the amount of electricity passed in the right direction. Fig. 35*c* refers to the case of mixed conduction in a crystal, when cations concentrate near the cathode and anions near the anode.

Let us remark in conclusion of this section that ionic conduction in dielectrics can be transformed into electron conduction, and the conductivity can then be raised by several hundred thousand times.

A factor which can do this is exposure to light. Light quanta of certain frequencies can be absorbed by the lattice ions and can knock electrons from them in the process so that the crystal is filled with free electrons. This is the internal photoelectric effect, which is identical to that which occurs in semiconductors and which we discussed above.

Another factor is an external electric field applied to the crystal. As before, let us designate the strength of this field E. The conductivities σ of both semiconductors and dielectrics do not depend on E if it is not too great, and we say that Ohm's law holds true:

$$\sigma = \text{const}. \tag{4.3}$$

Ohm's law is violated above a critical field strength. This happens earlier in semiconductors with electronic (or hole) conduction than in dielectrics, for which ionic conduction is characteristic. When E rises above the critical level, there is a region of instability. In a number of cases Poole's law starts being obeyed at high field strengths, viz.

$$\ln \sigma = \alpha E + \beta, \tag{4.4}$$

where α and β are constants. This law describes a rapid growth of the conductivity with the increase of the field. Figure 36 is a diagrammatic representation of the dependence of σ on the field strength E for the case of electronic or hole conduction (curve ABC) and for the case of ionic conduction (curve $A'B'$). The electronic conductivity in crystals is small compared with ionic conductivity at $E < E_1$. By contrast, the electronic component of conductivity becomes

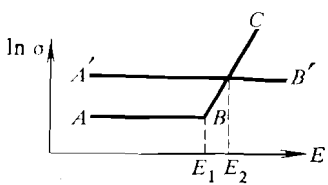


Fig. 36. Conductivity as a function of field strength E: $A'B'$ — ionic conductivity, ABC—electronic or hole conductivity

predominant and starts increasing at $E > E_2$. This increase in electric conductivity brings about, in the long run, a *breakdown* effect, which we shall discuss in the next section.

4.2. Dielectric Breakdown

When the strength of the electric field applied to a dielectric is increased, the current in the dielectric rises. This rise, however, cannot be unlimited. Sooner or later a catastrophe happens. The dielectric loses its insulating properties and is destroyed. This is called a dielectric breakdown, and the electric field strength at which the breakdown occurs is called the breakdown strength.

There are different dielectric breakdown mechanisms. Let us consider them.

The passage of current through any body (a conductor, a semiconductor, or a dielectric) is always accompanied by the evolution of heat. This is called *Joule's heat*, i.e. the heat evolved in compliance with Joule's law:

$$Q = c \frac{V^2}{R}, \tag{4.5}$$

where Q is the amount of heat evolved in the specimen per unit time, R is the resistance of the specimen, V is the potential difference applied across the specimen, and c is a constant which depends on the system of units being used.

When the heat is evolved, it is transferred from the hot parts of the dielectric to cooler ones. If a balance between the evolved and

abstracted heat is established at some stage, then the dielectric's temperature will not rise. The specimen will remain heated to a certain constant temperature.

However, if there is no balance between the evolved and abstracted heat, the temperature of the dielectric will rise until the dielectric melts and loses its dielectric properties. This is a case of dielectric breakdown called *thermal breakdown*.

The most essential and characteristic feature of a thermal breakdown is its temperature dependence. When a certain temperature is reached, any further growth in temperature is accompanied by a drop in the breakdown voltage. If we designate the breakdown voltage V_{br} and the amount of evolved heat required for the breakdown to occur Q_{br}, then, according to (4.5),

$$Q_{br} = c \frac{V_{br}^2}{R},$$

and hence the breakdown voltage V_{br} and resistance R are related, i.e.

$$V_{br} = A \sqrt{R},$$

where A is a coefficient we are not interested in at present.

Since the resistivity (the reciprocal of conductivity) and hence the resistance R of dielectrics drops very rapidly with temperature (i.e. the conductivity rises very rapidly, see Section 4.1), the breakdown voltage V_{br} must decrease considerably due to the heating. There is no doubt at all that the breakdown at high temperatures is purely thermal in origin.

If we take a transparent specimen (for instance, a crystal of rock salt), then the process of reaching a breakdown can be followed with a naked eye when the specimen is thick enough. The experiment is especially demonstrative if it is carried out at a temperature for which the specimen does not glow. Heat it a little, and it becomes dark-red and incandescent. First, when the voltage is just applied, the specimen starts glowing with a weak violet light, and some time later part of it becomes dark-red. This red glow, which characterizes the temperature of the more heated portion of the dielectric, becomes more and more intense, while the violet glow becomes weaker. Then all of a sudden there is a bright flare: this is the breakdown, and it produces a white-hot channel in the dielectric.

Another characteristic feature of thermal breakdown is that there is an appreciable dependence of the breakdown voltage on the duration of the voltage application. The process of heating and melting the entire dielectric is so slow that it is only complete after several minutes. That is why considerably greater breakdown voltages can be withstood in short pulses than when the potential difference is applied for a long time.

There is another form of breakdown which can be called *electrical breakdown*.

A dielectric always contains one or more free electrons. They are a 'seed charge' that initiate the development of a breakdown. These electrons roam randomly over the crystal, colliding with the atoms or ions of the lattice. When an external electric field E is applied, the electrons are

accelerated by the field and they accumulate kinetic energy. Suppose T is the kinetic energy that is accumulated between two consecutive collisions. Obviously

$$T = eEl,$$

where e is the absolute value of the electron charge and l is the distance between two collisions.

If this energy is sufficient to ionize an atom (ion), i.e.

$$T \geqslant I,$$

where I is the ionization potential, then one of the electrons belonging to the atom (ion) will be knocked out as a result of a collision. This

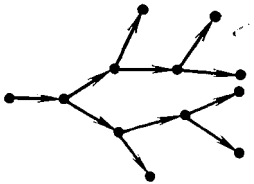


Fig. 37. Electric breakdown: the mechanism of collision ionization

new free electron will in turn ionize the atoms (ions) it encounters, given that the energy it will have accumulated by moving in the field is large enough to do so. In this way the number of free electrons and therefore the electric current in the dielectric can build up like an avalanche; it is diagrammatically presented in Fig. 37. After some time the current can be great enough to cause a breakdown.

Consequently, the basic parameter affecting an electric breakdown is not the potential difference V across the specimen, but the strength of the electric field. When the electric field is strong enough to give an electron the energy to ionize an atom while it is travelling between two collisions, a breakdown will happen. This is called the *collision mechanism* of ionization.

Breakdown voltage V_{br} and *breakdown strength* E_{br} in a uniform field are related as follows:

$$E_{br} = \frac{V_{br}}{d},$$

where d is the distance between the electrodes. Hence

$$V_{br} = E_{br}\, d.$$

Accordingly, the breakdown voltage for an electric breakdown should be proportional to the thickness of the specimen.

If the field is not uniform, and there may be sites within a crystal where the strength of the field is high, a breakdown can occur if the field strength at a site equals the breakdown strength. We can thus assert that the breakdown voltage depends on the heterogeneity of the dielectric.

On the other hand, the electric breakdown voltage, unlike the thermal sort, should not be noticeably dependent on temperature, nor should it be dependent on the duration of the voltage. This is because lattice ionization is instantaneous, so the development of an electron avalanche is an extremely rapid process. Therefore both short and long duration applications of a potential difference should result in more or less the same breakdown voltage.

Now let us consider collision ionization in terms of energy bands, the terms we used while discussing the conductivity of semiconductors (Sections 2.3 and 2.4).

Suppose W is the total energy of an electron in a crystal lattice in the absence of an external electric field. When a field is applied, it will become

$$W + eEx,$$

where eEx is an additional term due to the field. As before, e is the absolute value of the electron charge, E is the electric field strength, and x is the distance between the electron and an origin where the potential energy of the electron in the field is assumed to be zero. The x-axis is assumed to be at right angles to the flat electrodes between which the specimen is placed, and therefore the axis is in the direction of the field.

When the field is absent, the energy bands in (W, x) coordinates are represented by horizontal strips as shown in Fig. 38*a*. When the field is applied, the bands become sloped, as shown in Fig. 38*b*, the gradient being the steeper, the stronger the field. If in the absence of the electric field the forbidden energy band is between $W = W_1$ and $W = W_2$ for all x (see Fig. 38*a*), then in the presence of the electric field it is between

$$W = W_1 + eEx$$

and

$$W = W_2 + eEx,$$

i.e. it is different for different x, in other words it is different at different sites in the crystal

(see Fig. 38*b*). On the other hand, a given energy W is forbidden in the crystal within the interval from $x = x_1$ to $x = x_2$ (Fig. 38*b*) and is allowed outside this interval. The conduction band is

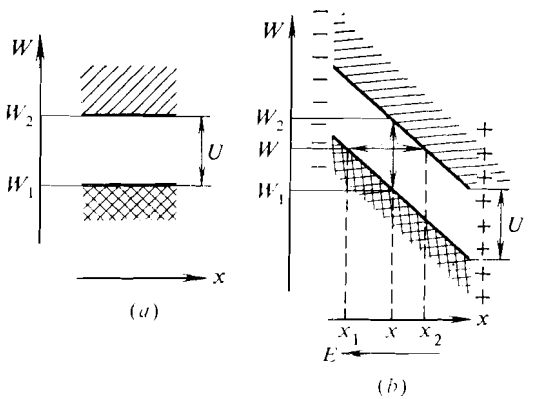


Fig. 38. Energy bands in a crystal: (*a*) in the absence of an electric field, (*b*) in the presence of an electric field

hatched while the valence band is double hatched in Figs. 38*a* and 38*b*.

Suppose there is an occasional free electron that moves in the conduction band in direction AB (Fig. 39). While moving, the electron accumulates a kinetic energy that can be shown in Fig. 39 as a segment T for each x. When T is equal or greater than the energy U (U is the width of the forbidden range between the bands), a free electron in collision with a bound electron in the valence band can push the bound electron into the conduction band by transferring an energy equal to or more than U. There are two

electrons in the conduction band now, each of which moving in the conduction band, accumulating kinetic energy, and able to ionize, i.e. push bound electrons into the free state. The free electron in Fig. 39 moves along the path $ABCD$, and the bound one moves along $B'C'D'$.

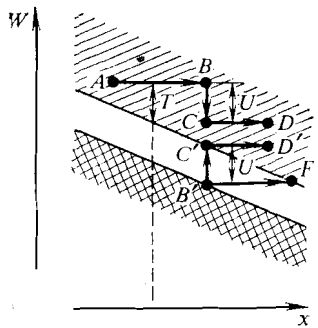


Fig. 39. Mechanisms of collision ionization and cold emission presented in a band diagram

This is the collision ionization mechanism as described in terms of the energy band diagram.

Besides collision ionization, there is also what is called *field ionization*. This is when an electron from the valence band passes into the conduction band directly due to the tunnel effect: this is shown by the arrow $B'F$ in Fig. 39. This mechanism is equivalent to cold emission, which we discussed in detail in the chapter on metals (see Section 1.4). However, and this should be emphasized, the field strengths at which this effect produces noticeable currents commonly exceed the strengths at which breakdowns occur, i.e. an electrical breakdown happens before a cold emission breakdown can occur.

4.3. Crystal Colouration

Suppose we have a transparent crystal, a crystal of rock salt for example. By exposing the crystal to light or putting it into an electric field, we can induce colouration in the crystal. The colouration is an evidence that electrons in the crystal have transferred from one energy level to another. A transition from a lower level to a higher one can take place if a light quantum is absorbed. And vice versa, a transition from a higher level to a lower one is accompanied by the emission of a quantum. The transfer of an electron from a donor localized level into the conduction band means the electron has gone from a bound state into a free state. The transition of an electron from the valence band to an acceptor localized level means a localized bound electron has appeared accompanied by the appearance of a free hole.

The frequencies absorbed by a crystal can be represented as a system of bands in the spectrum. The structure-stable and structure-sensitive bands should be distinguished (see Section 2.1).

Alkali halide crystals strongly absorb light in the remote ultraviolet, i.e. where the frequencies are very high (or the wavelengths are short). This is called *intrinsic* absorption and the frequencies are gathered into an intrinsic absorption band. This is a structure-stable band. Its intensity and position in the spectrum do not depend on any external factor.

It is typical that intrinsic absorption is photoelectrically inactive, i.e. light absorption in an intrinsic band does not induce photoconduction. When a crystal is exposed to frequen-

cies from its intrinsic band, the crystal remains an insulator as before.

Photoelectric inactivity seems to occur because absorption in the intrinsic band is exciton in origin. The absorption of a light quantum results in an electron being moved from a Cl^- ion to an adjacent Na^+ ion (for the sake of certainty, let us consider a crystal of rock salt NaCl, which is built of Na^+ and Cl^- ions). The result is two neutral atoms, Na and Cl, that are in the vicinity of each other and related by Coulomb forces. This is the Wannier-Mott exciton and is an electrically neutral formation that moves through a crystal as an entity without transporting electric charge.

There is another explanation for the photoelectric inactivity of the intrinsic band. The intrinsic band can absorb so much that all the incident light may be absorbed by the surface layer of the crystal without penetrating far inside.

Now let us take the structure-sensitive absorption bands. Lattice structure defects that act as absorption centers are responsible for the bands. We shall consider two main absorption bands: the *U*-band and *F*-band, which correspond to *U*-centers and *F*-centers (from the German words Ultravioletzentren and Farbenzentren, ultraviolet and colour centers). The *U*-band is in the near ultraviolet, at the 'tail' of the intrinsic band, and the *F*-band is in the visible part of the spectrum (Fig. 40). A crystal with an *F*-band is coloured in the literal sense of the word. Thus, the *F*-band in crystals of rock salt is responsible for their light blue colour.

There are various techniques to colour a crystal. One is to heat it for a long time at a high temperature in the vapour of an alkali metal. This is called *additive* colouration. The resulting colour depends on the composition of the hot crystal and does not depend on the alkali metal.

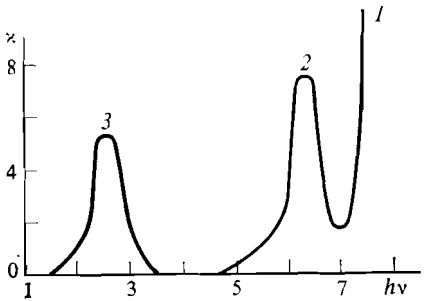


Fig. 40. Absorption bands in a crystal of NaCl: *1*—intrinsic band, *2*—*U*-band, *3*—*F*-band. Absorption constant $\varkappa$ (mm^{-1}) is plotted along the ordinate axis, energy $h\nu$ of the incident quanta (in electron volts) is along the abscissa axis

Another method is to expose a transparent crystal to X-rays or ultraviolet light. This is called *substractive* colouration. The same *F*-band appears in the spectrum of the crystal. If exposed to *U*-band frequencies, the *U*-centers are destroyed and *F*-centers appear. Remarkably the reverse phenomenon also occurs, i.e. exposing the crystal to the *F*-band frequencies attenuates or destroys the *F*-band, creating or amplifying the *U*-band.

It is possible to colour a crystal by putting it into an external electric field. At high temper-

atures, *F*-centers and *U*-centers appear near the cathode and move gradually towards the anode. This is observed directly when a blue colour (in rock salt) is seen to move from the cathode to the anode spreading over a greater and greater portion of the crystal. If the direction of the field is reversed, the *F*-centers will gradually 'return' into the cathode, i.e. the blue colour will be 'drawn' back into the cathode. As to the *U*-centers, they stay at their places, getting stuck, as it were, in the crystal.

When a coloured strip appears following additive or substractive colouration, and then an electric field is applied, the strip starts moving towards the anode. At high temperatures it moves as a single entity, without being deformed and leaving no colour behind it, while at low temperatures the trailing edge moves more slowly than the front one, and the strip gradually spreads. Thus as it moves the intensity of the whole cloud progressively declines due to *F*-centers being transformed into *U*-centers. The current that passes through a crystal in the process increases gradually and becomes constant when the coloured cloud reaches the anode. If the current flows for a long time, the colour of the crystal will gradually disappear uniformly over the entire bulk of it.

If a crystal contains an impurity, for instance a crystal of rock salt includes foreign atoms of gold or copper, then the blue cloud, when it reaches the area containing the impurity, colours the area depending on the impurity (for instance, red or brown). It is remarkable that if the direction of the electric field is reversed, then although

the blue cloud is drawn back into the cathode, the colour brought about by the impurity remains fixed. The crystal can only be discoloured completely by heating it. That there is an impurity in a crystal and what its distribution is over the bulk of the crystal can be determined with the naked eye.

Let us note the influence of temperature on the position and structure of the *F*-band. When a crystal is heated, the *F*-band is deformed so that its maximum shifts towards longer wavelengths and the whole band spreads out. This is illustrated in Fig. 41 for crystals of KBr. The process can be reversed: when the crystal is cooled, the band becomes narrower and its maximum shifts towards the shorter wavelengths.

It is characteristic that the *U*-band, like the intrinsic band, is photoelectrically inactive. Meanwhile, exposing the crystal to the frequencies belonging to the *F*-band influences pronounced photoconduction and the crystal becomes a good electronic conductor. This happens because electrons go from the *F*-level to the conduction band (see Fig. 42).

Let us discuss now the nature of colour centers. An *F*-center in an alkali halide crystal is a halogen vacancy with an electron localized either in or near it, thus maintaining the crystal as a whole electrically neutral. Experience has shown that this combination has an affinity for free electrons. A free electron travelling across the crystal and meeting an *F*-center adheres to it and forms what is called an *F'*-center. Accordingly, an *F'*-center is a halogen vacancy with two electrons localized around it.

A U-center is a negative hydrogen ion H^- which has substituted a halogen vacancy in a lattice point. If the H^- ion is drawn into an interstice, it is called a U_1-center. Then if the ion is neutralized, it becomes a U_2-center. Consequently, a U_2-center is an H atom in an interstice.

There are many possible colour centers other than these centers. We shall only mention V-centers, which are similar to the F-centers

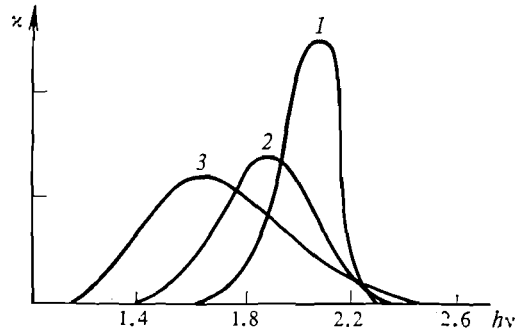


Fig. 41. Deformation of the F-band of a KBr crystal with heating: *1*— −245°C, *2*— +200°C, *3* — +600°C. The coordinate axes are the same as in Fig. 40

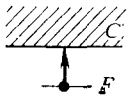


Fig. 42. Origination of photoconduction in the F-band. C—conduction band, V—valence band, F—level of an F-center

though a V-center is a metal vacancy combined with a hole localized nearby.

Many features of crystal colouration have now been established experimentally. However, so far there is no theoretical background that allows us to consider them all from a single standpoint.

4.4. Crystal Luminescence

We are going to discuss in this section the *luminescence* of dielectrics and semiconductors, which is the glow or light emitted from crystals. In order to produce the glow, the crystal has to be excited, i.e. supplied with additional energy that can later be released from the crystal in the form of light quanta. There are various ways of exciting crystals, and so there are the following kinds of luminescence:

(a) photoluminescence (light excitation),

(b) electroluminescence (electric field excitation),

(c) cathodoluminescence (excitation by an electron flux),

(d) chemiluminescence (excitation at the expense of energy produced in a chemical reaction).

Crystals that can luminesce are called *luminophors*. The ability to luminesce is brought about by impurities in the crystals which are called *activators*. Crystal luminophors are called *crystal phosphors*, or just *phosphors*. The spectral composition of the luminescence depends on the nature of the activators, and the radiation intensity depends on their concentration. We should point out that an activator is an impurity in the wide sense of the word, i.e. in the sense

attributed to the term in solid-state physics. Hence activators are not necessarily chemically foreign particles in the crystal. They may also be structural defects, i.e. local imperfections of the regular lattice structure. These include, in particular, vacancies and the lattice's own atoms or ions but existing in the interstices. The structural defects responsible for luminescence are conventionally called *activator atoms* or *luminescence centers*.

A crystal phosphor can be excited by either exciting the luminescence centers or ionizing them.

Let us consider the first case. Suppose a center of luminescence, as often happens, is an *acceptor-donor pair* in the crystal, i.e. a combination of an acceptor and a donor defect in the immediate vicinity of each other. Suppose further that the energy of electron affinity for an acceptor defect is greater than the ionization energy of a donor defect. Then normally an electron will be bound to the acceptor and a hole bound to the donor. Transition of the electron from the acceptor to the donor means that the pair is excited, and the reverse transition of the electron (from the donor to the acceptor), i.e. the return of the pair into its normal state, results in luminescence. Therefore, the excitation and the luminescence are transitions of an electron within the pair from one partner to the other.

Now let us consider a second case, when the excitation of a crystal phosphor is due to ionization of a luminescence center. The crystal structure of a phosphor permits us to describe the ionization and neutralization of a luminescence center in terms of energy bands (see Sections 2.3 and 4.3).

The ionization of an activator during a photoluminescent event can occur due to the absorption of a light quantum directly by a center of luminescence (this is the *impurity absorption*). An electron is thus knocked out of the center (if it is a donor defect) into the conduction band or it is knocked from the valence band to the luminescence center (if it is an acceptor defect). The reverse transitions, which result in the neutralization of the activator, result in the emission of a light quantum i.e. a luminescent event.

The ionization of activators and their subsequent luminescence not only result from light absorption by the luminescence centers directly, but also because of light absorption by the lattice itself, i.e. by the regular atoms (ions) of the lattice (this is called *intrinsic absorption*). An electron is then transferred from the valence band to the conduction band (see Fig. 43, arrow *1*). A hole in the valence band and an electron in the conduction band appear in the process. The return of the crystal to the normal state could then be the result of two consecutive acts, i.e. transition *2* and then transition *3* (see Fig. 43*a*) if the activator is an acceptor, or transition *2'* and then transition *3'* if the activator is a donor (see Fig. 43*b*). Each transition (or all of them together) can be radiative, i.e. can be accompanied by the emission of a light quantum, or luminescence.

A crystal phosphor can have what are called *adhesion centers* in addition to luminescence

centers. These are defects that trap electrons or holes released as the crystal is excited. The centers do not participate directly in luminescence, nevertheless they can play an essential role in the process. They correspond in the energy band spectrum to what are called *adhesion*

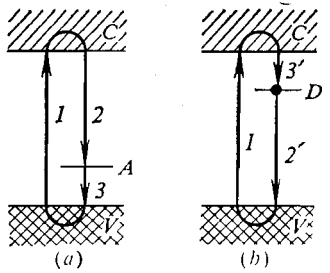


Fig. 43. Appearance of luminescence in intrinsic absorption. V—valence band, C—conduction band. A—acceptor center of luminescence, D—donor center of luminescence. Transition *1* is excitation, transitions *2*, *2'*, *3*, *3'* represent luminescence

levels. They can be acceptor levels (adhesion levels for electrons) or donor ones (adhesion levels for holes). The former are found immediately beneath the bottom of the conduction band, and the latter are immediately above the ceiling of the valence band (see levels T in Fig. 44).

An electron from the conduction band or a hole that appeared in the valence band due to the excitation of the phosphor can be captured by adhesion centers before they manage to recombine, i.e. unite with each other. The accumulation of electrons or holes on the adhesion

levels means transition of the system into an excited energy state (a *metastable state*). Leaving it requires what is called an *activation energy*. Indeed, the return of an electron to the conduction band or a hole to the valence band can

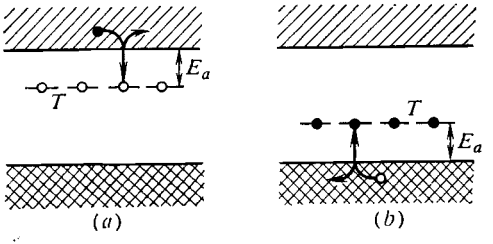


Fig. 44. Adhesion levels T: (*a*) for electrons, (*b*) for holes. E_a—activation energy: filled circles—electrons, blank circles—holes

only happen at the expense of heat or some external effect, for instance, due to exposure to infrared radiation. The location of adhesion levels in the energy band determines the duration of the *afterglow* of the phosphor, i.e. phosphorescence, or luminescence that persists after removal of the exciting source. The concentration of these levels determines the amount of stored light, i.e. the energy accumulated by the phosphor when it was being excited.

Besides luminescence centers and adhesion centers, a phosphor contains, as a rule, some defects that are centers of nonradiative recombination, i.e. events when electrons join holes without radiation emission. These centers are shown in the energy spectrum of a luminophor

as localized levels (see level R in Fig. 45) that can be either donors or acceptors depending on the nature of the centers. Recombination levels (R) are an intermediate stage in the recombination of electrons from the conduction band and holes from the valence band (transitions *1* and *2* in Fig. 45).

Consequently, the free electrons and holes that appeared as a result of transition *1* in Fig. 43 can either recombine through A or D atoms of the activator (transitions *2-3* or *2'-3'* in Fig. 43) and emit quanta, or recombine through R-centers (transitions *1-2* in Fig. 45) without emitting quanta. The two channels compete with each other. Obviously, recombination through R-centers results in luminescent decay (called *extrinsic decay*). Let us remark that recombination through activator atoms (A and D in Fig. 43) is not necessarily radiative. Generally, there is a nonzero probability for a recombination to be nonradiative. In other words, a certain fraction of the recombinations that occur through luminescence centers do not produce radiation (this phenomenon is called *intrinsic decay*).

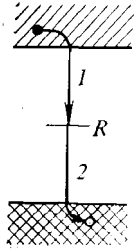


Fig. 45. Recombination level R

So far we have discussed photoluminescence. Now let us turn to electroluminescence for which the mechanism of excitation is quite different. Suppose that a free electron in a crystal has been accelerated in a strong electric field and has accumulated a large kinetic energy. If it collides with an activator atom, it knocks out one of the activator's electrons (the activator is ionized). This is the excitation mechanism in electroluminescence. A liberated electron then falls from the conduction band into the same or similar vacant place in the activator atom (it is neutralized) and a light quantum is emitted in the process (a luminescent event). Therefore, a crystal phosphor is excited in the case of electroluminescence by the mechanism of collision ionization, which we considered in Section 4.3.

In conclusion, let us take chemiluminescence. We shall only discuss a particular case of chemiluminescence, i.e. radical recombination luminescence, which was discovered and studied only recently. Suppose two atoms or two radicals have arrived at the surface of a crystal from the gas environment of the phosphor, and they meet. Recall that a radical is a molecule with an 'unsaturated' ('free') link. When the two atoms or two radicals join, they produce a 'valence-saturated' molecule. As a rule, a certain amount of energy is released in the process which can be spent to excite the phosphor and which subsequently appears as chemiluminescence. Consequently, this phosphor excitation and luminescence occur by utilizing the energy released in a chemical reaction that takes place on the crystal surface.

4.5. Electrets

Some dielectrics can have a special state called the *electret state*. This is characterized by the presence of an internal electric polarization caused by an external electric field and persisting for a long time after the field is removed. A dielectric in the electret state is called an electret. An electret is a charged dielectric with the poles apart: the electric analogue of a magnet. Naturally, this analogy is limited, since free magnetic charges have not so far been discovered, while free electric charges do exist.

If there is a negative charge on the surface of an electret facing the anode and a positive charge on the surface facing the cathode, the charge is called a *heterocharge*. In case of the opposite polarity, when the positive charge faces the anode and the negative charge faces the cathode, we speak of there being a *homocharge* in the dielectric. It is remarkable that in some cases an electret can undergo a spontaneous reversal of its polarization. After the polarizing field is removed, a heterocharge sometimes gradually drops to zero and then transforms into a homocharge that gradually increases to its maximum value. The lifetime of an electret state is limited though it varies over a wide range. Sooner or later the dielectric becomes depolarized, and the electret state disappears. However, the electret state can last several years.

An electret state that appears after the dielectric has been heated and then cooled in a strong electric field is called a *thermoelectret state*, and the respective electrets are called the

thermoelectrets. It has been established that the electret effect is a bulk and not a surface effect. Hence cutting or scraping off layers from the surface of an electret does not change its properties.

The *photoelectret state* deserves special attention. It appears when a photoconducting dielectric is put in a strong electric field and exposed to light. A bulk charge accumulates in the dielectric due to conduction and after both the field and light have been removed, the bulk charge persists in the dielectric for a long time. Dielectrics that can be in a photoelectret state are called photoelectrets.

The photoelectret state can be observed both in polycrystals and in a large number of monocrystals. Naturally, the mechanism leading to a stable polarization in photoelectrets is different from that in thermoelectrets. A prerequisite for the photoelectret state to appear is the presence of photoconduction in the dielectric. We shall only discuss photoelectrets below.

The appearance of a stable heterocharge is characteristic for photoelectrets. As for a homocharge, it is not usual in photoelectrets. The evolution of a heterocharge starts when a light quantum pushes an electron from the valence band or from an activator level to the conduction band. Recall that an activator level is a donor localized level in the forbidden range between the energy bands. If an electron is taken from an activator level, the activator atom remains without its electron, i.e. a hole is localized on the activator atom. If an electron is pushed from the valence band, a free hole appears that can

roam over the crystal and meet an activator atom with whose electron it recombines. In the long run, the result is the same: a hole localized on an activator atom.

As for the free electron that appears in the conduction band, it moves somewhat towards the anode because of the external electric field's influence and then gets stuck on the adhesion level, which is an acceptor localized level under the conduction band but above the activator level. We can obtain crystals in which the lower levels (activator levels) are devoid of electrons and the higher levels (adhesion levels) are full. This crystal is said to be in a *metastable state*, i.e. a state from which a transition to the normal state requires energy (activation energy) (see Section 4.4). The adhesion levels correspond to lattice defects that are found in a narrow region close to the electrodes. Those that are close to the anode are partially filled by electrons and yield a negative bulk charge. The crystal can thus be regarded as a dipole with oppositely charged poles, one of which contains an excess of electrons and the other an excess of holes, or, to put it better, too few electrons.

The depolarization of a photoelectret, i.e. the removal of the photoelectret state, can be brought about in many ways.

For instance, it is possible to remove the photoelectret state by exposing the dielectric to light quanta that knock electrons from the valence band to the vacant activator levels. The result is that the holes localized on the activators are transferred to the valence band, i.e. into the free state. Then these free holes recombine

with electrons on the adhesion levels, i.e. the electrons responsible for the photoelectret state. The result is that these electrons leave the adhesion levels, and the dielectric is depolarized.

Another way of depolarizing an electret is to expose it to light that will remove electrons directly from the adhesion levels and transfer them to the conduction band. Sometimes this process occurs with a transitory stage. An adhesion center is excited by the light and then the electron is knocked out of the excited state into the conduction band. The 'extra excitation' takes place when the photoelectret is heated. Depolarization in this case requires both light and heat at the same time. Let us remark that heating can be replaced by infrared radiation.

The adhesion levels responsible for the photoelectret state are, in fact, localized levels that are well below the conduction band. Some shallow adhesion levels can also exist in the crystal immediately below the conduction band and, very much like the adhesion levels, they are involved in photoluminescence (see Section 4.4). These levels also can play a role in electret phenomena, being responsible for what is called *dark polarization*, i.e. the polarization that appears in a dielectric when it is placed in an electric field although not exposed to light. While in the absence of a field the adhesion levels are filled uniformly, a field, displaces the conduction electrons which, in turn, results in their being unevenly distributed among the adhesion levels. It is this uneven distribution of localized electrons that is seen experimentally as dark polarization.

Photoelectret excitation is when short wavelength light pushes electrons from the valence band or from activator levels to the conduction band, and then the electrons start filling the adhesion levels, both the deep ones, which are responsible for the photoelectret state, and the shallow ones, which are responsible for dark polarization.

4.6. Dielectric Constant

The dielectric constant ε is another essential parameter of dielectrics, alongside electrical conductivity (or specific conductance) σ which we discussed in Section 4.1. The dielectric constant is the ratio of the forces between charges in a vacuum to the forces between the same charges, at the same distance apart, in the dielectric. Its value differs for different dielectrics and usually varies from 1 to 100.

However, some dielectrics have unusually high values of ε. One such dielectric is Seignette salt, or potassium sodium tartrate. Its dielectric constant at room temperature is 10,000. Dielectrics with unusually large values of ε are called *ferroelectrics*, or seignette-electrics and we shall discuss their properties in the next section.

The dielectric constant ε of most dielectrics changes little upon heating and is practically independent of the electric field strength.

The ε-fold decrease of the forces between charges in a dielectric and those in a vacuum is caused by what is called dielectric *polarization.* This phenomenon is characteristic for dielectrics, -unlike electric conduction which occurs in crys

tals of all three types, i.e. metals, semiconductors, and dielectrics. As mentioned in Section 4.1, the conductivity of a dielectric is much less than that of a semiconductor or a metal because dielectrics do not contain significant numbers of free charged particles whose directed motion in an electric field constitutes the electric current.

Let us discuss dielectric polarization in detail. Suppose we have a dielectric in the form of a rectangular parallelepiped. Suppose it is in the uniform field of a flat condenser, one of whose plates is charged positively, the other negatively, and the base of the parallelepiped is parallel to the condenser plates. Thus the electric field in the condenser is perpendicular to the base of the parallelepiped.

The positive and negative charges in the molecules move apart a small distance due to the electric field. Each molecule thus turns into an electric dipole that can be represented as two point charges, $+q$ and $-q$, which are rigidly connected to each other at a distance of d.

The electric moment of the dipole is a vector $\mathbf{p}$ equal to the product of the charge q and the displacement vector $\mathbf{d}$, which points from the negative charge to the positive one, i.e.

$$\mathbf{p} = q\mathbf{d}.$$

If there are already polarized molecules in a dielectric, they are turned and oriented by the field.

The displacement of charges within the dielectric molecules and the orientation of the dipole molecules in the electric field bring about the appearance of bound surface charges. The

dielectric face touching the positive plate carries a charge $-Q$, and the one touching the negative plate carries a charge $+Q$*. The positive and negative charges compensate each other within the dielectric.

The process leading to the appearance of bound charges with a dielectric is called *polarization.*

Polarization is quantified by its *polarization vector* **P**, which is the electric moment of a dielectric per unit volume. The vector points in the same direction as the electric field, and its absolute value equals the sum of the projections of the electric moments of all the elementary dipoles per dielectric unit volume onto the field direction.

The electric moment M of the entire dielectric is the product of P and the volume of the dielectric:

$$M = PSh,$$

where S is the area of the base of the parallelepiped, and h is its height.

On the other hand, M is the product of the surface bound charge Q and the height of the parallelepiped:

$$M = Qh.$$

Therefore

$$PSh = Qh, \quad P = \frac{Q}{S}.$$

* Do not mix this up with the Joule heat designated by the same letter Q in Section 4.2.

Consequently, the modulus of the dielectric polarization is equal to the surface charge density Q/S

The charge that appears on the surface of a dielectric due to its polarization produces in the dielectric a field $\mathbf{E}_1$ that is opposite in its direction to the external field and equals in the SI units

$$E_1 = \frac{Q}{S\varepsilon_0} = \frac{P}{\varepsilon_0},$$

where the electric constant $\varepsilon_0 = 8.85 \times 10^{-12}$ C^2/(N·m^2) $= 8.85 \times 10^{-12}$ F/m.

The field strength E in a dielectric is equal to the difference between the field strength E_0 in the absence of the dielectric and the strength E_1 of the polarization field, i.e.

$$E = E_0 - E_1 = E_0 - \frac{P}{\varepsilon_0}. \tag{4.6}$$

The ε-fold reduction in the force between charges in a dielectric is tantamount to an ε-fold reduction in the electric field strength when the dielectric is placed in the field:

$$E_0 = \varepsilon E. \tag{4.7}$$

Substituting (4.7) into (4.6), we can find the dependence between the polarization P and the electric field strength E in a dielectric:

$$P = \varepsilon_0 (\varepsilon - 1) E. \tag{4.8}$$

Hence, the modulus of the dielectric polarization is proportional to the strength of the electric field. The proportionality factor in (4.8)

is commonly designated as $\varkappa$:

$$\varkappa = \varepsilon_0 (\varepsilon - 1), \tag{4.9}$$

and is called the dielectric susceptibility (electric susceptibility). It does not in practice depend on E and can be considered a property of the material.

Equation (4.9) relates two of the parameters of a dielectric: $\varkappa$ and ε.

We have only discussed the simplest mechanisms of polarization, but in real dielectrics polarization is more complex. It occurs due to a limited displacement of the charged particles in dielectrics due to the effects of both external and internal fields. A result is that each element of the dielectric's bulk, within which the displacement took place, aquires a nonzero electric moment. There are different types of polarization depending on the particles involved and the actual displacements.

The particles that can be displaced or oriented in an electric field fall into two categories, i.e. *elastically bound* and *weakly bound* particles. The displacement of an elastically (strongly) bound particle is opposed by an elastic force that tends to return it to the equilibrium position around which the particle oscillates thermally. An elastically (strongly) bound particle is shifted a small distance from the equilibrium position by an electric field. A weakly bound particle has more than one equally probable equilibrium positions in the absence of an electric field. Thus while it oscillates around one equilibrium position it can be moved to another position due to thermal fluctuation. When an electric field

is applied, the probability of the transition increases: for a positively charged particle in the direction of the field, and for a negatively charged particle in the opposite direction. Oppositely charged particles are shifted farther apart when they are weakly bound than when elastically bound. Two types of polarization correspond to these two kinds of particles, i.e. *elastic* and *relaxation* (or thermal) *polarization.*

The polarization does not occur the moment an electric field is applied, it takes some time to be completed. The time for relaxation polarization to build up is essentially longer than for elastic polarization to build up. The following types of polarization are commonly distinguished.

Electronic elastic polarization. This is the polarization that results from the displacement of the electron shells of atoms, molecules, and ions with respect to heavy 'fixed' nuclei to a distance that is less than the dimensions of the atoms, molecules, or ions. It occurs in all dielectrics for any state of aggregation, and takes 10^{-14}-10^{-15} s, to build up.

Ionic elastic polarization. This occurs in ionic crystals. It is brought about by the elastic displacement of positive and negative ions from their equilibrium positions to a distance less than the distance between adjacent ions. It takes 10^{-11}-10^{-14} s to build up.

Dipole elastic polarization. This occurs in molecular crystals which consist of dipole molecules. Since these molecules are fixed and cannot rotate freely, they can only turn through a small angle due to the effect of an electric field. This type of polarization builds up relatively slowly.

Electronic relaxation polarization. This occurs in certain crystals in which the electrons can move in the vicinity of the defects to which they are attached. They move a small distance of the order of several atomic spacings and the polarization builds up relatively slowly.

Ionic relaxation polarization. This occurs in a number of ionic crystals. It is brought about by the displacement of weakly bound ions from their equilibrium positions to a distance of the order of several atomic spacings. The polarization builds up relatively slowly.

Dipole relaxation polarization. This occurs in dipole dielectrics in their liquid and gaseous states. The molecules can then rotate freely around their equilibrium positions. They are oriented randomly in the absence of an electric field, but polarization appears in a field due to the existence of a prevailing orientation of the dipoles. The build-up time is of the order of 10^{-10} s or more.

Structural polarization. This includes certain kinds of polarization that are due to the appearance of bulk charges in the dielectric and is closely related to the process of electrical conduction. It occurs when free current carriers move under the influence of an electric field, but are hindered by some obstacles and so not all the carriers reach the electrodes or those that do approach them cannot be fully discharged. Bulk charges thus appear near the electrodes in homogeneous dielectrics (the *high-voltage polarization* we have already discussed in Section 4.1). The advance of the free charges can be hindered by defects of the crystal lattice that capture the charge carriers. In heterogeneous dielectrics, the charges can accumulate at the interfaces between phases (*interlayer polarization*). It is characteristic of structural polarization that the charges are displaced much farther than in other types of polarization, and the build-up time is very great, sometimes a dozen minutes or more.

Spontaneous polarization. This occurs in some dielectrics in the absence of an electric field. Ferroelectrics are examples.

Persistent polarization. Given certain conditions, the polarization persists for a long time after the removal of the electric field in some dielectrics. These are called electrets and have low electrical conductivity (see Section 4.5).

As we mentioned above, after the electric field has been set up not every type of polarization builds up at the same rate. If a dielectric is in an alternating electric field, only those types of polarization whose build-up time τ is less than half the period T of the electric field can be established. As a result the dielectric constant ε becomes dependent on the electric field's frequency, the dielectric constant falling as the frequency increases.

A phase difference between polarization and the electric field occurs at $\tau \geqslant T$. This phase difference is brought about because the moving particles lag behind the forces that cause the motion. Particle movement in the dielectric is accompanied, as it were, by 'friction'. This results in the dielectric becoming heated and therefore in a loss of energy. The energy dissipated per unit time in a dielectric in an electric field is called the *dielectric loss*, or dielectric absorp-

tion. The loss due to electrical conduction, which is accompanied by the evolution of Joule heat, is distinguished from the loss due to a lag in polarization, which is called the relaxation dielectric loss.

4.7. Ferroelectrics and Piezoelectrics

Ferroelectrics are crystal dielectrics that possess, in a certain temperature range, a number of unusual properties.

These properties were discovered for the first time in Seignette salt, and hence ferroelectrics are sometimes called seignette-electrics. Crystals of Seignette salt have clearly pronounced *anisotropy*, i.e. their properties depend on the orientation of the crystals with respect to the direction of the electric field. The unusual properties can only be observed when the crystal is in a special position with respect to the condenser plates. In any other position the crystal will have the same properties as ordinary dielectrics.

Ferroelectrics include such dielectrics as potassium phosphate, potassium niobate, and barium metatitanate. Over a hundred ferroelectrics are known at present.

The primary feature of a ferroelectric is the anomalous dependence of its dielectric constant on temperature. We noted in Section 4.6 that the dielectric constant ε of common dielectrics changes very little with heating. The dielectric constant temperature diagrams for a ferroelectric has one or more very sharp maxima where ε can reach several thousands. The temperatures at which these maxima occur are called Curie temperatures, or Curie points. For instance, Seignette salt has two Curie temperatures (+22.5°C and −15°C), while barium metatitanate has only one Curie temperature (+80°C). If a ferroelectric has just one Curie temperature, its anomalous properties show up at temperatures below it, whereas for one with two Curie temperatures, its properties are anomalous in the temperatures between the Curie points.

The second feature of a ferroelectric is a nonlinear dependence of its polarization on the strength E_0 of the external electric field. It follows from Equations (4.7) and (4.8), that the polarization P is proportional to E_0 in ordinary dielectrics. The relationship in ferroelectrics is more complex in nature, being different in different ferroelectrics. The dielectric parameter ε is then clearly not a constant but depends on E_0.

The third and the most remarkable feature of ferroelectrics is that the dependence of the polarization on E_0 has the appearance of a hysteresis loop, like the magnetization curve of a ferromagnet. Suppose a ferroelectric that is not polarized is put into the electric field between two capacitor plates and the strength of the field is increased. The polarization P will rise nonlinearly and reach saturation at a certain value of P_s. The polarization will not change even if E_0 is increased further. If the field in the capacitor then decreases, the P versus E_0 graph will not be the same as the curve obtained for rising E_0: the new curve will lie above the old one. When E_0 is down to zero, the polarization will be nonzero and equal to a certain value P_0 that is called the *residual polarization*. The resi-

dual polarization persists in the ferroelectric even in the absence of an external electric field. To reduce the polarization to zero, a field E_c must be applied in the opposite direction; this field is called an electric *coercive force*.

There is a clear similarity between the electric properties of ferroelectrics and the magnetic properties of ferromagnets, which is exactly why the dielectrics are called ferroelectrics.

The behaviour of ferroelectrics can be explained as follows. Strong interaction between the particles making up the crystal results in the appearance of separate *domains* within the ferroelectric, domains being regions with nonzero polarization. These domains were first discovered in barium metatitanate, and they can be seen through a polarizing microscope.

The polarization vector has different directions in different domains, so that the electric moment of the whole specimen is close to zero. This corresponds to the lowest energy state, since otherwise, polarization charges would appear on the surface of the specimen, and a field would be induced around it containing additional energy.

The polarization directions of individual domains change in the external electric field to the direction of the field. In addition, the most advantageously oriented domains can grow by pushing their boundaries into neighbouring domains. The resulting electric moment of the ferroelectric becomes nonzero. The presence of polarized domains in ferroelectrics has an essential effect on their electric properties.

The domains only exist within a certain temperature range where the ferroelectric properties appear.

So far we have only considered cases when the polarization of a dielectric was brought about by an asymmetry in the localization of positive and negative charges and was caused by an external electric field. This asymmetry can be produced in some crystals by deforming them mechanically. If such a crystal is compressed or stretched in a certain direction, it becomes polarized, and polarization charges appear on its surface. They can then be registered, for instance, by an electrometer. This phenomenon is called the *piezoelectric effect*. The electric moment that appears in the crystals due to mechanical deformations depends on the crystal lattice direction along which the deformation occurs and on the crystal lattice type. The piezoelectric effect only occurs in crystals whose structure does not possess a center of symmetry.

Crystals that exhibit the piezoelectric effect are called *piezoelectrics*. They include Seignette salt, ammonium phosphate, potassium phosphate, ethylenediamine tartrate, quartz, and tourmaline.

Let us discuss the piezoelectric effect in some detail, taking quartz as an example. Quartz crystals have four symmetry axes. An axis is a symmetry axis when a rotation of the crystal around it through an angle $\alpha = 360°/n$ brings the crystal into a position indistinguishable from its original position. If $n = 2$, the symmetry axis is said to be a two-fold axis, if $n = 3$, it is a three-fold axis, etc. Quartz crystals have one three-fold symmetry axis called the *optic*

axis and three two-fold axes called the *electrical axes*. The latter are at right angles to the optic axis and at 120° to each other.

Suppose we have a specimen in the form of a parallelepiped shown in Fig. 28 cut from a quartz crystal so that its z-axis is the optic axis of the crystal, and its x-axis is one of the electric axes.

If we apply a force f_x to the specimen compressing it in the direction of the x-axis, polarization charges $+Q$ and $-Q$ will appear on the faces perpendicular to this axis. The value of Q is proportional to the compressing force f_x and does not depend on the specimen's dimensions, i.e.

$$Q = kf_x, \tag{4.10}$$

where k is the *piezoelectric constant* of the material. This is called *longitudinal piezoelectric effect*.

However, if we apply a compressive (or tensile) force f_z to the crystal in the direction of the z-axis, no polarization will occur.

If a compressive force f_y is applied in the direction of the y-axis, polarization charges will appear like those in the first case and on the faces perpendicular to the x-axis. However, the signs of the charges will be opposite to the ones in the first case. The charges will also be proportional to the value of compressive force f_y, but now they also depend on the dimensions a and b of the specimen along the x- and y-axes respectively, i.e.

$$Q = k\frac{b}{a}f_y,$$

where k is the same constant as in (4.10).

This is called *transverse piezoelectric effect*. It appears because the force compressing the specimen along the y-axis causes a secondary expansion of the specimen along the x axis.

If the force applied to the crystal is not compressive but tensile, the signs of the polarization charges change.

The *reverse piezoelectric effect* occurs as well. If an electric field is applied to metal electrodes fixed onto opposite faces of a crystal, the dimensions of the crystal change a little, the crystal becoming longer or shorter depending on the direction of the field.

The inevitability of the reverse piezoelectric effect can easily be shown to follow from the law of the conservation of energy. Suppose a compressive force f is applied to a dielectric specimen. When there is no piezoelectric effect, the work of the external forces is equal to the potential energy of the specimen that is elastically compressed. If there is a piezoelectric effect, polarization charges appear on the specimen faces, and an electric field appears that contains an additional energy. According to the law of the conservation of energy, the work performed when the specimen is compressed will be greater, and therefore an additional force counteracting the compression will appear. This is the force of the reverse piezoelectric effect. It acts in a direction opposite to that of the force of the direct piezoelectric effect, i.e. if the charges $+Q$ and $-Q$ appear on the specimen faces *in compression*, the specimen will *expand* if the same charges appear due to an applied electric field, and vice versa.

It follows from what we have said in this and all the preceding sections that the properties of crystals and their behaviour under the influence of external factors are determined by the behaviour of the electrons in the crystals. These electrons are the true masters of crystals.

Remarks in Conclusion: Theory and Experiment

To conclude this book, let us glance at the ground we have covered. We have presented a variety of experimental facts and patterns and interpreted some of these patterns theoretically. The union of theory and experiment is the wonder-working alloy of which the modern physics is constructed.

Experiment is the culture medium in which theory sprouts. The theory is like theatre floodlights for the experiment. A theory divorced from experiments is nonsense. An experiment without a theory is a blind alley. Experiment is the origin we always come back to, and it is the natural judge of the theory. The theorist always follows the experimenter while at the same time showing him the path. But we must not forget that nature is far more complicated than any theory and it will never tire of challenging our perception.

Any physicist, like any researcher, is forever confronted by 'how' questions and 'why' questions. How does a physical phenomenon develop, and what are the laws it obeys? Why does it develop the way it does, and not in another way? To answer a 'how' question is to *describe* the phenomenon. To answer a 'why' question is to *explain* it. The first sort of questions is answered by the experimental physicist, and the theoretical physicist is called in to answer the second one. In the course of time some of 'how' and

'why' questions are answered but then new 'how' and 'why' questions emerge. This is the cause of the continuous evolution of science, 'the daughter of astonishment and curiosity'. I will be very glad if the feelings of astonishment and curiosity are aroused in you as you read this book.

To the reader

Mir Publishers would be grateful for your comments on the content, translation, and design of this book.
We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics