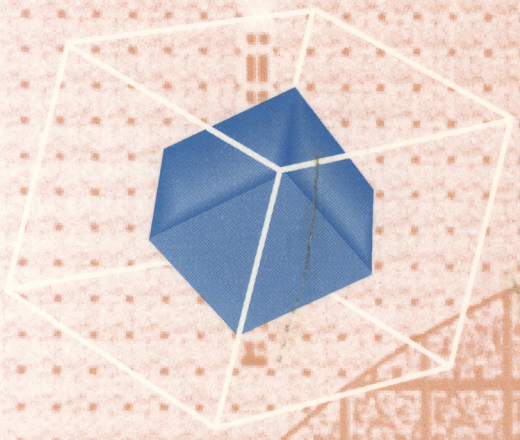
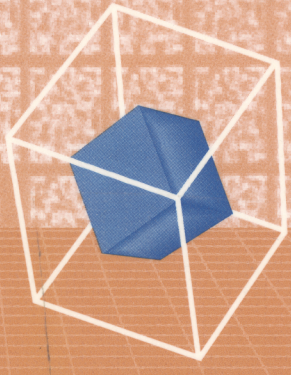
PHYSICS OF SEMICONDUCTORS

B. Sapoval • C. Hermann





Springer-Verlag

Physics of Semiconductors

B. Sapoval C. Hermann

Physics of Semiconductors

With 157 illustrations

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

B. Sapoval
C. Hermann
Laboratoire de Physique de la
Matière Condensée
Unité de Recherche Associée du
Centre de la Recherche Scientifique
Ecole Polytechnique
91128 Palaiseau Cedex
France

Translator
A.R. King
Department of Astronomy
University of Leicester
Leicester LE1 7RH
United Kingdom

French edition published by Ellipses, Paris, 1988.

Library of Congress Cataloging-in-Publication Data
Sapoval, B.
[Physique des semi-conducteurs. English]
Physics of semiconductors B. Sapoval, C. Hermann.
p. cm.
Translation of: Physique des semi-conducteurs.
Includes bibliographical references and index.
ISBN 0-387-94024-3 (New York). – ISBN 3-540-94024-3 (Berlin)
1. Semiconductors I. Hermann, C. II. Title.
QC611.S2613 1993
537.6'22 – dc20 93-286

Printed on acid-free paper.

Production managed by Henry Krell; manufacturing supervised by Jacqui Ashri.
Photocomposed copy prepared from the authors' LaTeX files.
Printed and bound by R.R. Donnelley & Sons, Harrisonburg, VA.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-94024-3 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-94024-3 Springer-Verlag Berlin Heidelberg New York

Preface

The discovery of semiconductors is one of the great scientific and technological breakthroughs of the 20th century. It has caused major economic change, and has perhaps changed civilization itself. Silicon, for example, now plays as important a role in our lives as carbon did in the 19th century. Most of the information technology depends on the properties of semiconductors. One can only be struck by the contrast between our world and one without transistors, computers, rockets, medical image processing and heart pacemakers.

We can see that this development is built on the combination of new and old concepts: miniaturization and printing. If we wish to handle information by a machine, it is clear that the machine's "moving parts" must be as small as possible. Here the parts are the electrons. The invention of the transistor at the end of the 1940s, that made use for the first time of the physics of semiconductors, was the key to miniaturization.

Producing such a machine was difficult because of its small size, and it was not feasible to produce on a mass scale. The introduction of planar technology at the beginning of the 1960s changed that situation. It allowed for the use of photogravure techniques that resemble the printing process. Instead of having to link components one by one, like the individual letters were before Gutenberg, we can now make an entire machine such as a microprocessor through a limited number of processes. The very cheap mass production of these machines has begun to cause industrial and cultural changes that stem from, and are limited by, the physics of semiconductors.

Why should we teach the physics of semiconductors in a course at the Ecole Polytechnique? The main reason is that it applies the most fundamental concepts of quantum and statistical mechanics. We hope to show that it is possible to use these concepts easily so as to meet the needs of the engineer. For this reason, several devices that make use of this physics are described. We give a simple explanation of the principle of the most common systems based on semiconductors.

The appendices that we have included serve two distinct functions: they may give detailed justifications for results given in the main text or illustrate various applications. This book can therefore be used in either an elementary or a more advanced manner. In the first form it is at the level of the second cycle of the Ecole, while the more advanced form is at the third cycle in French universities.

The contents of this book are more in depth than what is currently taught at the Ecole Polytechnique. This follows the tradition of the courses at the Ecole that provide the engineer of tomorrow with a scientific basis for much of his career.

This work owes much to the remarks and criticisms of Yves Quéré, Henri Alloul, Hervé Arribart, Henri-Jean Drouhin, Guy Fishman, Georges Lampel, Gilles de Rosny, Jacques Schmidt and Claude Weisbuch. Some of the problems are based on work by Hervé Arribart, Maurice Bernard, Jean-Noël Chazalviel and Georges Lampel. Very special thanks are due to Jean-Noël Chazalviel and François Ozanam for their numerous comments.

The French version of this book was produced by the Ecole Polytechnique Press and published by Edition Marketing Ellipses (Paris).

Contents

1.

Simple Ideas about Semiconductors

1.1 Definition and Importance of Semiconductors

The solids known as semiconductors have been the subject of very extensive research over recent decades, not simply because of their intrinsic interest but also because of ever more numerous and powerful applications: rectifiers, transistors, photoelectric cells, magnetometers, solar cells, reprography, lasers, and so forth.

A main feature of many of these applications is the possibility of miniaturization of the devices. Miniaturization is more than a convenience: if we are faced with the problem of coding and transmitting messages from a satellite, the complete system of computer and transmitter must be made small. It must work properly for long periods without maintainance. The power available on board the satellite must come from radiation, the only source possible in space. Semiconductor devices, transistors, and solar cells provide solutions to these problems. Similarly the electronic components of a heart pacemaker have to consume little power and be very small. But the most spectacular and most important application of semiconductors is the development of information technology. These developments have only been possible because of the miniaturization of the logic elements allowing the construction of compact systems with great computing power or memory.

Miniaturization has become possible through the perfection of "planar" fabrication techniques. These allow "integration" of circuits and thereby the production of devices containing thousands of elements on a few mm^2. All this industrial development has come into existence only because physics allows us to understand the specific properties of semiconductors, and then use this understanding to create "electron machines" in the form of semiconductor devices.

Semiconductors, as we shall see, are insulators whose "forbidden bands" or "gaps" are sufficiently narrow that thermal excitation allows a small number of electrons to populate the "conduction band." The working element in a semiconductor is this small number of electrons. It is clear that

this small number of electrons can be influenced by a small number of chemical impurities or even by the surface of the crystal.

This sensitivity hindered the understanding of the properties of these materials for a long time. For several decades the crystals studied were not pure enough, so that a number of their properties appeared to be impossible to reproduce with other apparently "identical" crystals. Thus the development of semiconductor physics has had to await progress in semiconductor chemistry, and indeed the chemistry of solids in general. Semiconductors are now the purest and most reproducible solids we can make. The techniques perfected in their manufacture are frequently applied in other branches of chemistry and solid-state physics. There is a very close connection in this subject between industrial requirements, control of materials, and the understanding of the phenomena.

The definition of semiconductors as "insulators with narrow forbidden bands" should be supplemented by a description of the essential physical properties of these materials, namely:

1. their resistivity decreases as the temperature rises, at least for a certain temperature range, unlike metals;

2. semiconductors are sensitive to visible light but transparent in the infrared. When irradiated their resistivity decreases. If they are inhomogeneous, an induced electric field may appear;

3. they often give rise to rectifying or non-ohmic contacts;

4. they exhibit a strong thermoelectric effect, i.e., an electric field induced by a temperature gradient;

5. their resistivity lies between 10^{-5} and 10^{6} ohm·cm.

The materials possessing such properties are the elements of column IV of the Periodic Table, silicon and germanium; III – V compounds of the type GaAs, GaSb, InSb, InP, and so forth; IV – VI compounds such as PbS, PbSe, PbTe, and so forth; II – VI compounds such as CdSe, CdTe, and Cu_2S; ternary compounds such as $Al_xGa_{1-x}As$; and quaternary compounds.

There are several important dates in the history of semiconductor physics.

1897: Discovery of the Hall effect: When a magnetic field is applied to a conductor carrying a current perpendicular to the field, an electric field appears in the direction perpendicular to the current and the magnetic field. The strength of the electric field allows one to measure the number of mobile charge carriers carrying the electric current. Measurements made at the beginning of the century show the existence of a small or very small number of mobile charges varying from sample to sample in an apparently incoherent way. This number is of the order of 10^{-3} to 10^{-7} per atom. The sign of the Hall electric field also allows one to determine the sign of the charge carriers. Surprisingly, in certain crystals this sign was observed to be positive, suggesting that these charges were cations. However, the observed mobilities were very large, much greater than the mobilities of

ions in liquid electrolytes. Indeed, they were comparable to those measured in apparently identical crystals but in which the negative sign of the mobile charges indicated that the carriers were electrons. This was an unexplained paradox.

1926: Bloch's theorem, with its fundamental consequence: a Bloch wave packet can traverse a perfect crystal without colliding with the crystal ions. Collisions result only from crystal defects or vibrations in a perfect crystal. This idea allowed an understanding of the large mobilities observed for electrons.

1931: Wilson lays the foundations of the modern theory of semiconductors as insulators with narrow forbidden bands and introduces the idea of a hole.

1948: Discovery of the transistor effect by John Bardeen and William H. Brattain.

1960: Appearance of planar technology.

1982: World production of 3×10^{13} binary units of active memory in the form of 64 kilobyte (512 kilobit) units alone.

1990: Manufacture of Dynamical Random Access Memories ("DRAM") of 4 megabits per chip.

1991: High Definition Television camera with 2 Megapixel Charge Coupled Device Sensor.

1992: Semiconductor component sales, worldwide: \$60 billion.

1993: World production of transistors: 2.10^{17}

1995: Manufacture of 64-megabit DRAM (estimated).

1.2 A Chemical Approach to Semiconductors

Even though some properties of semiconductors were discovered experimentally in the course of the 19th century, an understanding of the origin of this behavior had to await the advent of quantum mechanics.

The first, classical, theory of electrical conductivity in solids, proposed by Drude in 1900, assumed the current to be transported by a fixed number of electrons that behave like classical particles obeying Maxwell–Boltzmann statistics. In the presence of an applied electric field the electrons attain a velocity proportional to the field (Ohm's law) as they constantly undergo collisions that brake their motion (see Chap. 5). While a number of properties of metals could be understood, nothing in this model predicted the increase in the number of charge carriers with temperature in semiconductors observed via the Hall effect. One might appeal to thermal ionization of the electrons from individual atoms of the solid, but since the ionization energies are of order 10 eV, this effect is too weak at room temperature to account for the observed concentrations.

In any case, the Drude model had no explanation at all for the fact that in some samples the mobile charges were positive.

1.2a The Contribution of Quantum Mechanics

We know that in a physical system, here a solid, electrons occupy stationary energy levels which are the solutions of the Schrödinger equation

$$\mathcal{H}\psi = E\psi, \tag{1.1}$$

where the Hamiltonian

$$\mathcal{H} = \frac{p^2}{2m} + V(\mathbf{r}) \tag{1.2}$$

takes account of both the kinetic energy $p^2/2m$ of the electrons, and the potential $V(\mathbf{r})$ of their interaction with the ions of the solid. Here ψ is the wave function and E the energy associated with ψ. The mass of a free electron is m.

The electrons have half-integer spin and are thus fermions: at most two electrons of opposite spins can occupy each orbital state ψ satisfying Schrödinger's Eq. (1.1). In a solid in thermal equilibrium at temperature T, the energy levels must be populated according to Fermi–Dirac statistics. To understand the properties of semiconductors (or solids in general) we therefore have to first find the energy levels satisfying Eq. (1.1). The second step is to find the state of the system at temperature T by populating these levels according to Fermi–Dirac statistics. We can then examine the properties of this system, which is the aim of this book. This procedure is simple in principle, but its implementation encounters considerable difficulties, and we have to resort to approximations.

1.2b Qualitative Description

One can imagine two main ways of understanding at least qualitatively the properties of quantum electron states in solids. We can think of the crystal constructed from atoms, which we bring together by introducing a coupling between them: this is the "chemical" approach we shall follow initially. Or instead we can start with a solid viewed as a "box of electrons," initially empty of ions, and progressively impose the attraction of the ions. This is the "nearly free electron" method, which we shall develop in Sect. 1.3.

1.2c The Chemical Approach

In the chemical approach we first consider two distant atoms each having one electron. As the atoms are brought closer together, the electrons around each of the nuclei will begin to feel the potential caused by the other nuclei. This potential is a perturbation which lifts the degeneracy more and more effectively as the distance between the atoms is decreased (Fig. 1.1). This holds both for the ground and excited states.

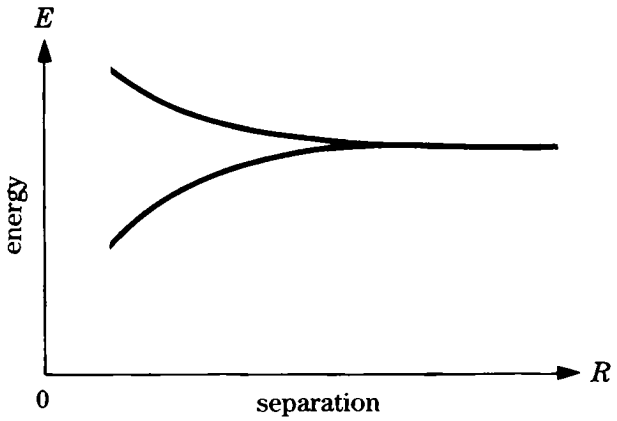


Fig. 1.1. Splitting of an energy level through coupling between two atoms.

In the ground state, the descending branch of Fig. 1.1 is populated by two electrons of opposite spins. This causes a decrease of the system energy when the two atoms are brought closer together, and thus gives rise to a chemical bond, the **covalent bond**. If there are three atoms we start with triply degenerate levels; and in a crystal, with N-fold degeneracy. When the atoms are close, the degeneracy is lifted but several energy levels of the same atom are mixed. This is shown schematically in Fig. 1.2 where the vertical bars show the allowed energies as a function of the separation R of the atoms. The permitted energy bands are separated by forbidden bands. This picture corresponds to **purely covalent** semiconductors such as silicon and germanium. When the atoms constituting the crystal are different, for example in gallium arsenide GaAs, we start with non-degenerate atomic levels. The bond is then **partially covalent**. We shall return in Chap. 2 to the covalent bond. If the relative positions of the bands and their filling by the electrons are such that at zero temperature a band is just full and the band immediately above is empty, we have an insulator, as in such a system a weak electric field cannot increase the electron energies by accelerating them. In fact the only available states are very far away in energy because of the existence of a forbidden energy band: energies near to the initial value are not allowed. Acceleration is thus impossible. The electrons cannot respond to a small electric field, and the system is an insulator. The lower, full, band is called the valence band, while the upper empty band is the conduction band. The forbidden energy region between them is often called the band gap.

The exact determination of the energy bands is difficult; worse, the physical behavior of the solid will be determined by the nature and properties of the levels in small fringes ΔE_c and ΔE_v around the forbidden band. In fact in thermal equilibrium at finite temperature a number of electrons

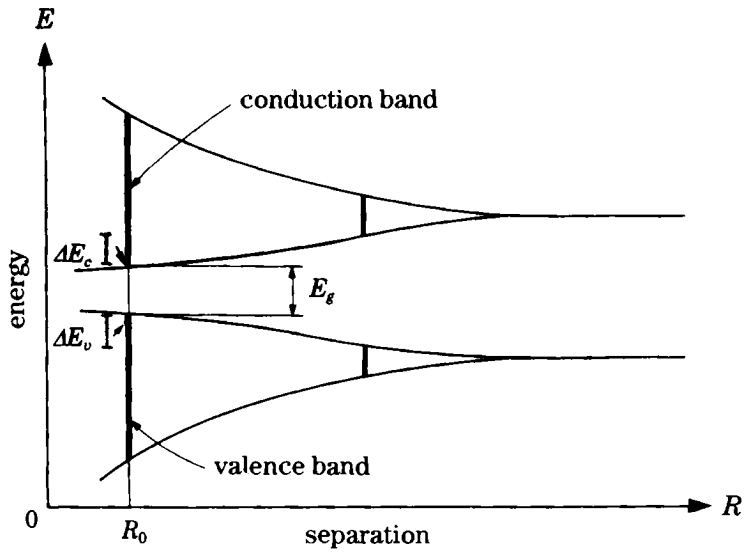


Fig. 1.2. Energy levels in a system of N atoms as a function of the separation R between the atoms. The equilibrium atomic separation is R_0.

populate the conduction band and leave empty states in the valence band. It is clear that the most populated states in the conduction band will be those at the lowest energies, while the empty places in the valence band will correspond to the highest energies allowed in that band. We thus see that observable electron properties will depend on the details of the band structure around the energy extrema on each side of the gap. We thus need to know the band structure very well in the neighborhood of the band gap, something which is very difficult in view of the approximations we have to adopt.

The other approach to understanding the properties of quantum states of the crystal starts from solution of Eq. (1.1) in the case where the crystal potential $V(\mathbf{r})$ is zero, and introduces this potential as a perturbation. We thus get solutions which are linear combinations of plane waves. But the same difficulty appears: the energy scale of the plane waves we have to consider is of the order of atomic energies, i.e., 10 eV. Correct states are linear combinations of a large number of states distributed over a wide energy range, while we are interested in the detailed behavior of only one part of the spectrum. Here too we are faced with approximations which are in conflict with the desired accuracy.

1.3 Quantum States of a Perfect One-Dimensional Crystalline Solid

We consider the case of a linear chain of period a, whose total length L contains N periods, where N is an integer. We seek the eigenstates of the Hamiltonian

$$\mathcal{H} = \frac{p_x^2}{2m} + V(x) \tag{1.3}$$

with a periodic potential

$$V(x + a) = V(x). \tag{1.4}$$

1.3a Nearly Free Electron Model

One way to proceed is to assume that $V(x)$ is small, and regard this term as a perturbation by comparison with the kinetic energy term. For the free electron Hamiltonian

$$\mathcal{H}_0 = \frac{p_x^2}{2m}, \tag{1.5}$$

the eigenfunction of the state $|k>$ is a plane wave:

$$\psi_k(x) = \frac{1}{\sqrt{L}} \exp ikx \tag{1.6}$$

and the corresponding energy is

$$E = \frac{\hbar^2 k^2}{2m}. \tag{1.7}$$

The dispersion curve $E(k)$ is a parabola. To express the fact that the chain of atoms has a finite length $L = Na$, we use the periodic boundary conditions of Born and von Kármán, which consist of joining the chain to itself. We then have

$$\psi(x + L) = \psi(x). \tag{1.8}$$

From Eq. (1.6) we find that the allowed values of k are quantized:

$$k = l_x \frac{2\pi}{Na} \tag{1.9}$$

with l_x an integer.

In a qualitative approach to bringing out the characteristic properties of a crystalline solid, we may confine ourselves to the simplest approximation of a periodic potential: a sine potential of period a:

$$V(x) = 2V_1 \cos \frac{2\pi x}{a}, \tag{1.10}$$

$$= V_1 \left[\exp\left(i\frac{2\pi x}{a}\right) + \exp\left(-i\frac{2\pi x}{a}\right) \right]. \tag{1.11}$$

For a macroscopic crystal N is very large, of the order of $(10^{23})^{1/3}$, and the allowed values of k are thus extremely close: the characteristic distance in k space, defined starting from the period (1.10) in x, is of the order of $2\pi/a$, so that the allowed values of k given by Eq. (1.9) are spaced by an amount $2\pi/L$ and are indeed very close.

The origin of the abscissa is chosen as one of the ions, and the sign of the Coulomb interaction between the ions and the electrons (attractive) requires $V_1 < 0$. We assume that V_1 is small compared with the relevant kinetic energies and use perturbation theory to find its effect on a state $|k>$. Let us look for states $|k'>$ coupled to $|k>$ by V: the matrix element of V between the two states $|k>$ and $|k'>$ is

$$< k'|V|k > = \frac{1}{L} \int_0^L \exp(-ik'x)\, V(x)\, \exp(ikx)\, dx$$

$$= \frac{V_1}{L} \int_0^L \left[\exp i\left(k - k' + \frac{2\pi}{a}\right) x + \exp i\left(k - k' - \frac{2\pi}{a}\right) x \right] dx. \tag{1.12}$$

The two exponentials are periodic functions, and the integral is non-zero only for

$$k - k' = \pm\frac{2\pi}{a}, \tag{1.13}$$

for which its value is V_1. The potential V thus only couples states differing in k by $\pm 2\pi/a$, values of the wave vectors which appear in the Fourier series for V and which express the periodicity of the crystal. In particular, the first-order correction $< k|V|k >$ to the state $|k>$ vanishes.

If the energy of the state $|k'>$ before the introduction of the crystalline potential differs from that of $|k>$, the introduction of the periodic potential has essentially no effect. If by contrast the two independent states coupled by V have the same unperturbed energy, this simultaneously imposes

$$\frac{\hbar^2 k^2}{2m} = \frac{\hbar^2 k'^2}{2m} \tag{1.14}$$

and

$$k' = -k.$$

Then Eq. (1.13) requires

$$k = \pm\frac{\pi}{a} \tag{1.15}$$

and the presence of V lifts the degeneracy between the states $|\pi/a>$ and $|-\pi/a>$. Projected onto this basis the Hamiltonian becomes

$$\mathcal{H} = \begin{pmatrix} \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 & V_1 \\ V_1 & \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 \end{pmatrix}. \tag{1.16}$$

The eigenstates of $\mathcal{H}$ are

$$\psi_- = \frac{1}{\sqrt{2}}\left(\psi_{\frac{\pi}{a}} + \psi_{-\frac{\pi}{a}}\right) = \sqrt{\frac{2}{L}}\cos\frac{\pi}{a}x,$$
$$\text{with } E_- = \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 - |V_1|, \tag{1.17}$$

$$\psi_+ = \frac{1}{\sqrt{2}}\left(\psi_{\frac{\pi}{a}} - \psi_{-\frac{\pi}{a}}\right) = i\sqrt{\frac{2}{L}}\sin\frac{\pi}{a}x,$$
$$\text{with } E_+ = \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 + |V_1|. \tag{1.18}$$

The lowest energy state has an enhanced probability density at the ions (Fig. 1.3).

We see that in the energy interval

$$\frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 - |V_1| < E < \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 + |V_1| \tag{1.19}$$

there is no stationary energy eigenstate. The presence of the periodic electrostatic potential $V = 2V_1\cos(2\pi x/a)$ leads to the existence of a forbidden energy region, or band gap, near $k = \pm\pi/a$. Figure 1.4 shows the resulting lifting of the degeneracy.

We note that applying the same reasoning to the periodic potential

$$V = 2V_l\cos\frac{2\pi}{a}lx, \quad \text{with } l \text{ an integer} \tag{1.20}$$

we obtain coupling between the states

$$k = \pm l\frac{\pi}{a} \tag{1.21}$$

producing a forbidden band for energies:

$$\frac{\hbar^2}{2m}\left(\frac{l\pi}{a}\right) - |V_l| \quad < E < \quad \frac{\hbar^2}{2m}\left(\frac{l\pi}{a}\right)^2 + |V_l|. \tag{1.22}$$

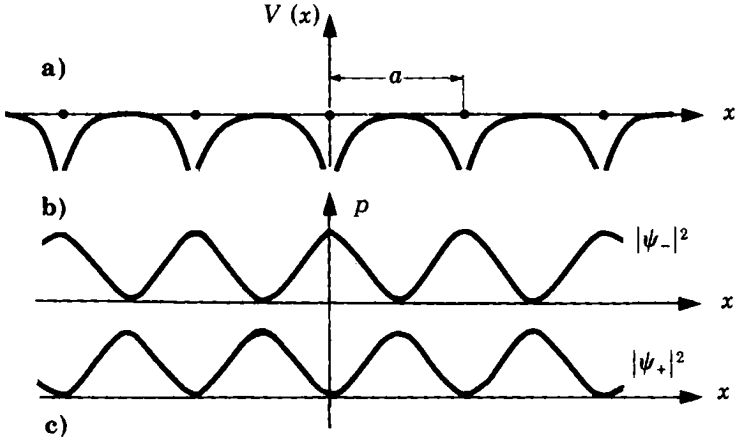


Fig. 1.3. (a) Variation of the electrostatic potential energy of an electron in the Coulomb field of the ions of a linear chain of atoms; (b) and (c) probability densities for the electron for the states ψ_- and ψ_+ (Eqs. (1.17) and (1.18)) after the introduction of a small periodic potential of the form $V(x) = 2V_1 \cos 2\pi x/a$. In the lowest-energy ψ_- state the probability density is maximal at the ions.

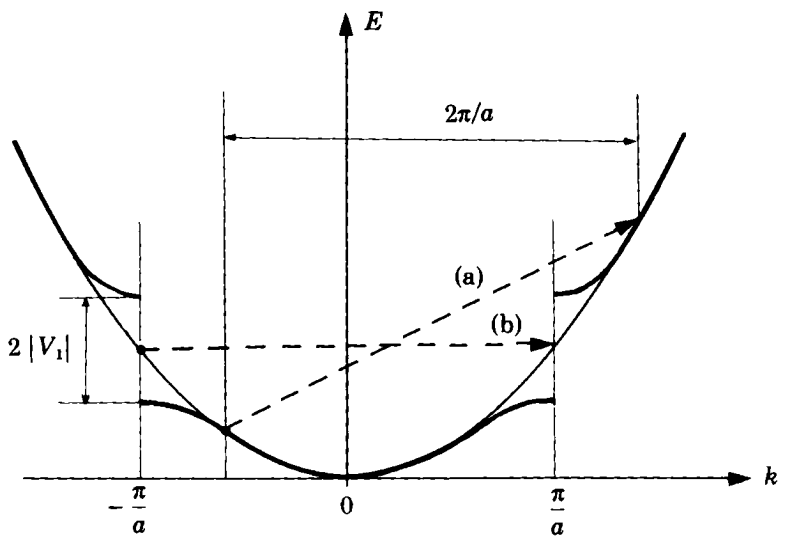


Fig. 1.4. The lightly drawn parabola is the dispersion curve for free electrons. The heavy curve takes account of the effect of the crystalline potential $V = 2V_1 \cos 2\pi x/a$. States coupled by this potential (relation (1.13)) are linked by a dashed arrow: (a) states which are non-degenerate in the absence of the periodic potential; (b) states which are degenerate in the absence of $V(x)$.

More generally, any periodic potential with period a can be written as

$$V = \sum_{l'} 2V_{l'} \cos \frac{2\pi l' x}{a}, \text{ with } l' \text{ an integer.} \tag{1.23}$$

The wave vectors $2\pi l'/a$ appearing in the Fourier decomposition (1.23) of V define a periodic set of period $2\pi/a$ in the space of wave vectors (reciprocal space). The set of points $2\pi l'/a$ constitutes the reciprocal lattice whose generalization to three dimensions we will see in Sect. 2.1. The vectors k defined by Eq. (1.21), around which a band gap opens, are the basic wave vectors of the decomposition (1.23) of V.

By an approximate method, completely different from the chemical approach described in Sect. 1.2b, we have just shown the existence of band gaps in crystalline solids. We will now show that band gaps are expected on general grounds on the basis of Bloch's theorem.

1.3b Bloch's Theorem

In the Bloch theory the electrons, assumed independent, feel a **periodic** crystalline potential $V(x)$. According to Bloch's theorem, the eigenfunctions of

$$\left[\frac{p^2}{2m} + V(x)\right] \psi(x) = E\psi(x), \tag{1.24a}$$

where $V(x+a) = V(x)$ have the form

$$\psi_k(x) = \exp(ikx)\, u_k(x). \tag{1.24b}$$

The term $\exp(ikx)$ describes the variation at large scales, while $u_k(x)$, which has the periodicity of the crystal,

$$u_k(x+a) = u_k(x) \tag{1.25}$$

expresses the variation of the wave function within an elementary cell. The wave vector k, which is real, no longer plays the same role as for the plane wave (1.6). We note that the plane wave is a solution of the type (1.24) for which $u_k(x) =$ constant.

Applying the Hamiltonian to the Bloch function yields the equation satisfied by $u_k(x)$:

$$\left[\left(\hbar k - i\hbar \frac{\partial}{\partial x}\right)^2 + V(x)\right] u_k(x) = E_k\, u_k(x). \tag{1.26}$$

In this equation k plays the role of a parameter, and we can limit x to the interval $[0, a[$ since $u_k(x)$ is periodic. The equation has a form similar to the Schrödinger equation for an atom, and we know that this has discrete

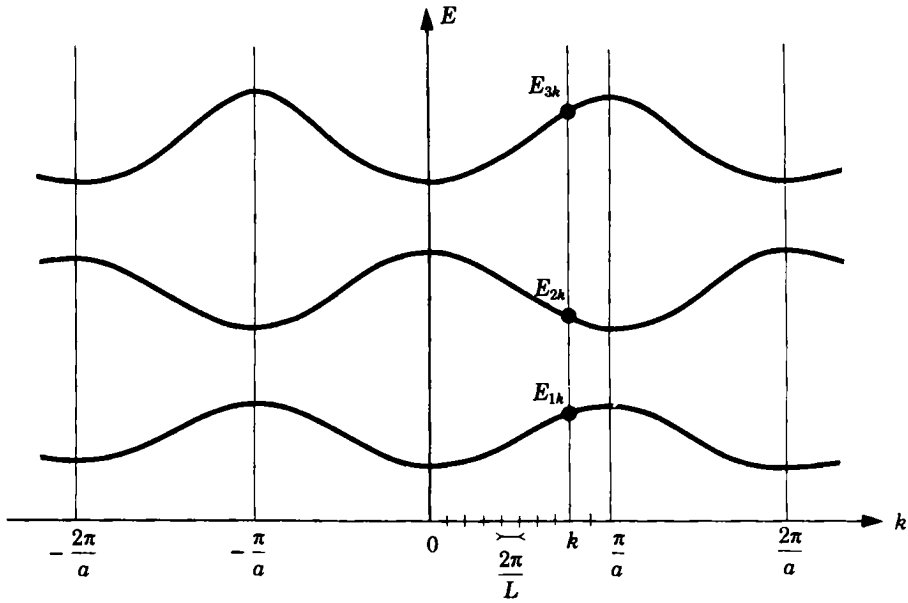


Fig. 1.5. General form of the solutions of the Schrödinger equation (1.24) for a periodic crystalline potential. The solutions $E(k)$ are periodic, with period $2\pi/a$.

eigensolutions, denoted $u_{n,k}(x)$, of energy $E_{n,k}(x)$ $(n = 1, 2, ...)$ (Fig. 1.5). As k is varied, the energy eigenvalues describe an "energy band" for each integer value of n. Whenever there is no overlap in energy between bands with different indices n there appear energy ranges with no stationary values $E_{n,k}$. These are the band gaps.

Their definitions show that the two quantum numbers n and k have quite different meanings. The band index n is an integer. The allowed values of the wave vector k are fixed by the surface boundary conditions of the solid: we could take once again the periodic Born–von Kármán conditions (1.8), which impose the same values (1.9) on k as before; in that case, k varies in a quasi-continuous fashion, with jumps of $2\pi/L$.

The energies $E_{n,k}$ of the Bloch states are periodic functions of k. The Bloch function $\psi_{n,k}$ can also be written as

$$\psi_{n,k}(x) = \exp i\left(k + l\frac{2\pi}{a}\right)x \;\; u_{n,k}(x) \;\; \exp\left(-il\frac{2\pi}{a}x\right), \tag{1.27}$$

where l is an integer, or equivalently,

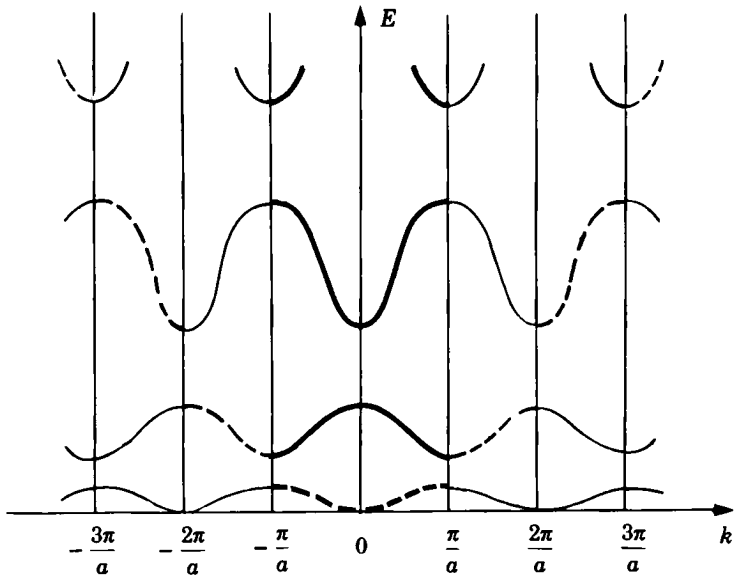


Fig. 1.6. Dispersion curves $E(k)$ in a crystal. The periodicity of the solutions $E(k)$ appears in the lighter curves. The heavy curves show the reduced zone scheme, while the dashed curves represent the extended zone scheme.

$$\psi_{n,k}(x) = \exp i\left(k + l\frac{2\pi}{a}\right)x \cdot u_{n,k+l2\pi/a}(x). \tag{1.28}$$

As $u_{n,k+l2\pi/a}$ is periodic, $\psi_{n,k}$ is also a Bloch function for $k' = k + l2\pi/a$. The associated energy $E_{n,k}$ is thus also an eigenvalue for the state $|k + 2\pi l/a >$, and the energies are periodic in the space of wave vectors k, with period $2\pi/a$.

There are thus two equivalent ways of classifying the eigenstates of the crystal Hamiltonian. Using the periodicity of the solutions of its eigenvalue equation we can arrange to confine ourselves simply to the variation of k over the interval $[-\pi/a, \pi/a[$: for each value of k in this interval there exist discrete solutions labelled by the integer index n. This situation is described by the heavy curves in Fig. 1.6; it is called the **reduced zone scheme**. By contrast we may retain just one branch of the successive curves $E(k)$ for each value of k: the lowest for $-\pi/a \leq |k| < \pi/a$, the second branch for $\pi/a < |k| \leq 2\pi/a$, and so on. We thus obtain the dashed curves in Fig. 1.6; this is called the **extended zone scheme**.

These two descriptions are strictly equivalent. The extended zone description is reminiscent of the parabola $E(k)$ of free electrons in a box. The reduced zone description carries the discrete index n and is reminiscent of the quantization of atomic levels: for $k = 0$, Eq. (1.26) actually gives solutions of atomic type.

We have used quantum mechanics to show that the energy levels of a solid group themselves into permitted and forbidden bands. Let us now see how Fermi–Dirac statistics dictate the filling of these levels.

1.3c Level Filling

The boundary conditions (1.8) on the wave function at the sample surface confine k to the values $k = l_x 2\pi/Na$, where l_x is an integer. Moreover the discussion of crystalline solids in Sect. 1.3b allows us to characterize each electron eigenstate by its band index n and its wave vector k. As we have seen, restricting k to the interval $[-\pi/a, \pi/a[$ gives a full description of the properties of the solid. l_x takes exactly N values over this range. For each value of l_x and thus of k there are two independent allowed spin states, leading to $2N$ possible states per band.

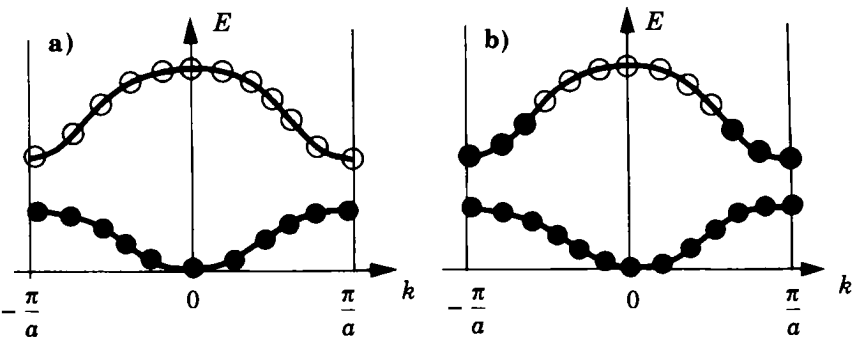


Fig. 1.7. Filling of crystal energy levels at zero temperature. (a) The electron number z per cell is even, and the solid is an insulator; (b) z is odd, and the solid is a metal. Filled circles represent occupied states and hollow circles empty states.

For a linear crystal composed of a medium with z electrons per elementary cell, the total number of electrons in the line is Nz. At zero temperature, the Pauli principle dictates the filling of the levels, starting with the states of lowest energy. If the Nz electrons completely fill one or more bands but leave others empty, i.e., if z is even and the bands do not overlap (Fig. 1.7(a)), it would require an electric field E of order E_g/ea, where E_g is the width of the band gap and e the electron charge, to give

an electron enough energy to lift it into the first available energy state, at the bottom of the first empty band. For a normal medium, where E_g is of the order of 1 eV, this requires an electric field $\mathcal{E} \simeq 3 \times 10^9$ V/m, more than the breakdown field! In such a medium there is no acceleration in an external field, thus no conduction, and the medium is an **insulator**.

We denote as a semiconductor an insulator whose band gap between the last filled band, called the valence band, and the first empty band, called the conduction band, is narrow enough to allow thermal excitation of charge carriers at room temperature, i.e., a significant concentration of electrons in the conduction band.

If by contrast z is an odd number (Fig. 1.7(b)), the filling leaves one or more bands only partly full at zero temperature. The first excited state is very close to the last filled state, and the system can respond to even a weak electric field. The medium is therefore metallic, as the electrons can be accelerated by an electric field.

In summary, this discussion of a one-dimensional crystal introduces the following basic ideas:

– an electron state is characterized by two quantum numbers (n, k);

– the range of variation of k can be restricted to the reduced zone $[-\pi/a, \pi/a[$;

– because of the crystalline potential, not all energies are accessible to the electrons. Permitted bands are separated by band gaps;

– the filling of the electron energy levels according to the Pauli principle shows that, depending on the medium and the form of the energy bands, the solid behaves as an insulator or a metal. If the number of electrons per cell is even and the bands do not overlap, the solid is an insulator. It is a metal in the opposite case. A semiconductor is an insulator with a "narrow" band gap.

The existence of a band gap thus explains three characteristics of semiconductors:

– The conductivity rises with temperature since thermal excitation gives a conduction band population which increases with T. In this band there are many empty electron states able to accomodate electrons which can be accelerated by even a weak electric field.

– The existence of a forbidden energy band explains the transparency of semiconductors to infrared radiation; photons with energy $h\nu$ smaller than E_g cannot be absorbed, as the electron cannot reach a final state within the forbidden band.

– In contrast, if $h\nu$ exceeds E_g, electrons can be excited into the conduction band by absorbing photons. This explains the existence of "photoconductivity" in semiconductors.

2.

Quantum States of a Perfect Semiconductor

We have seen in the case of a one-dimensional periodic solid that quantum mechanics provides a basis for understanding the properties of semiconductors. Here we discuss the physics of electrons in perfect semiconductors, i.e., those without defects or impurities. We will see in later chapters that certain properties of these materials depend quite directly on the presence of such impurities or defects in the crystal lattice.

The characteristic properties of a solid are determined by the distribution of energy levels called the band structure. From this band structure we can say whether the medium is an insulator, a conductor, or a semiconductor.

We have to know the band structure for other reasons. In the presence of the periodic crystal potential the response of an electron to an external force is no longer determined by its mass, but by an "effective mass" imposed by the band structure. The effective mass of an electron in a crystal may be very different from the electron mass itself: it can even be negative.

2.1 Quantum States of a Three-Dimensional Crystal

A crystal is made up of the periodic repetition of an elementary basis. More precisely, if a particular point of the structure in equilibrium has coordinates $\mathbf{r}$, all the points in equilibrium having the same physical and chemical "environment" as $\mathbf{r}$ are expressible as

$$\mathbf{r}' = \mathbf{r} + m_1\mathbf{a}_1 + m_2\mathbf{a}_2 + m_3\mathbf{a}_3, \tag{2.1}$$

where the m_i $(i = 1, 2, 3)$ are whole numbers and the $\mathbf{a}_i$ are three noncoplanar vectors specific for the particular crystal. The rhombohedron constructed from $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ makes up the cell. The "smallest" cell, with the smallest volume, is called the primitive or unit cell. The vectors $\mathbf{a}_i$ which generate this primitive cell are the periods of the Bravais lattice, or direct lattice, of the crystal.

Semiconductors formed from elements of column IV of the Periodic Table (Si, Ge), and most semiconductors formed from elements of column III and elements from column V (GaAs, InP) crystallize in the face-centered cubic system shown in Fig. 2.1. Some semiconductors of type II–VI like CdS crystallize in a hexagonal form.

In the face-centered cubic system the primitive cell is not a cube but the rhombohedron shown in Fig. 2.1 (or any equivalent cell, e.g., that formed from the basis vectors $-\mathbf{a}_1, \mathbf{a}_2, -\mathbf{a}_3$). For the most usual semiconductors the cell consists of two atoms, one at the position $(0,0,0)$, and the other one-quarter of the way along the diagonal of the cube (Fig. 2.2). The primitive cell thus contains a basis of two atoms. For simplicity, and to show the crystal symmetries, we generate the crystal from a cubic cell which is not the primitive cell. **The lattice of points generated by the translations (2.1) of the primitive cell is called the Bravais lattice.**

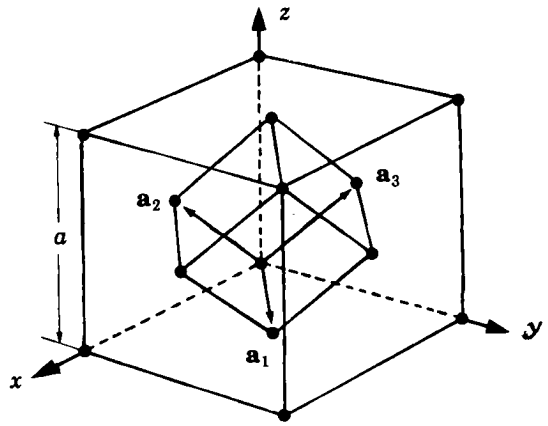


Fig. 2.1. Face-centered cubic lattice. The rhombohedron constructed from the vectors $\mathbf{a}_1 = (a/2)(\mathbf{i}+\mathbf{j})$, $\mathbf{a}_2 = (a/2)(\mathbf{j}+\mathbf{k})$, $\mathbf{a}_3 = (a/2)(\mathbf{i}+\mathbf{k})$ forms the primitive cell.

A macroscopic solid, of characteristic size L contains a very large number of elementary cells, of the order of $(L/a)^3$, where a is the characteristic size of the cell.

In an infinite crystal, electrons at displacements of

$$\mathbf{T} = m_1\mathbf{a}_1 + m_2\mathbf{a}_2 + m_3\mathbf{a}_3 \tag{2.2}$$

feel the same crystal potential, implying that the Hamiltonian is invariant under translations $\mathbf{T}$ of the lattice.

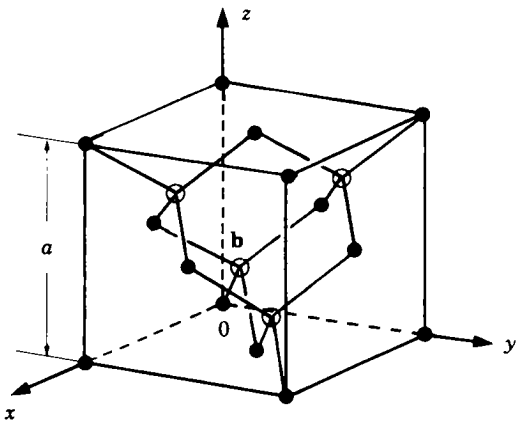


Fig. 2.2. Crystalline structure of silicon. The basis consists of two atoms, that at the origin and atom **b** with coordinates $(a/4)(1, 1, 1)$. The Bravais lattice is face-centered cubic.

2.1a Bloch's Theorem

Solutions of the Schrödinger equation

$$\left[\frac{p^2}{2m} + V(\mathbf{r})\right] \psi = E\psi, \tag{2.3}$$

where the crystal potential $V(\mathbf{r})$ has the periodicity of the crystal, are called "Bloch functions," and have the form

$$\psi_{n,\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k}.\mathbf{r})\, u_{n,\mathbf{k}}(\mathbf{r}), \tag{2.4}$$

where the function $u_{n,\mathbf{k}}$ is periodic in the direct lattice:

$$u_{n,\mathbf{k}}(\mathbf{r} + \mathbf{T}) = u_{n,\mathbf{k}}(\mathbf{r}). \tag{2.5}$$

A state is thus specified by four quantum numbers, n and the three components k_x, k_y, k_z of the vector $\mathbf{k}$. From Eq. (2.5),

$$\psi_{n,\mathbf{k}}(\mathbf{r} + \mathbf{T}) = \exp(i\mathbf{k}.\mathbf{T})\, \psi_{n,\mathbf{k}}(\mathbf{r}). \tag{2.6}$$

The points $\mathbf{r}$ and $\mathbf{r}+\mathbf{T}$ thus have the same physical properties, the functions differing only by a phase factor independent of $\mathbf{r}$. We note further that $|\psi_{n,\mathbf{k}}(\mathbf{r})|^2$ is periodic in space.

As in Chap. 1, we choose as boundary conditions at the edges of the macroscopic solid the periodic boundary conditions of Born and von Kármán:

$$\psi(\mathbf{r} + \mathbf{L}) = \psi(\mathbf{r}). \tag{2.7}$$

We assume that the vector $\mathbf{L}$ is a vector of the Bravais lattice, corresponding to a very large number of periods. For simplicity we choose an elementary cell and a solid in the shape of a rectangular parallelepiped. Imposing condition (2.7) on the form (2.4) of the Bloch function requires the values of $\mathbf{k}$ to be quantized:

$$k_x = l_x \frac{2\pi}{L_x}, \quad k_y = l_y \frac{2\pi}{L_y}, \quad k_z = l_z \frac{2\pi}{L_z}, \tag{2.8}$$

where (l_x, l_y, l_z) are whole numbers or zero. For a macroscopic solid these values of $\mathbf{k}$ are very close on the scale of $1/a$, since $L >>> a$. The vectors $\mathbf{k}$ play an essential role in the description of the properties of solids. We can regard them as belonging to a space "reciprocal" to the crystal itself: **the reciprocal space is generated by the basis vectors $\mathbf{a}_i^*$ derived from the basis vectors $\mathbf{a}_j$ of the Bravais lattice by**

$$\mathbf{a}_i^* \cdot \mathbf{a}_j = 2\pi\delta_{ij}. \tag{2.9}$$

Any vector $\mathbf{G}$ of the reciprocal lattice has the form

$$\mathbf{G} = h\mathbf{a}_1^* + k\mathbf{a}_2^* + l\mathbf{a}_3^* \tag{2.10}$$

(h, k, l) being whole numbers. In one dimension we would have $\mathbf{G} = (2\pi/a)\mathbf{i}$.

The reciprocal lattice of a face-centered cubic lattice with side a is a body-centered cubic lattice with side $4\pi/a$ (as in Si, GaAs, ...) (Fig. 2.3). The reciprocal lattice of a hexagonal lattice is a hexagonal lattice (as in CdS, ...).

From the definition (2.9) it follows that for any translation $\mathbf{T}$ of the direct lattice defined by Eq. (2.2) and for any vector $\mathbf{G}$ of the reciprocal lattice given by Eq. (2.10), we have

$$\exp i\mathbf{G} \cdot \mathbf{T} = 1. \tag{2.11}$$

The relation (2.11) shows the symmetrical roles of the direct and reciprocal lattices (the reciprocal of the reciprocal lattice is the direct lattice). We recall that the real crystal is a lattice of atoms or molecules, or more generally of bases. The reciprocal lattice is a network of points independent of the basis in the real crystal.

Let us consider a Bloch function relative to a wave vector $\mathbf{k}$, such that $\mathbf{k} = \mathbf{k}_0 + \mathbf{G}$, where $\mathbf{G}$ is a vector of the reciprocal lattice, and show that this Bloch function $\psi_{n,\mathbf{k}}(\mathbf{r})$ is also a Bloch function for the wave vector $\mathbf{k}_0$:

$$\begin{aligned} \psi_{n,\mathbf{k}} &= \exp(i\mathbf{k} \cdot \mathbf{r})\ u_{n,\mathbf{k}}(\mathbf{r}) \\ &= \exp(i\mathbf{k}_0 \cdot \mathbf{r})\ \exp(i\mathbf{G} \cdot \mathbf{r})\ u_{n,\mathbf{k}}(\mathbf{r}). \end{aligned} \tag{2.12}$$

The function $\exp i\mathbf{G}\cdot\mathbf{r}\ u_{n,\mathbf{k}}(\mathbf{r})$ has the property (2.5). In fact by using Eqs. (2.5) and (2.11) we deduce

$$\exp i\mathbf{G}\cdot(\mathbf{r}+\mathbf{T})\ u_{n,\mathbf{k}}(\mathbf{r}+\mathbf{T}) = \exp i(\mathbf{G}\cdot\mathbf{r})\ u_{n,\mathbf{k}}(\mathbf{r}) \tag{2.13}$$

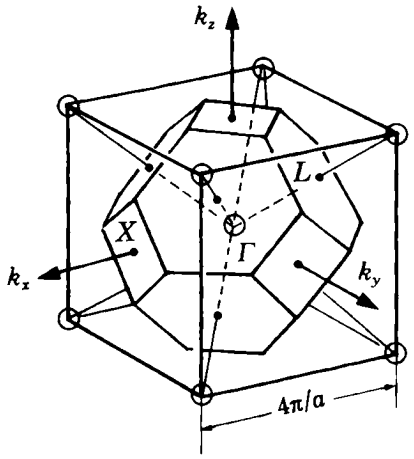


Fig. 2.3. Reciprocal lattice of the face-centered cubic lattice of silicon. This is a body-centered cubic lattice of side $4\pi/a$. We also show the first Brillouin zone. This is bounded by the perpendicular bisector planes of the vectors $(2\pi/a)(1,1,1)$ and their equivalents, producing the hexagonal faces with center L and also by the perpendicular bisector planes of the vectors $(4\pi/a,0,0)$ and their equivalents, corresponding to the square faces with center X. The origin Γ is chosen at the atom of the cubic lattice situated at the center of the cube.

which shows that the latter function is periodic. In consequence the function (2.12) $\exp(i\mathbf{k}_0 \cdot \mathbf{r})\ \exp(i\mathbf{G} \cdot \mathbf{r})\ u_{n,\mathbf{k}}\ (\mathbf{r})$ is a solution of the Schrödinger equation simultaneously for $\mathbf{k}$ and for $\mathbf{k}_0 = \mathbf{k} - \mathbf{G}$. The reciprocal space is thus inconveniently "too large" for classifying the Bloch functions, since it contains all the points $\mathbf{k}$ and $\mathbf{k} - \mathbf{G}$, for any $\mathbf{G}$, whose quantum states are identical.

2.1b The Brillouin Zone

We can thus reduce the area of study of Bloch states to those values of $\mathbf{k}$ belonging to the **"first Brillouin zone." The first Brillouin zone is the volume of the reciprocal space closer to the original node $k = 0$ than to any other point of the reciprocal lattice.** This first Brillouin zone is bounded by the perpendicular bisector planes of the shortest vectors $\mathbf{G}$ of the reciprocal lattice. It has the same volume as the elementary cell of the reciprocal lattice. In one dimension it corresponds to the interval $[-\pi/a, \pi/a[$. In three dimensions one of the edges is part of the first zone, but not the one opposite to it.

The first Brillouin zone is shown for a face-centered cubic Bravais lattice (Si, Ge, GaAs, ...) in Fig. 2.3. If the side of the cube is a, the coordinates

of the points of high symmetry of the reciprocal lattice shown on the figure are $(\pi/a)(1,1,1)$ for L and $(2\pi/a,0,0)$ for X. The coordinates of points equivalent under the symmetries of the cube are easily deduced.

If we restrict the range of $\mathbf{k}$ to the first Brillouin zone we say we are in the "restricted zone scheme," in contrast to the "extended zone scheme." Similarly, as in one dimension, for a fixed value of $\mathbf{k}$ the Schrödinger equation has a discrete set of solutions, which define, as $\mathbf{k}$ varies, the permitted energy bands, separated by forbidden bands. The index n of the Bloch function (2.4) is thus the band index; when $\mathbf{k}$ varies the eigenvalue $E_{n,\mathbf{k}}$ of Eq. (2.3) spans the nth energy band. The accessible values (2.8) of $\mathbf{k}$ are very close on the scale of the Brillouin zone, so we can regard $E_{n,\mathbf{k}}$ as a quasi-continuous function $E_n(\mathbf{k})$.

We assume without proof that **the functions $\psi_{n,\mathbf{k}}(\mathbf{r})$ form a complete orthonormal basis:**

$$< n,\mathbf{k}|n',\ \mathbf{k}' >= \delta_{n,n'}\delta_{\mathbf{k},\mathbf{k}'}$$

$$\sum_{n\mathbf{k}} |n\mathbf{k}><n\mathbf{k}| = \mathbf{1}. \tag{2.14}$$

The matrix elements of a periodic operator between states $|n,\mathbf{k}>$ and $|n',\mathbf{k}'>$ of Bloch function form vanish unless $\mathbf{k}=\mathbf{k}'$. This is shown in Appendix 2.1.

2.1c Inversion Symmetry of Constant Energy Surfaces in k-space

If $\psi_{n,\mathbf{k}}(\mathbf{r})$ satisfies

$$\left[\frac{p^2}{2m} + V(\mathbf{r})\right]\psi_{n,\mathbf{k}}(\mathbf{r}) = E_{n,\mathbf{k}}\,\psi_{n,\mathbf{k}}(\mathbf{r}), \tag{2.15}$$

we see on taking the complex conjugate of this equation

$$\left[\frac{p^2}{2m} + V(\mathbf{r})\right]\psi^*_{n,\mathbf{k}}(\mathbf{r}) = E_{n,\mathbf{k}}\,\psi^*_{n,\mathbf{k}}(\mathbf{r}) \tag{2.16}$$

that $\psi^*(\mathbf{r})$ is an eigenstate with the same eigenvalue. But $\psi^*_{n,\mathbf{k}}(\mathbf{r}) = u^*_{n,\mathbf{k}}(\mathbf{r})\exp(-i\mathbf{k}\cdot\mathbf{r})$ is a Bloch function for the point $-\mathbf{k}$ of the Brillouin zone, and thus orthogonal to $\psi_{n,\mathbf{k}}(\mathbf{r})$, with the corresponding eigenvalue $E_{n,-\mathbf{k}}$. We deduce that

$$E_{n,\mathbf{k}} = E_{n,-\mathbf{k}}. \tag{2.17}$$

We can thus confine ourselves to studying the dispersion relations $E_n(\mathbf{k})$ in one-half of the zone. This property is a consequence of the reality of the Hamiltonian, which arises from the invariance under time reversal of the laws of microscopic mechanics. We deduce that at the point $\mathbf{k} = 0$ the

relations $E_n(\mathbf{k})$ have the forms 1, 2, or 3, the set of curves $E_n(\mathbf{k})$ being even (Fig. 2.4).

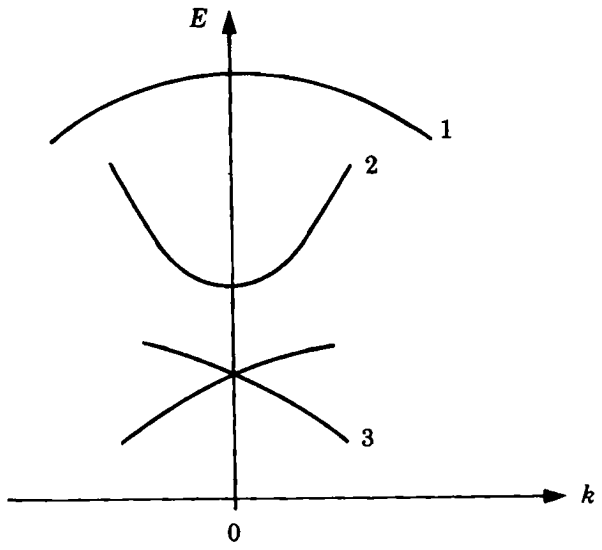


Fig. 2.4. Behavior of dispersion relations in the neighborhood of $\mathbf{k} = 0$.

2.1d Symmetries of Constant Energy Surfaces

Besides translation symmetry, crystals possess point symmetries, i.e., they are invariant under certain symmetry transformations: rotations through $\pi/2, 2\pi/3, \ldots$, symmetries with respect to a point, an axis, or a plane. Use of the symmetry properties of the Hamiltonian is fundamental in simpifying the study of the system eigenstates, but it requires the use of group theory, which is beyond the scope of this book. Appendix 2.2 gives a proof of the following fundamental result: **the surfaces of constant energy have the symmetries of the crystal.**

Because of these crystal symmetries and the inversion symmetry of the constant energy surfaces we need only study a volume of $\mathbf{k}$ space equal to the volume of the Brillouin zone divided by twice (because of the inversion symmetry) the number of symmetries of the crystal.

2.2 Dynamics of a Bloch Electron. The Crystal Momentum

We begin by noting that the wave vector $\mathbf{k}$ introduced by the Bloch function (2.4) is a quantum number with three components which is not related in a simple way to the momentum operator $\mathbf{p} = -i\hbar\nabla$, unlike in the case of plane waves. In fact, if the potential $V(\mathbf{r})$ were constant, the solution of the Schrödinger equation would be a pure plane wave:

$$\psi_{\mathbf{k}} = C \exp(i\mathbf{k} \cdot \mathbf{r}), \tag{2.18}$$

the constant C ensuring the normalization of the wave function over the macroscopic solid. In this case the mean value of the operator $\mathbf{p}$ would be $\hbar\mathbf{k}$:

$$\mathbf{p}\ \psi_{\mathbf{k}} = -i\hbar\nabla(C \exp i\mathbf{k} \cdot \mathbf{r}) = \hbar\mathbf{k}\ \psi_{\mathbf{k}}\,. \tag{2.19}$$

By contrast, for a Bloch function,

$$\begin{aligned} < \mathbf{p} >_{n,\mathbf{k}} &=< n, \mathbf{k} \,|\mathbf{p}|\, n, \mathbf{k} >, \qquad (2.20)\\ &= \int \psi^*_{n,\mathbf{k}}(\mathbf{r})\ (-i\hbar\nabla)\psi_{n,\mathbf{k}}(\mathbf{r})\ d^3\mathbf{r},\\ &= \int u^*_{n,\mathbf{k}}(\mathbf{r})[\hbar\mathbf{k} - i\hbar\nabla]u_{n,\mathbf{k}}(\mathbf{r})\ d^3\mathbf{r}\\ &= \hbar\mathbf{k} - i\hbar \int u^*_{n,\mathbf{k}}(\mathbf{r})\nabla u_{n,\mathbf{k}}(\mathbf{r}) d^3\mathbf{r}\\ &\neq \hbar\mathbf{k}. \qquad (2.21) \end{aligned}$$

In a crystal there is no simple relation between the mean value of the momentum in a Bloch state and the wave vector $\mathbf{k}$ which defines the state. However the vector $\hbar\mathbf{k}$ plays a very special role, as we shall see. We call this vector the **crystal momentum.**

2.2a True Momentum, Group Velocity.

We show here that the mean value of the momentum is related in a simple manner to the behavior of the dispersion relations $E_n(\mathbf{k})$. Using a first-order Taylor series we write the energy at the point $\mathbf{k}+\mathbf{q}$:

$$E_n(\mathbf{k}+\mathbf{q}) = E_n(\mathbf{k}) \ + \ \mathbf{q} \cdot \nabla_{\mathbf{k}}\ E_n(\mathbf{k}) \ + \cdots. \tag{2.22}$$

Now the second term can be easily calculated by perturbation theory. Trivially generalizing Eq. (1.26) to three dimensions we write the Hamiltonian $\mathcal{H}_{\mathbf{k}}$ satisfied by the periodic part $u_{n,\mathbf{k}}\,(\mathbf{r})$ of the Bloch function:

$$\mathcal{H}_{\mathbf{k}} = \frac{1}{2m}(\hbar\mathbf{k} - i\hbar\nabla)^2 + V(\mathbf{r}). \tag{2.23}$$

Similarly, at the point $(\mathbf{k}+\mathbf{q})$ we can write

$$\mathcal{H}_{\mathbf{k}+\mathbf{q}} = \mathcal{H}_{\mathbf{k}} + \frac{\hbar^2}{m}\mathbf{q}\cdot(\mathbf{k}-i\nabla) + \frac{\hbar^2 q^2}{2m}. \tag{2.24}$$

We see that for small $|\mathbf{q}|$ the second term is small and the third is negligible. We can thus obtain $E_n(\mathbf{k}+\mathbf{q})$ from $E_n(\mathbf{k})$ by using first-order perturbation theory, which gives

$$E_n(\mathbf{k}+\mathbf{q}) = E_n\,(\mathbf{k}) + \int u^*_{n,\mathbf{k}}\,\frac{\hbar^2}{m}\mathbf{q}\cdot(-i\nabla+\mathbf{k})\,u_{n,\mathbf{k}}\,d^3\mathbf{r} + \cdots \tag{2.25}$$

or equivalently:

$$E_n\,(\mathbf{k}+\mathbf{q}) = E_n\,(\mathbf{k}) + \int \psi^*_{n,\mathbf{k}}\,\mathbf{q}\cdot\left(-\frac{\hbar^2}{m}i\nabla\right)\psi_{n,\mathbf{k}}\,d^3\mathbf{r} + \cdots. \tag{2.26}$$

Equating the coefficients of $\mathbf{q}$ in Eqs. (2.22) and (2.26) we obtain the relation

$$\nabla_{\mathbf{k}}\,E_n(\mathbf{k}) = \frac{\hbar}{m} < \mathbf{p} >, \tag{2.27}$$

where the mean is taken over $\psi_{n,\mathbf{k}}$. Ehrenfest's theorem gives the group velocity as

$$\mathbf{v} = \frac{d < \mathbf{r} >_{n,\mathbf{k}}}{dt} = \frac{1}{m} < \mathbf{p} >_{n,\mathbf{k}} \tag{2.28}$$

from which we deduce the velocity of an electron in the Bloch state $\psi_{n,\mathbf{k}}$:

$$\mathbf{v} = \frac{1}{\hbar}\nabla_{\mathbf{k}}\,E_n\,(\mathbf{k}). \tag{2.29}$$

In fact the motion of an electron regarded as a particle should be described by a wave packet. For an electron in a crystal this is a packet of Bloch waves centered on $\mathbf{k} = \mathbf{k}_0$, which is constructed by introducing other neighboring states $\mathbf{k}$ belonging to the same band n.

We note that, as the Bloch states are eigenstates of $\mathcal{H}$, the velocity of an electron in a Bloch state is constant: an electron in such a state suffers no collisions in the crystalline potential included in $\mathcal{H}$. This is a fundamental discovery: **a periodic potential does not scatter Bloch electrons**; it determines their constant velocity through Eq. (2.29).

In a perfectly periodic crystal electrons suffering no collisions would thus have infinite conductivity. The deviations from periodicity determine the finite value of the conductivity. The defects that are most effective in producing scattering are the presence of impurities and the fact that at a finite temperature the crystal undergoes thermal vibrations which deform the perfect crystalline lattice. The latter excitations, called phonons, are not studied in this book, but play an important role in limiting the mobility of electrons and holes in semiconductors.

2.2b Acceleration Theorem in the Reciprocal Space

Under the effect of an external electric field $\mathcal{E}$ the energy of an electron is modified. Let us assume that $\mathcal{E}$ varies little over the scale of the cell, and only slowly with time at the scale of the transition frequencies between permitted energy bands. The work dW done on an electron of speed $\mathbf{v}$ and charge $-e$ over the time interval dt changes its energy $E_n(\mathbf{k})$ by modifying the value of $\mathbf{k}$ and thus the crystal momentum. Hence we have the relation

$$dW = -e\mathcal{E} \cdot \mathbf{v}dt = \frac{dE_n(\mathbf{k})}{dt}\, dt. \tag{2.30}$$

Using expression (2.29) for $\mathbf{v}$ we deduce

$$-e\mathcal{E} \cdot \frac{1}{\hbar}\nabla_{\mathbf{k}}\, E_n(\mathbf{k}) = \nabla_{\mathbf{k}}\, E_n(\mathbf{k}) \cdot \frac{d\mathbf{k}}{dt} \tag{2.31}$$

so that

$$\hbar\frac{d\mathbf{k}}{dt} = -e\mathcal{E} = \mathbf{F}, \tag{2.32}$$

where $\mathbf{F}$ is the applied force. This is the **acceleration theorem in the reciprocal space.** The essential result is that **the response to an external force varying slowly in space and time is equal to the derivative of the crystal momentum and not the derivative of the electron momentum.** In the presence of an electric field and a magnetic field $\mathbf{B}$ it is possible to generalize Eq. (2.32) to

$$\frac{\hbar d\mathbf{k}}{dt} = -e\,(\mathcal{E} + \mathbf{v} \times \mathbf{B}) = \mathbf{F}. \tag{2.33}$$

In fact, Eqs. (2.32) or (2.33) describe the behavior of a packet of Bloch waves, localized to about Δr in real space, and to about $\Delta k \sim 1/(\hbar\Delta r)$ in the reciprocal space. If the force $\mathbf{F}$ varies in time at the scale of the transition frequencies between bands the evolution of the system can no longer be described by the motion of the point $\mathbf{k}$ within a given band, but by transitions between bands. This is true of the effect of light on a semiconductor, studied in Chaps. 6 and 7: optical frequencies are of the order of 10^{14} Hz.

2.2c Effective Mass and Acceleration in Real Space

Differentiating the velocity $\mathbf{v}$ given by Eq. (2.29) with respect to time, and using the acceleration theorem (2.33), we obtain

$$\frac{d\mathbf{v}}{dt} = (\nabla_{\mathbf{k}}\, \mathbf{v}) \cdot \frac{d\mathbf{k}}{dt} = \frac{1}{\hbar^2}\nabla_{\mathbf{k}}\, [\nabla_{\mathbf{k}}\, E_n(\mathbf{k})] \cdot \mathbf{F} \tag{2.34}$$

or

$$\frac{dv_\alpha}{dt} = \sum_\beta \left(\frac{1}{m^*}\right)_{\alpha\beta} F_\beta \tag{2.35}$$

with

$$\left(\frac{1}{m^*}\right)_{\alpha\beta} = \frac{1}{\hbar^2}\frac{\partial^2 E_n(\mathbf{k})}{\partial k_\alpha \partial k_\beta} \tag{2.36}$$

which defines the **effective mass tensor** at a point $\mathbf{k}$ of a given band n. Expressions (2.35) and (2.36) constitute the **acceleration theorem in real space**, which is subject to the same restrictions of slow variations in space and time on the force $\mathbf{F}$ as Eq. (2.33) from which it results. From its definition we see that the effective mass tensor is symmetric.

The notion of effective mass is of most interest in the vicinity of an extremum of the band, where, to lowest order,

$$E_n(\mathbf{k}) - E_n(\mathbf{k}_0) \simeq \sum_{\alpha,\beta} \frac{\hbar^2}{m_{\alpha\beta}} \Delta k_\alpha \Delta k_\beta. \tag{2.37}$$

Here $\mathbf{k}_0$ is the wave vector of the extremum, and $\mathbf{k} = \mathbf{k}_0 + \Delta\mathbf{k}$.

At the zone center of a cubic crystal, if the energy is not degenerate, the constant energy surfaces are spheres; the effective mass is thus isotropic and has a value m^*. It is positive near a band minimum and negative near a band maximum. A negative effective mass implies, from Eq. (2.35), that the velocity resulting from the action of F_β is in the opposite direction from that acquired by an electron in vacuum acted on by F_β (cf. the definition of a hole in Chap. 3). This apparently paradoxical behavior should be compared with Hall effect experiments which in certain materials imply the existence of positive charge carriers. It shows that in a solid the response of an electron to an applied force is strongly influenced by the reaction to the crystal potential. Even when the effective mass has the same sign as m, the mass of a free electron, we can have values of m^*/m very different from unity. While in metals $m^*/m \sim 1$, this is not always true of semiconductors: the effective mass of the conduction band is $+0.067m$ in GaAs, and $+0.014m$ in InSb.

We note that the momentum formula for the simplest case can be written

$$m^* \frac{d\mathbf{v}}{dt} = \mathbf{F} \tag{2.38}$$

so that it is not the derivative of the ordinary momentum $m d\mathbf{v}/dt$ which is equal to the external force. In this special case, the velocity (2.29) and crystal momentum are related by

$$\mathbf{v} = \frac{\hbar\mathbf{k}}{m^*}. \tag{2.39}$$

2.3 Metal, Insulator, Semiconductor

The results we have obtained for the Bloch states have the following fundamental consequence: **a full band cannot carry a current.** In fact from the definition of the current density,

$$\mathbf{j} = -e \sum_{\mathbf{k} \in \text{full band}} \mathbf{v}(\mathbf{k}), \tag{2.40}$$

$$= -\frac{e}{\hbar} \sum_{\mathbf{k} \in \text{full band}} \nabla_{\mathbf{k}} E_n(\mathbf{k}) = 0, \tag{2.41}$$

as the functions $E_n(\mathbf{k})$ are even in $\mathbf{k}$ (2.17): from Eq. (2.29),

$$\mathbf{v}(-\mathbf{k}) = -\mathbf{v}(\mathbf{k}) \tag{2.42}$$

and the total current is zero. Let us now apply an electric field $\mathcal{E}$. Using the acceleration theorem in the reciprocal space, each vector $\mathbf{k}$ is modified by $d\mathbf{k} = -e\mathcal{E}dt$ in time dt, the shift being the same for all the vectors. From the definition of the first Brillouin zone, states leaving this zone are equivalent up to a vector $\mathbf{G}$ of the reciprocal lattice to those which become empty, and the band remains full. A full band does not react to an applied electric field and does not participate in the current. If the occupied bands of a solid are completely full, the solid is an insulator.

To find the filling factor of the band states of a three-dimensional solid we will use reasoning similar to that used for a one-dimensional system in Sect. 1.3c; the situation is slightly complicated by the fact that we have to consider all directions of the wave vector $\mathbf{k}$.

2.3a Density of States in the Reciprocal Space

The Born–von Kármán boundary conditions (2.8) result in a uniform density of states in the space of wave vectors $\mathbf{k}$: we consider again the example of a crystal of macroscopic dimensions L_x, L_y, L_z. The accessible wave vectors have components:

$$k_i = l_i \frac{2\pi}{L_i} \quad \text{with } i = x,\ y,\ z. \tag{2.43}$$

Two spin states correspond to each wave vector. The number of accessible states in a volume $d^3\mathbf{k}$ of the reciprocal space, assumed large compared to the volume $(2\pi)^3/L_xL_yL_z$ per orbital state, is thus:

$$\underset{\text{spin}}{2} \quad \times \quad \underset{\text{number of orbital states}}{\frac{L_xL_yL_z}{(2\pi)^3}\, d^3\mathbf{k}} \quad = \quad n\,(\mathbf{k})\, d^3\mathbf{k}. \tag{2.44}$$

The number of orbital states contained in a band is equal to the number of elementary cells contained by the crystal, independent of the number of atoms per elementary cell.

Let us show this for the face-centered cubic lattice, corresponding to the most common semiconductors. If a is the side of the cube, the volume of the primitive cell of Fig. 2.1 is $a^3/4$, since in each cube we have $(8 \times 1/8) + (6 \times 1/2) = 4$ nodes of the Bravais lattice. A macroscopic parallelepiped L_x, L_y, L_z of the crystal thus contains $N = 4L_xL_yL_z/a^3$ primitive cells. The reciprocal lattice is a body-centered cubic of side $4\pi/a$, with $(8\times1/8)+1 = 2$ nodes, so the volume of the primitive cell, which is also that of the first Brillouin zone, is $(1/2)(4\pi/a)^3 = 32\pi^3/a^3$. The density of orbital states in the reciprocal space is, by Eq. (2.44), $L_xL_yL_z/8\pi^3$. The number of states in a primitive cell of the reciprocal lattice is thus $(32\pi^3/a^3) \times (L_xL_yL_z/8\pi^3) = 4L_xL_yL_z/a^3 = N$, the number of primitive cells in the crystal. Hence each band of index n contains N orbital states, or $2N$ states taking account of spin, where N is the number of primitive cells of the crystal.

To find the state of the whole crystal we fill up the states at zero temperature in accordance with the Pauli principle, starting with the states of lowest energy. The number of electrons per elementary cell, and also the position of the bands in the various directions of the reciprocal space and any overlaps, will fix the behavior, conducting or insulating, of the material. Diamond, silicon, germanium, and grey tin have four valence electrons, and crystallize in a face-centered cubic lattice, with two atoms per elementary cell. We thus have eight electrons for each elementary cell. Among the bands constructed from the valence states for C, Si, Ge, the four lowest bands have no energy overlap with the higher bands. We can thus fill them with the $8N$ electrons with two electrons per orbital. These materials are insulators at zero temperature.

The band gap E_g of diamond is 5.4 eV, that of silicon 1.1 eV, and that of germanium 0.67 eV. As we shall see in Chap. 4 the thermal excitation probability of conduction electrons at temperature T is proportional to $\exp(-E_g/2kT)$ where k is the Boltzmann constant. At room temperature (for $T = 300$ K, i.e., $kT = 25$ meV) this probability is 1×10^{-47} for diamond, 3×10^{-10} for silicon, 1.5×10^{-6} for germanium, while the number of electrons per m^3 in a solid is of order 10^{28}. Diamond is thus a very good insulator, silicon and germanium are semiconductors. By contrast in grey tin the valence band overlaps a higher partially filled band. This makes it a metallic conductor. The conduction properties of semiconductors are determined by the states with energies close to the extrema of the bands since these are the states most easily populated by thermal excitation. We therefore have to count these states.

2.3b Density of States in Energy

Starting from the dispersion relation for a given band $E_n(\mathbf{k})$, we deduce the number dN of states in the band n whose energy is between E and $E+dE$:

$$dN = n(E)\, dE = \int_{\delta v(E)} n(\mathbf{k})\, d^3\mathbf{k} = \frac{L_x L_y L_z}{4\pi^3} \int_{\delta v(E)} d^3\mathbf{k}. \tag{2.45}$$

This relation defines the density of states in energy $n(E)$; $\delta v(E)$ is the volume of the reciprocal space contained between the constant energy surfaces $S(E)$ and $S(E + dE)$. We can easily find the density of states in energy $n(E)$ for a free electron: from the dispersion relation

$$E\,(\mathbf{k}) = \frac{\hbar^2 k^2}{2m} \tag{2.46}$$

we deduce

$$dE = \frac{\hbar^2}{m} k\, dk. \tag{2.47}$$

If the electron moves in three dimensions, the volume $\delta v(E)$ between two constant energy spheres is $4\pi k^2 dk$. Relation (2.44) is then

$$2\left(\frac{L}{2\pi}\right)^3 \cdot 4\pi k^2 dk = n(E)\, dE \tag{2.48}$$

and we get

$$n(E) = 4\pi \left(\frac{L}{h}\right)^3 (2m)^{3/2} \sqrt{E}. \tag{2.49}$$

For an electron in a periodic solid the dispersion relation does not in general have a simple analytic form. The density of states of silicon is shown in Fig. 2.5. The volume element $\delta v(E)$ of Eq. (2.45) can be decomposed into d^2S, the elementary area on $S(E)$, multiplied by the distance along the normal to the surface

$$\begin{aligned} d^3\mathbf{k} &= d^2S \cdot \frac{d\mathbf{k}}{dE} dE \\ &= d^2S \cdot \frac{1}{|\nabla_{\mathbf{k}} E|} dE \end{aligned} \tag{2.50}$$

so that, taking account of the spin,

$$n(E) = \frac{L_x L_y L_z}{4\pi^3} \int_{S(E)} \frac{d^2S}{|\nabla_{\mathbf{k}} E|}. \tag{2.51}$$

If we can define an isotropic effective mass m_e in the vicinity of the minimum E_c of the conduction band, the density of states there can be found directly

from expression (2.49) applying for the free electron, by substituting m_e for m:

$$n(E) = 4\pi \left(\frac{L}{h}\right)^3 (2m_e)^{3/2}\sqrt{E - E_c}. \tag{2.52}$$

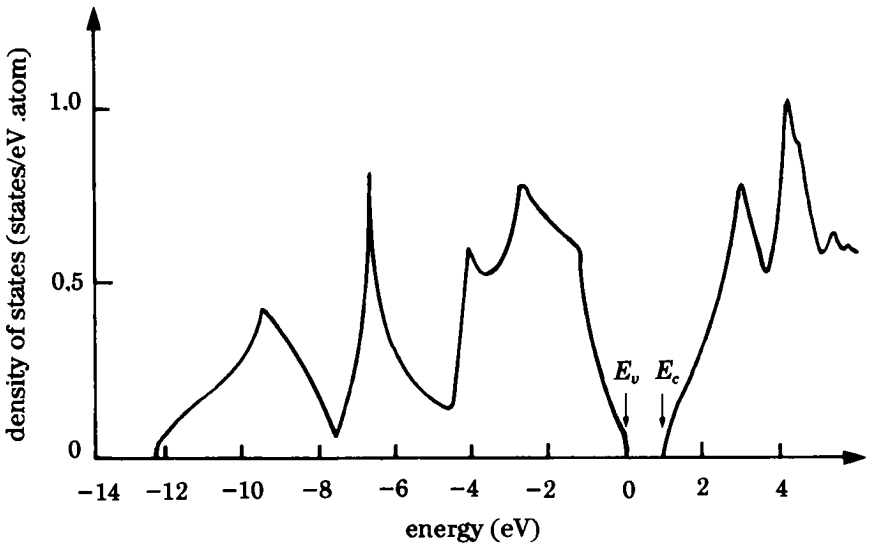


Fig. 2.5. Density of states of the valence and conduction bands of silicon, calculated by J.R. Chelikowsky and M.L. Cohen, Physical Review B **14**, 556 (1976). The energy origin is at the maximum E_v of the valence band. In the neighborhood of E_v, the maximum of the valence band, and E_c, the minimum of the conduction band, the density of states varies parabolically with energy.

2.4 Theoretical Determination of Band Structure

We have seen that one can deduce the dynamical properties of the electrons and the density of states from the relations $E_n(\mathbf{k})$. To understand the properties of semiconductors, we therefore have to determine their band structure. This is a complex task, simultaneously involving first principles, semi-empirical calculations, and experimental data.

We sketch below several methods used to calculate the band structure of semiconductors. These can be skipped in a first reading.

2.4a The Tight Binding Approximation

This is also called the LCAO ("linear combination of atomic orbitals") method. It corresponds to a "chemical" point of view, i.e., to deducing the crystal properties from the already known eigenstates of the constituent atoms and their chemical bonds, which are calculated now.

Simple Cubic Crystal with One Atom per Cell

We consider first the hypothetical case of a simple cubic crystal having one monovalent atom per cell, and assume that the Hamiltonian $\mathcal{H}_{\text{at}}$ of an isolated atom only has one eigenvalue E_0, associated with a non-degenerate eigenstate $\phi(\mathbf{r})$:

$$\mathcal{H}_{\text{at}}\ \phi(\mathbf{r}) = \left[\frac{p^2}{2m} + V(\mathbf{r})\right] \phi(\mathbf{r}) = E_0\ \phi(\mathbf{r}). \tag{2.53}$$

The crystal potential is the sum of the atomic potentials centered at the various sites j of the lattice:

$$\mathcal{H} = \frac{p^2}{2m} + \sum_j V(\mathbf{r} - \mathbf{R}_j) \tag{2.54}$$

and we seek a solution of the Schrödinger equation for the crystal of the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_j A_{\mathbf{k},j}\ \phi(\mathbf{r} - \mathbf{R}_j). \tag{2.55}$$

For simplicity, we will sometimes write

$$\phi(\mathbf{r} - \mathbf{R}_j) = \phi_j. \tag{2.56}$$

Bloch's theorem fixes the form of the $A_{\mathbf{k},j}$: $\psi_{\mathbf{k}}$ must be a Bloch function, so that for each translation $\mathbf{T}$ of the direct lattice we have

$$\begin{aligned}\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{T}) &= \exp\ (i\mathbf{k} \cdot \mathbf{T})\ \ \psi_{\mathbf{k}}\ (\mathbf{r}) \\ &= \sum_j A_{\mathbf{k},j}\ \phi(\mathbf{r} - \mathbf{R}_j + \mathbf{T}).\end{aligned} \tag{2.57}$$

Replacing $\psi_{\mathbf{k}}\ (\mathbf{r})$ by expression (2.55) and setting $\mathbf{R}_{j'} = \mathbf{R}_j - \mathbf{T}$ we obtain

$$A_{\mathbf{k},j+\mathbf{T}} = \exp\ (i\mathbf{k} \cdot \mathbf{T})\ \ A_{\mathbf{k},j}. \tag{2.58}$$

For this relation to hold for all $\mathbf{T}$ we require

$$A_{\mathbf{k},j} = C \exp\ (i\mathbf{k} \cdot \mathbf{R}_j), \tag{2.59}$$

where C is a constant. The normalization of the function $\psi_{\mathbf{k}}$ can thus be written as

$$< \psi_{\mathbf{k}} \, | \psi_{\mathbf{k}} > = C^2 \sum_{j,j'} \exp(-i\mathbf{k} \cdot \mathbf{R}_{j'}) \exp(i\mathbf{k} \cdot \mathbf{R}_j) \times$$

$$\int \phi^*(\mathbf{r} - \mathbf{R}_{j'}) \, \phi(r - \mathbf{R}_j) \, d^3r, \tag{2.60}$$

$$= C^2 \sum_{j,j'} \exp i\mathbf{k} \cdot (\mathbf{R}_j - \mathbf{R}_{j'}) < \phi_{j'} | \phi_j > . \tag{2.61}$$

If we make the hypothesis that states centered on different atoms overlap very little, then

$$< \phi_j | \phi_{j'} > = \delta_{jj'} \tag{2.62}$$

and

$$< \psi_{\mathbf{k}} \, | \psi_{\mathbf{k}} > = C^2 \sum_{j} < \phi_j | \phi_j > = NC^2 = 1, \tag{2.63}$$

hence

$$\psi_k \, (\mathbf{r}) = N^{-1/2} \sum_{j} \exp(i\mathbf{k} \cdot \mathbf{R}_j) \quad \phi \, (\mathbf{r} - \mathbf{R}_j). \tag{2.64}$$

The matrix representing the Hamiltonian (2.54) on the basis of Bloch states is diagonal since the Hamiltonian is a periodic operator (cf. Appendix 2.1). Its eigenvalues are therefore directly given by

$$E(\mathbf{k}) = < \mathbf{k} | \mathcal{H} | \mathbf{k} > = N^{-1} \sum_{j, \, j'} \exp i\mathbf{k} \cdot (\mathbf{R}_j - \mathbf{R}_{j'}) < \phi_{j'} | \mathcal{H} | \phi_j > . \tag{2.65}$$

To evaluate the terms of Eq. (2.65) we decompose the crystal Hamiltonian:

$$\mathcal{H} = \frac{p^2}{2m} + V(\mathbf{r} - \mathbf{R}_j) + \sum_{j'' \neq j} V(\mathbf{r} - \mathbf{R}_{j''}) \tag{2.66}$$

so that

$$E(\mathbf{k}) = E_0 + N^{-1} \sum_{j,j',j'' \neq j} \exp i\mathbf{k} \cdot (\mathbf{R}_j - \mathbf{R}_{j'}) < \phi_{j'} | V(\mathbf{r} - \mathbf{R}_{j''}) | \phi_j >. \tag{2.67}$$

The summation (2.67) contains terms of two types: integrals involving three different centers $j \neq j' \neq j''$, and terms where $j'' = j'$. The terms with three different centers are small as the functions ϕ and the potential V decrease with distance, so we neglect them. There remain either diagonal terms:

$$< \phi_j | \sum_{j' \neq j} V(\mathbf{r} - \mathbf{R}_{j'}) | \phi_j > = -\alpha < 0 \tag{2.68}$$

or non-diagonal terms

$$<\phi_{j'}|V\ (\mathbf{r}-\mathbf{R}_{j'})|\phi_j>= \int \phi^*(\mathbf{r}-\mathbf{R}_{j'})\ V(\mathbf{r}-\mathbf{R}_{j'})\ \phi(\mathbf{r}-\mathbf{R}_j)d^3\mathbf{r} = I_{jj'}. \tag{2.69}$$

Here again, as the atomic functions decrease rapidly with distance, the only non-negligible terms come from the nearest neighbors, for which we set $I_{jj'} = -\gamma$. We thus write $E(\mathbf{k})$ as

$$E(\mathbf{k}) = E_0 - \alpha + N^{-1} \sum_{j' \text{ next to } j} \exp i\mathbf{k}\cdot(\mathbf{R}_j - \mathbf{R}_{j'})\ (-\gamma). \tag{2.70}$$

The quantities α and γ are the fundamental parameters of the chemical bond.

We can calculate γ for the $1s$ states of hydrogen. We find

$$\gamma = 2\ E_1 \left(1 + \frac{a}{a_1}\right) \exp\left(-\frac{a}{a_1}\right), \tag{2.71}$$

where E_1 is the binding energy of the $1s$ level, a_1 the Bohr radius for this level ($E_1 = 13.6$ eV, $a_1 = 0.53$ angstrom) and a the distance between nearest neighbors.

The order of magnitude of the energy spread of the band $E(\mathbf{k})$ is about 2γ. For deep electron states this spread is extremely small as $a >> a_1$. Let us take the case of silicon: $a_1 = (0.53/Z)$ angstrom with $Z = 14$, $a = 2.53$ angstrom (distance between nearest neighbors). We find $\gamma = 6 \times 10^{-24}$ eV, showing that there is only an infinitesimal overlap between orbitals: the $1s$ level remains essentially an atomic level within the solid, without energy dispersion. The valence band of silicon is formed by $3s$ and $3p$ states, but because of the screening of the nuclear charge by electrons in levels $n = 1$ and $n = 2$, we can regard the effective charge as $Z' = 4$. For a $3s$ state, $a_3 = (3 \times 0.53/Z')$ Å $\simeq 0.4$ Å. We thus obtain a broadening or bandwidth of the order of one electron volt.

For the case of a simple cubic lattice where $\mathbf{R}_{j'} - \mathbf{R}_j$ takes the six values $\pm a$ along the three axes

$$E(\mathbf{k}) = E_0' - 2\gamma\left[\cos k_x a + \cos k_y a + \cos k_z a\right] \tag{2.72}$$

with

$$E_0' = E_0 - \alpha.$$

Exercise: Show that the constant energy surfaces given by Eq. (2.72) are orthogonal to the faces of the Brillouin zone.

The width of the band is 12γ. For small $|k|$,

$$E(\mathbf{k}) \simeq E_0' - 6\gamma + \gamma a^2(k_x^2 + k_y^2 + k_z^2). \tag{2.73}$$

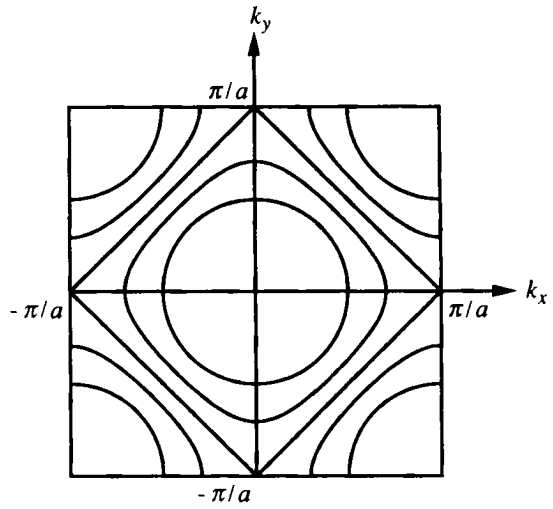


Fig. 2.6. Representation of the constant energy curves for a square lattice in two dimensions. We see that for small k these are circles (spheres in three dimensions).

The constant energy surfaces are therefore spheres. The effective mass (2.36) is isotropic and takes the simple form

$$m^* = \frac{\hbar^2}{2\gamma a^2}. \tag{2.74}$$

For larger values of the energy the form of the constant energy surfaces is shown schematically in Figs. 2.6 and 2.7.

Exercise: Using the above formulas and appropriate formulas from quantum mechanics, find the effective mass at the bottom of the lowest band formed from the lowest hydrogen-like orbital (principal quantum number $n = 1$) for an atom of atomic number Z:

$$\textit{Result:}\quad m^*/m = 2^{-1}(a_1/Za)^2[1 + (a_1/Za)]^{-1}\exp(aZ/a_1). \tag{2.75}$$

Applying the result to silicon, assuming a simple cubic lattice as above, gives for the $1s$ level an enormous value $m^*/m \sim 1.9 \times 10^{21}$! This expresses the fact that the energy is effectively independent of $\mathbf{k}$.

In this calculation, each atomic level of energy E_0 has a corresponding permitted energy band centered around $E_0 - \alpha$. **Permitted bands and forbidden gaps appear only if we couple systems with several atomic levels** (cf. Fig. 1.1). This is the case when several orbitals are taken into account. It is also true for crystals built from dissimilar atoms, or for crystals built from the same kind of atom, but with an asymmetric cell.

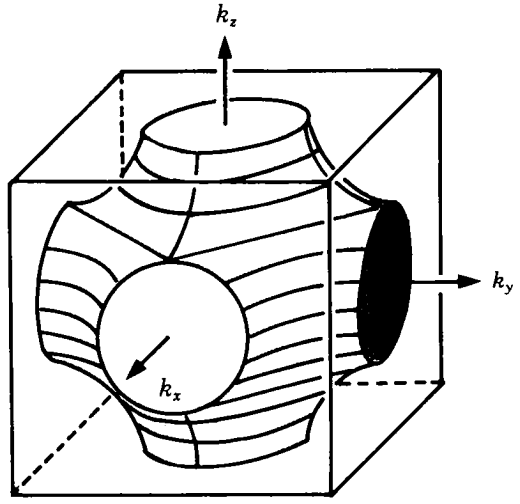


Fig. 2.7. Constant energy surface in the tight binding approximation for a cubic lattice, for the energy value $E(\mathbf{k}) = E'_0$ (Eq. (2.72)).

One-Dimensional AB Crystal

We will bring out some basic ideas for the example of a one-dimensional AB crystal by following step-by-step the above reasoning for the cubic crystal (Eqs. (2.53) to (2.72)).

The crystal is shown schematically in Fig. 2.8. The elementary cell has period $2d$. We take as the quantum state of the valence electron of an isolated atom A (respectively B) the non-degenerate orbital ϕ_A of energy E_A (respectively ϕ_B of energy E_B):

$$\mathcal{H}_A\,\phi_A = \left[\frac{p^2}{2m} + V_A\,(\mathbf{r}-\mathbf{R}_A)\right]\phi_A(\mathbf{r}-\mathbf{R}_A) = E_A\,\phi_A(\mathbf{r}-\mathbf{R}_A),$$
$$\mathcal{H}_B\,\phi_B = \left[\frac{p^2}{2m} + V_B\,(\mathbf{r}-\mathbf{R}_B)\right]\phi_B(\mathbf{r}-\mathbf{R}_B) = E_B\,\phi_B(\mathbf{r}-\mathbf{R}_B). \tag{2.76}$$

The Hamiltonian $\mathcal{H}$ of the crystal is

$$\mathcal{H} = \frac{p^2}{2m} + \sum_{j=1}^{N} V_{Aj} + \sum_{j=1}^{N} V_{Bj}. \tag{2.77}$$

We seek an eigenstate $|\phi_k>$ as a linear combination of the atomic orbitals ϕ_{Aj} and ϕ_{Bj}, centered on the atoms A_j and B_j:

$$|\psi_k> = \sum_{j=1}^{N} \xi_{kj}|\phi_{Aj}> + \eta_{kj}|\phi_{Bj}>\,. \tag{2.78}$$

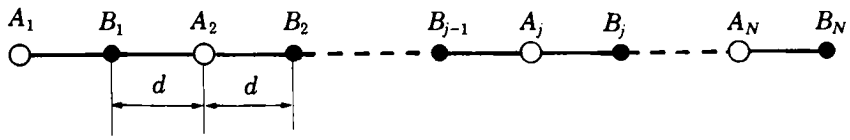


Fig. 2.8. Schematic view of a linear crystal formed by the repetition of atoms A and B at distances d.

The use of Bloch's theorem allows us to write

$$\xi_{kj} = a_k \exp(ik2jd) \quad \text{and} \quad \eta_{kj} = b_k \exp(ik2jd). \tag{2.79}$$

We assume again that the wave functions of neighboring sites overlap only weakly, i.e.,

$$< \phi_{Ai}|\phi_{Aj} > = \delta_{ij}, \quad < \phi_{Bi}|\phi_{Bj} > = \delta_{ij}, \quad < \phi_{Ai}|\phi_{Bj} > = 0. \tag{2.80}$$

The normalized wave function is therefore

$$|\phi_k > = N^{-1/2} \sum_{j=1}^{N} \exp(ik2jd) \ (a_k|\phi_{Aj} > + b_k|\phi_{Bj} >) \tag{2.81}$$

with

$$|a_k|^2 + |b_k|^2 = 1. \tag{2.82}$$

The coefficients a_k and b_k express the weights of the atomic states A and B in ψ_k, i.e., the degree of hybridization. We seek solutions of the eigenvalue equation

$$[\mathcal{H} - E(k)]|\psi_k > = 0. \tag{2.83}$$

It is convenient, as in Eq. (2.66), to write the Hamiltonian $\mathcal{H}$ by separating out one atomic Hamiltonian $\mathcal{H}_{Aj}$:

$$\mathcal{H} = \mathcal{H}_{Aj} + \sum_{j'' \neq j} V_{Aj''} + \sum_{j} V_{Bj}. \tag{2.84}$$

Acting with $< \phi_{Aj}|$ on the left of Eq. (2.83) and using Eqs. (2.78) and (2.84) we write

$$[E_A - E(k)]a_k \exp(2ikjd) + < \phi_{Aj}| \sum_{j'' \neq j} V_{Aj''} + \sum_{j} V_{Bj} \ |\psi_k > = 0. \tag{2.85}$$

The most important terms in the summations involve only the nearest neighbors if we neglect the integrals with three centers and the overlap of distant states. The remaining matrix elements are of four types:

$$\begin{aligned} &< \phi_A \,|\, V_B \,|\, \phi_A > = -\alpha_A \,, \quad < \phi_B \,|\, V_A \,|\, \phi_B > = -\alpha_B, \\ &< \phi_A \,|\, V_A \,|\, \phi_B > = -\gamma_A \,, \quad < \phi_B \,|\, V_B \,|\, \phi_A > = -\gamma_B, \end{aligned} \tag{2.86}$$

where the four quantities $\alpha_A, \alpha_B, \gamma_A$, and γ_B are positive. Moreover we note that $\gamma_A = \gamma_B = \gamma$ since

$$\begin{aligned} < \phi_A \,|\, \frac{p^2}{2m} + V_A + V_B \,|\, \phi_B > &= E_A < \phi_A \,|\, \phi_B > + < \phi_A \,|\, V_B \,|\, \phi_B > \\ &= -\gamma_A \\ &= E_B < \phi_A \,|\, \phi_B > + < \phi_A \,|\, V_A \,|\, \phi_B > \\ &= -\gamma_B. \end{aligned} \tag{2.87}$$

Replacing the wave function $|\psi_k >$ by its expression (2.81) in Eq. (2.85), and limiting ourselves to coupling between nearest neighbors, we get

$$\begin{aligned} a_k \, \exp(ik2jd) \; [E_A - E(k) \; > + < \phi_{Aj} \,|\, V_{Bj-1} + V_{Bj+1} \,|\, \phi_{Aj} >] + \\ b_k \, [\exp(ik2jd) < \phi_{Aj} \,|\, V_{Aj} \,|\, \phi_{Bj} > + \\ \exp[ik2(j-1)d] < \phi_{Aj} \,|\, V_{Aj} \,|\, \phi_{Bj-1} >] = 0 \end{aligned} \tag{2.88}$$

or

$$[E_A - 2\alpha_A - E(k)] \; a_k - \gamma \; [1 + \exp \; (-2ikd)] b_k = 0. \tag{2.89}$$

Similarly the projection of Eq. (2.83) on to $< \phi_{Bj}|$ gives

$$-\gamma \; [1 + \exp(2ikd)] a_k + [E_B - 2\alpha_B - E(k)] b_k = 0. \tag{2.90}$$

Equations (2.89) and (2.90) are two homogeneous equations for two unknowns a_k and b_k. For them to be compatible requires their eigenvalues $E(k)$ to be solutions of the quadratic equation:

$$\begin{aligned} [E_A - 2\alpha_A - E(k)][E_B - 2\alpha_B - E(k)] - \\ \gamma^2 [1 + \exp(-2ikd)][1 + \exp(2ikd)] = 0. \end{aligned} \tag{2.91}$$

This equation has two solutions; there are therefore two energy bands for k in the first Brillouin zone, namely,

$$\begin{aligned} E(k) = \frac{E_A - 2\alpha_A + E_B - 2\alpha_B}{2} \pm \\ \frac{1}{2} \sqrt{(E_A - 2\alpha_A - E_B + 2\alpha_B)^2 + 16\gamma^2 \cos^2 kd}. \end{aligned} \tag{2.92}$$

At $k = 0$ the state of lowest energy E_- is a binding orbital, the state E_+ an antibonding orbital. At the edge of the Brillouin zone, the energies

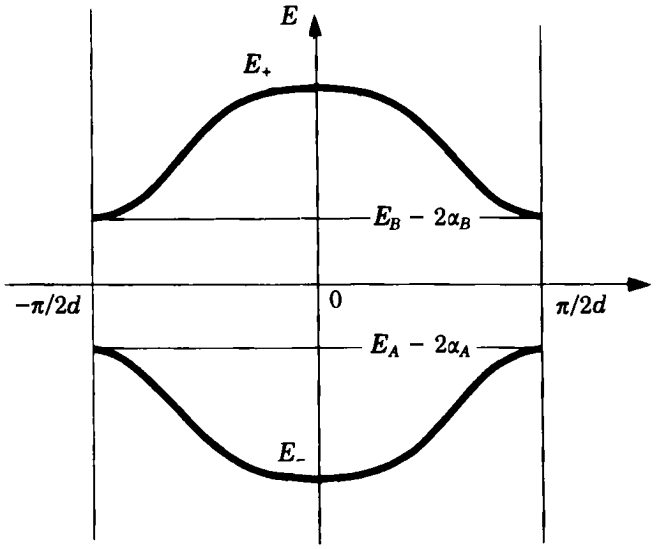


Fig. 2.9. Band structure of the one-dimensional compound AB. We have assumed $E_A < E_B$.

$E_A - 2\alpha_A$ and $E_B - 2\alpha_B$ differ little from those of isolated atoms. The band gap, situated at the edge of the zone, has a width of $|E_B - E_A - 2(\alpha_B - \alpha_A)|$.

By this simple example we have shown that if the elementary cell of a crystal has two different atoms, a band gap appears. Depending on how the bands are filled, i.e., on the number z of electrons per elementary cell, we have a metal (z odd) or an insulator (z even). These two bands are symmetrical, as shown in Fig. 2.9.

Distorted Linear Chain

Let us now take the example of the crystal shown in Fig. 2.10. It consists of just one type of atom A but is distorted.

The period is still $2d$, and we now have $E_A = E_{A'}$, but the values of the matrix elements must be reconsidered: we set

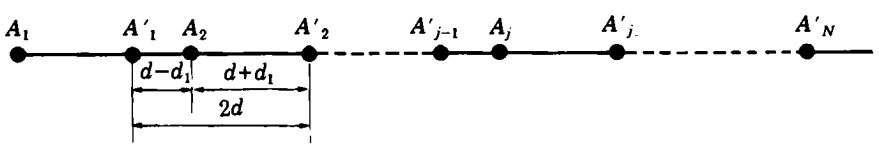


Fig. 2.10. Linear chain made up of one type of atom, but distorted.

$$<\phi_{Aj}|V_{A'j} + V_{A'j-1}|\phi_{Aj}> = -2\alpha_A,$$
$$<\phi_{A'j}|V_{Aj} + V_{Aj+1}|\phi_{A'j}> = -2\alpha_{A'}, \tag{2.93}$$
$$<\phi_{Aj}|V_{Aj}|\phi_{A'j}> = -\gamma(1-\epsilon) = <\phi_{Aj}|V_{A'j}|\phi_{A'j}>,$$
$$<\phi_{Aj}|V_{Aj}|\phi_{A'j-1}> = -\gamma(1+\epsilon) = <\phi_{Aj}|V_{A'j-1}|\phi_{A'j-1}>. \tag{2.94}$$

The equations analogous to Eqs. (2.89) and (2.90) that yield a_k and a'_k, the coefficients of ψ_k, have one non-vanishing solution: we obtain from it the equation giving the eigenvalues $E(k)$:

$$E(k) = E_A - (\alpha_A + \alpha_{A'}) \pm \sqrt{(\alpha_A - \alpha_{A'})^2 + 4\gamma^2[\cos^2 kd\,(1-\epsilon^2) + \epsilon^2]}. \tag{2.95}$$

A band gap opens between the two solutions at $k = \pi/2d$ which exists only if α_A differs from $\alpha_{A'}$ and/or ϵ is non-zero (Fig. 2.11).

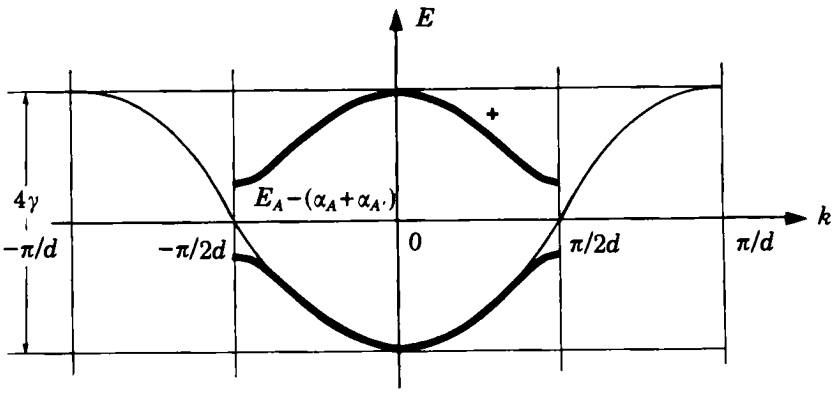


Fig. 2.11. Band structure of the distorted compound of Fig. 2.9. The first Brillouin zone is the segment $[-\pi/2d, \pi/2d[$. $|\alpha_A - \alpha'_A|$ is assumed to be small with respect to $2|\gamma|$.

In the absence of distortion the period of the lattice is d and the first Brillouin zone consists of the segment $[-\pi/d, \pi/d[$. The distortion doubles the spatial period, which becomes $2d$, and reduces the first zone to the segment $[-\pi/2d, \pi/2d[$. The band structure of the compound can be regarded as a folding of the band: the upper branch denoted + then arises from the regions $|k| > \pi/2d$ of the dispersion curve $E(k)$ of the undistorted compound.

If the periodic chain has one electron per cell, the electron states are filled up to $k = \pi/2d$. In the absence of distortion the chain is a conductor. The distortion opens a band gap at precisely this value of k, and the system becomes an insulator. This is the "Peierls transition."

2.4b Plane Wave Expansion

We are interested in states close to the band gap and seek a solution of the Schrödinger equation of the form

$$\psi = \sum_{\mathbf{k}} a_{\mathbf{k}} |\mathbf{k}>, \tag{2.96}$$

where $|\mathbf{k}>$ is any plane wave of the form $\Omega^{-1/2} \exp i\mathbf{k}\cdot\mathbf{r}$, where Ω the volume of the crystal. Substituting this into the crystal Hamiltonian and multiplying by $<\mathbf{k}'|$ we get a set of homogeneous linear equations in $a_{\mathbf{k}}$:

$$\frac{\hbar^2}{2m} k'^2 a_{\mathbf{k}}{}' + \sum_{\mathbf{k}} <\mathbf{k}'|V|\mathbf{k}> a_{\mathbf{k}} = E\, a_{\mathbf{k}}{}'. \tag{2.97}$$

Expanding V in a Fourier series,

$$V = \sum_{\mathbf{G}} V_{\mathbf{G}} \exp i\, \mathbf{G}\cdot\mathbf{r}, \tag{2.98}$$

where $\mathbf{G}$ is a vector of the reciprocal lattice (cf. Appendix 2.1), and regarding the crystal as a parallelepiped of sides L_x, L_y, L_z, we can write

$$\begin{aligned} <\mathbf{k}'|V|\mathbf{k}> &= \frac{1}{\Omega} \sum_{\mathbf{G}} V_{\mathbf{G}} \int \exp\left[i(\mathbf{k}-\mathbf{k}'+\mathbf{G})\cdot\mathbf{r}\right]\, d^3\,\mathbf{r} \\ &= \sum_{\mathbf{G}} V_{\mathbf{G}} \prod_{x,y,z} \frac{2\sin\,(k_x - k'_x + G_x)L_x/2}{(k_x - k'_x + G_x)L_x} \\ &= \sum_{\mathbf{G}} V_{\mathbf{G}}\, \delta_{\,\mathbf{k}-\mathbf{k}'+\mathbf{G}}\,. \end{aligned} \tag{2.99}$$

The matrix element $<\mathbf{k}'|V|\mathbf{k}>$ is then zero unless $\mathbf{k}'-\mathbf{k}$ is a vector of the reciprocal lattice. In Eq. (2.97) we can limit ourselves to summing over those $\mathbf{k}$ which differ from $\mathbf{k}'$ by a vector of the reciprocal lattice. This divides the number of terms by the number of cells in the crystal and is an enormous simplification. But the remaining sum is still in principle infinite. In practice, however, the Fourier components $V_{\mathbf{G}}$ of the potential decrease for large vectors $\mathbf{G}$, and we can truncate the summation at a few hundred terms. There thus remain for each zone point several hundred equations giving several hundred energy eigenvalues and the first few hundred bands. Even with modern computers this method remains relatively inaccurate.

We can greatly improve the convergence by looking for states as linear combinations of **orthogonalized plane waves** $\Phi_{\mathbf{k}}$ defined as

$$\Phi_{\mathbf{k}} = \exp(i\,\mathbf{k}\cdot\mathbf{r}) + \sum_{c} b_c\, \psi^c{}_{\mathbf{k}}\,(\mathbf{r}), \tag{2.100}$$

where $\psi^c{}_{\mathbf{k}}\,(\mathbf{r})$ is a Bloch state corresponding to the core of the ions (hence with very negative energy). We require $\Phi_{\mathbf{k}}$ to be orthogonal to the states $\psi^c{}_{\mathbf{k}}$, i.e.,

$$\int d^3\mathbf{r}\ \psi_{\mathbf{k}}^{c*}\ (\mathbf{r})\ \Phi_{\mathbf{k}} = 0$$

or

$$b_c = -\int d^3\mathbf{r}\ \psi_{\mathbf{k}}^{c*}\ (\mathbf{r})\ \exp(i\ \mathbf{k}\cdot\mathbf{r}). \tag{2.101}$$

The $\Phi_{\mathbf{k}}$ functions are known precisely if we know the $\psi^c{}_{\mathbf{k}}$ accurately. This is true if we use the tight binding method for calculating the deep Bloch states, as the overlap of atomic functions of very negative energies is small. The tight binding approximation is excellent in this case.

The $\Phi_{\mathbf{k}}$ functions have, by construction, the properties we seek: They are orthogonal to the deep states. They have a very localized part which oscillates rapidly, like the wave functions of the atomic core, and between atoms they appear like plane waves. The method then consists of seeking a solution to the Schrödinger equation for the crystal of the form

$$\psi = \sum_{\mathbf{k}} c_{\mathbf{k}}\ \Phi_{\mathbf{k}}. \tag{2.102}$$

Again, the $\Phi_{\mathbf{k}}$ are only coupled by the periodic potential to functions $\Phi_{\mathbf{k}+\mathbf{G}}$ so that this method is analogous to the plane wave method described above. But we obtain excellent results using expansions of ψ in a few tens of orthogonalized plane waves. This is at present the most powerful method of calculating the band structure.

Appendix 2.4 describes another semiempirical method of determining the band structure, the so-called $\mathbf{k}\cdot\mathbf{p}$ method.

2.5 The True Band Structure

Before describing real band structure for silicon we note that Appendix 2.3 shows, using the tight binding method, that crystals of the face-centered cubic form like silicon have triply degenerate bands at $\mathbf{k} = 0$ (levels x_2 and x_6). This occurs at the top of the valence band. We also know from the $\mathbf{k}\cdot\mathbf{p}$ method (Appendix 2.4) that for $\mathbf{k} = 0$ the states resemble atomic states, and are thus possibly degenerate. In fact, we have up to now neglected the role of spin in the electron Hamiltonian by taking account of its influence only via the Pauli principle. It can be shown that in its motion in the electric potential, the spin sees a magnetic field which results in a Hamiltonian of the form

$$\mathcal{H}_{\text{s.o.}} = \frac{1}{4\ m^2\ c^2}\ \boldsymbol{\sigma} \times \mathbf{grad}\ V(\mathbf{r}) \cdot \mathbf{p}, \tag{2.103}$$

where $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli matrices.

This so-called "spin-orbit" interaction has the effect of partially lifting the degeneracy mentioned above. This is seen in Fig. 2.12 which shows the

band structure of silicon around the band gap of width E_g. We note that the maximum of the valence band is doubly degenerate (the point $\Gamma^{25'}$). The spin-orbit interaction has split one of the valence bands by an amount $\Delta E_{\text{s.o.}}$.

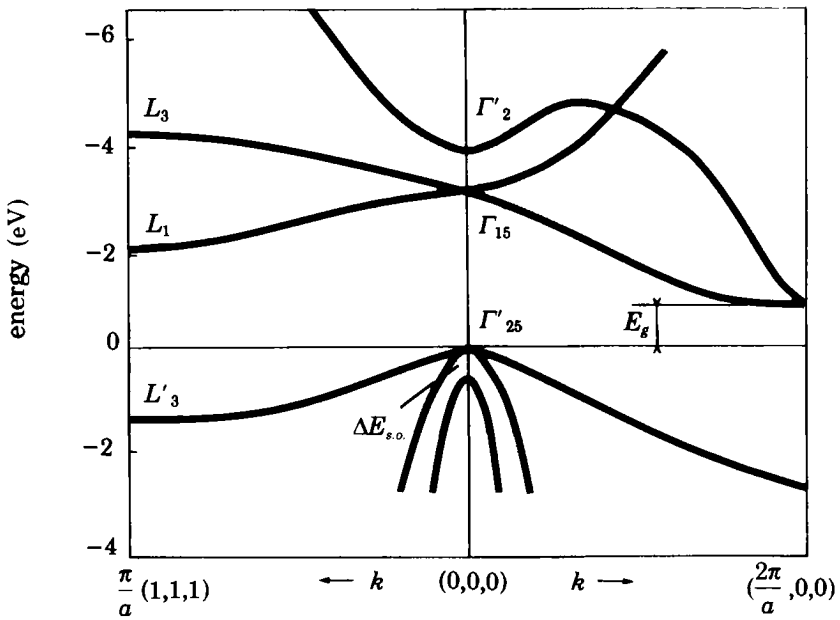


Fig. 2.12. Band structure of silicon: the letters L, Γ denote particular points of high symmetry in the zone. Γ is the center of the zone. L is the point at the edge of the zone in the direction (111). The distance $L_1 L'_3$ is experimentally determined, while the absolute energies of the L'_3 and L_1 levels are found by a calculation by the method of orthogonalized plane waves (D. Brust, J.C. Phillips, and F. Bassani, Physical Review Letters **9**, 94, 1962).

Some of the energy values given in the figure are found from experiment and others from calculations based on the method of orthogonalized plane waves. The form of the constant energy surfaces near the top of the valence band is complex (to describe it we need a $\mathbf{k} \cdot \mathbf{p}$ theory for a degenerate level). We assume that it consists of two spheres, one called "heavy holes" and the other "light holes" (the concept of a hole will be introduced in Chap. 3).

We note that the minimum of the conduction band is not at $\mathbf{k} = 0$, in contrast to the maximum of the valence band. In such a case the gap is called **indirect**. The value of E_g is 1.12 eV for silicon. The mass of the heavy holes is $m_{\text{hh}} = 0.49m$ and the mass of the light holes is $m_{\text{lh}} = 0.16m$.

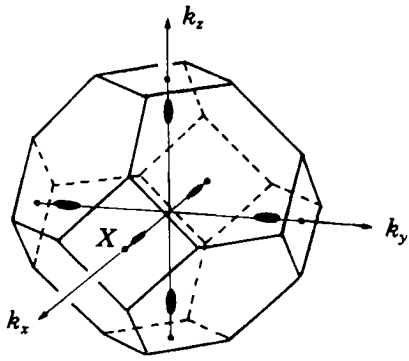


Fig. 2.13. Constant energy surfaces near the bottom of the conduction band of Si.

The minimum of the conduction band is in the direction [100], and by symmetry on the equivalent directions $< 100 >$. There are thus six minima of the conduction band around $|k| \sim 0.8 \times (2\pi/a)$.

By symmetry each ellipsoid of constant energy in the conduction band must have two equal axes. These are prolate ellipses as shown in Fig. 2.13. The dispersion relation near a minimum of the conduction band has the form

$$E(\mathbf{k}) = \frac{\hbar^2}{2}\left(\frac{k_x^2 + k_y^2}{m_T} + \frac{\Delta k_z^2}{m_L}\right) \tag{2.104}$$

for the ellipsoids [001] and [00$\bar{1}$]. Here we have set $\Delta k_z = k_z - k_{z0}$, with $k_{z0} = (0, 0, 1.6\pi/a)$. There appears a longitudinal effective mass $m_L = 0.98m$ and a transverse mass $m_T = 0.19m$.

This shows the complexity of the real situation. However to understand many properties it is often enough to consider a band structure with a direct gap as shown below, using appropriate effective masses. We will call this representation "standard band structure" in the remainder of this book (Fig. 2.14).

2.6 Experimental Study of Band Structure

We confine ourselves to indicating two important methods.

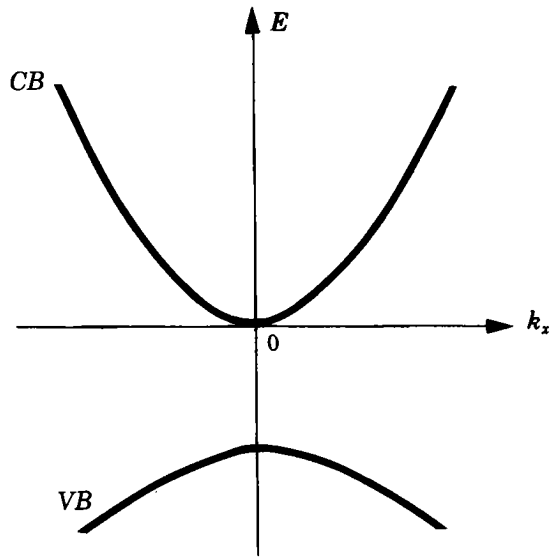


Fig. 2.14. Standard band structure: the maximum of the valence band (VB) and the minimum of the conduction band (CB) are at $k = 0$.

2.6a Optical Methods: Energy Level Determinations

In the same way that atomic spectroscopy helps to determine atomic energy levels, spectroscopy of semiconductors allows us to fix the energy levels of the crystal. Consider first the effect of a light beam of wavelength λ incident on the surface of a semiconductor. If the frequency of the light is such that $h\nu < E_g$ the light beam will traverse the crystal without attenuation. If by contrast $h\nu > E_g$ the photons can be absorbed by excitation of valence-band electrons into the conduction band. In addition a part of the incident beam is reflected.

The intensity of the light beam I varies with distance as

$$I = I_0 \exp(-\alpha x), \tag{2.105}$$

where α is the absorption coefficient (Fig. 2.15). One can show (e.g., Wooten: Optical Properties of Solids, Academic Press, New York, 1972) that the absorption coefficient can be expressed as

$$\alpha \sim (h\,\nu - E_g)^{\gamma}, \tag{2.106}$$

where γ is a constant which depends on the nature of the transitions. A calculation of the absorption coefficient is presented in Sect. 6.1.

Figure 2.16 shows the various possible transitions: (a) vertical permitted transitions between extrema, called direct transitions (see Chap. 4): $\gamma =$

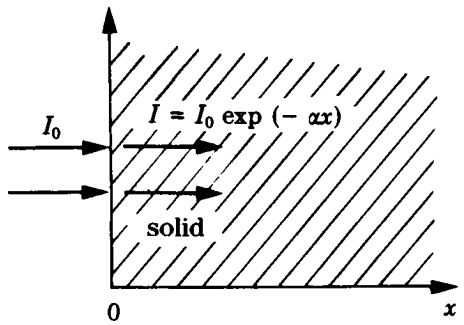


Fig. 2.15. Absorption of a light beam by a solid.

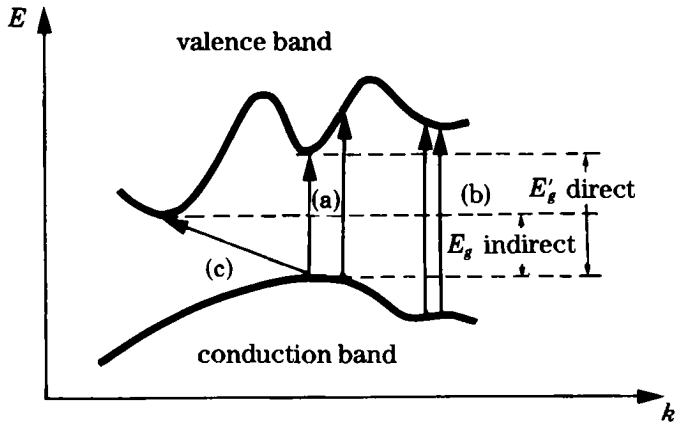


Fig. 2.16. Optical transitions between the valence and conduction bands.

1/2; (b) vertical forbidden transitions: $\gamma = 3/2$; (c) indirect transitions between extrema situated at different points in the zone: $\gamma = 2$.

The latter transitions can only occur if accompanied by the emission or absorption within the crystal of sound waves known as phonons.

When $h\nu$ is much larger than the fundamental absorption these effects overlap for the various gaps E_{g1}, E_{g2} which occur in the band structure.

Absorption curves for various semiconductors are shown in Fig. 2.17. We see that absorption rises very rapidly with energy. We reach absorption coefficients of order 10^5 to 10^6 cm^{-1}, i.e., the beam decreases in intensity by a factor $1/e$ over a distance of 100 or 1000 angstroms. It is then difficult to measure the absorption, which requires a very thin sample, and it is preferable to study the reflectivity of the crystal surface. Each time the photon energy reaches a critical value for the band structure (the distance between band extrema) a structure is observed in the reflectivity. Figure

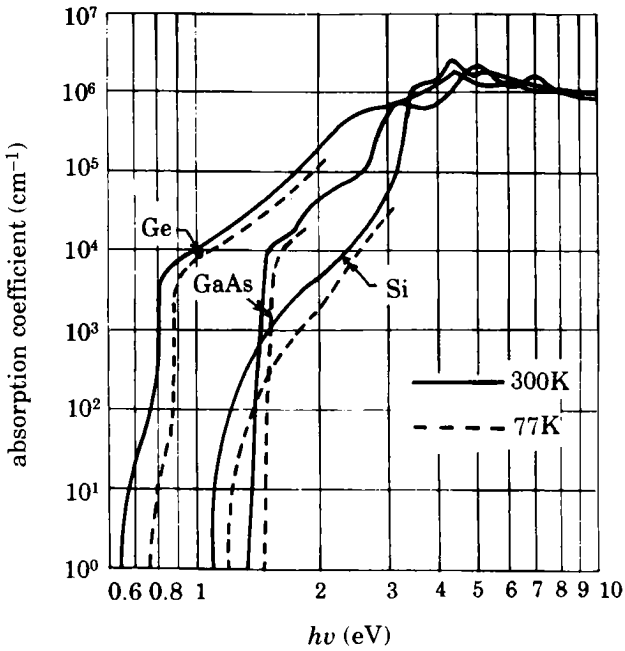


Fig. 2.17. Absorption coefficients for common semiconductors.

2.18 shows a schematic view of a reflectivity experiment, which may use a light source with a non-constant wavelength intensity.

An elliptical mirror with foci A and B is mounted on a rotating axis. The monochromator supplies at A a beam of wavelength λ but whose intensity may vary with λ. As the elliptical mirror rotates, the beam from A alternately falls on the sample under study, when the detector only receives light reflected from the sample, or directly on the detector after a rotation

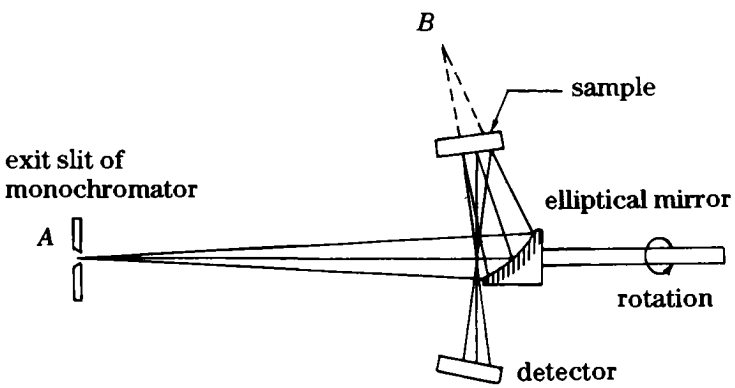


Fig. 2.18. Principle of a reflectivity experiment.

of the mirror through π. The rotation thus produces a sequence of reflected and incident signals at the detector, allowing one to measure their ratio as a function of wavelength.

2.6b The Cyclotron Resonance: Measuring the Effective Mass

Let us consider for example a carrier of charge q placed in a constant magnetic field $\boldsymbol{\mathcal{B}}$ and an alternating electric field orthogonal to $\boldsymbol{\mathcal{B}}$ with frequency ω. Let us assume that its effective mass m^* is isotropic. The equation of motion (2.38) is

$$m^* \frac{d\mathbf{v}}{dt} = q\,(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \boldsymbol{\mathcal{B}}). \tag{2.107}$$

In the absence of the electric field the motion is circular in the plane perpendicular to $\boldsymbol{\mathcal{B}}$ or helical around $\boldsymbol{\mathcal{B}}$ with frequency ω_c:

$$\omega_c = \frac{q\,B}{m^*}. \tag{2.108}$$

We call these orbits "cyclotron orbits." The orbits are helices around the magnetic field. In the presence of an electric field we have to take account of collisions, which we can do by introducing an average frictional force; in the Drude model (Sect. 5.2) this is $-m^*\mathbf{v}/\tau$.

The equation of motion is then

$$m^* \left(\frac{d\,\mathbf{v}}{dt} + \frac{\mathbf{v}}{\tau} \right) = q\,(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \boldsymbol{\mathcal{B}}). \tag{2.109}$$

Setting $B_x = B_y = 0$; $B_z = B$; $\mathcal{E}_y = \mathcal{E}_z = 0$; $\mathcal{E}_x = \mathcal{E}_0 \exp i\omega t$; $v_x = A_1 \exp i\omega t$; $v_y = A_2 \exp i\omega t$, substitution in Eq. (2.109) gives

$$A_1 \left(i\,\omega + \frac{1}{\tau} \right) = \frac{q\,\boldsymbol{\mathcal{E}}_0}{m^*} + A_2\,\omega_c,$$

$$A_2 \left(i\,\omega + \frac{1}{\tau} \right) = -A_1\,\omega_c,$$

and the conductivity for this carrier

$$\sigma_{xx} = \frac{j_x}{\mathcal{E}_x} = \frac{q\,v_x}{\mathcal{E}_x} = \frac{qA_1}{\mathcal{E}_0} = \frac{q^2\,\tau}{m^*} \frac{1 + i\,\omega\,\tau}{1 + \tau^2(\omega_c^2 - \omega^2) + 2\,i\,\omega\,\tau}. \tag{2.110}$$

A resonance appears for $\omega = \omega_c$, in the form of a conductivity increase. This corresponds to a resonant absorption of energy by the carriers, which

occurs when the charge rotates at exactly the same frequency as the electric field. Measurement of the resonant frequency gives a determination of the effective mass (in practice one works at fixed frequency and scans the magnetic field). For this phenomenon to be important requires $\omega_c \tau > 1$, i.e., we have to work with strong magnetic fields and at low temperatures where collisions are rarer. Appendix 3.1 describes the results of a cyclotron resonance experiment for silicon.

Appendix 2.1

Matrix Element of a Periodic Operator between Two Bloch States

Periodicity in Three Dimensions and Fourier Expansion

We show first that the Fourier expansion of a function $f(\mathbf{r})$ which is periodic on a crystal picks out the vectors $\mathbf{G}$ of the reciprocal lattice and no others.

The Fourier expansion of $f(\mathbf{r})$, for any $\mathbf{r}$, can be written

$$f(\mathbf{r}) = \sum n_{\mathbf{K}} \exp i\mathbf{K} \cdot \mathbf{r}. \tag{2.111}$$

The periodicity of $f(\mathbf{r})$, for any translation $\mathbf{T}$ of the direct lattice given by Eq. (2.2), can be written as

$$f(\mathbf{r} + \mathbf{T}) = f(\mathbf{r}) \tag{2.112}$$

or

$$\sum_{\mathbf{K}} n_{\mathbf{K}} \exp(i\mathbf{K} \cdot \mathbf{r}) \exp(i\mathbf{K} \cdot \mathbf{T}) = \sum_{\mathbf{K}} n_{\mathbf{K}} \exp(i\mathbf{K} \cdot \mathbf{r}). \tag{2.113}$$

The relation (2.113) will hold for all $\mathbf{r}$ and $\mathbf{T}$ only for those vectors $\mathbf{K}$ such that

$$\exp i\mathbf{K} \cdot \mathbf{T} = 1. \tag{2.114}$$

This is precisely the property (2.11) of the reciprocal lattice, and $\mathbf{K}$ can be identified with the vector $\mathbf{G}$ defined by Eq. (2.10). Hence the crystal potential has a Fourier expansion involving only the reciprocal lattice vectors. Similarly the periodic part of a Bloch function has the expansion

$$u_{n,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} U_{n,\mathbf{k},\mathbf{G}} \exp i\mathbf{G} \cdot \mathbf{r}. \tag{2.115}$$

Action of a Periodic Operator on a Bloch Function

A periodic operator $A(\mathbf{p},\mathbf{r})$ is an operator for which

$$A(\mathbf{p},\ \mathbf{r}+\mathbf{T}) = A(\mathbf{p},\mathbf{r}). \tag{2.116}$$

If we write it in the form

$$A(\mathbf{p},\mathbf{r}) = \sum_i f_i(\mathbf{r})\ A_i(\mathbf{p}), \tag{2.117}$$

the definition (2.116) requires that the $f_i(\mathbf{r})$ should be periodic functions. Expanding these functions in a Fourier series, we can always write a periodic operator in the form

$$A(\mathbf{p},\mathbf{r}) = \sum_{\mathbf{G}} \exp(i\mathbf{G}\cdot\mathbf{r})\ A_{\mathbf{G}}(\mathbf{p}). \tag{2.118}$$

The action of such an operator on a Bloch function $\exp(i\mathbf{k}\cdot\mathbf{r})\ u_{n,\mathbf{k}}(\mathbf{r})$ will produce a periodic function multiplied by $\exp(i\mathbf{k}\cdot\mathbf{r})$. In fact $A_{\mathbf{G}}\ (\mathbf{p})$ is a combination of derivative operators which transform a periodic function into a periodic function, while retaining a factor $\exp(i\mathbf{k}\cdot\mathbf{r})$. In general, we will have

$$A(\mathbf{p},\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r})\ u_{n,\mathbf{k}}\ (\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r})\ \ \mathcal{U}_{n,\mathbf{k}}(\mathbf{r}), \tag{2.119}$$

where $\mathcal{U}_{n,\mathbf{k}}(\mathbf{r})$ is a periodic function which can be expanded in a Fourier series:

$$\mathcal{U}_{n,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} \mathcal{U}_{n,\mathbf{k},\mathbf{G}}\ \exp(i\mathbf{G}\cdot\mathbf{r}). \tag{2.120}$$

Matrix Elements of a Periodic Operator

We wish to calculate

$$\begin{aligned} <n'\mathbf{k}'|A|n\ \mathbf{k}> &= \int_\Omega u'^{*}_{n',\mathbf{k}'}(\mathbf{r})\ \mathcal{U}_{n,\mathbf{k}}(\mathbf{r})\ \exp[-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}]\ d^3\mathbf{r} \\ &= \sum_{\mathbf{G},\mathbf{G}'} U'^{*}_{n',\mathbf{k}',\mathbf{G}}\ ,\ \mathcal{U}_{n,\mathbf{k},\mathbf{G}}\times \\ &\quad \int_\Omega d^3\mathbf{r}\ \exp i(\mathbf{k}-\mathbf{k}'+\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}. \end{aligned} \tag{2.121}$$

We assume a cubic lattice with lattice constant a, and let $\Omega = L_xL_yL_z$ be the volume of the crystal, taken to be a parallelepiped of edges L_x, L_y, L_z. We calculate the integral

$$\frac{1}{\Omega}\int \exp i(\mathbf{k}-\mathbf{k}'+\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}\; d^3\mathbf{r}$$

$$= \frac{1}{\Omega}\prod_{x,y,z}\int_{-Lx/2}^{+Lx/2} dx \exp i(k_x - k'_x + G_x - G'_x)x$$

$$= \prod_{x,y,z}\frac{2\sin\frac{1}{2}(k_x - k'_x + G_x - G'_x)L_x}{(k_x - k'_x + G_x - G'_x)L_x}. \qquad (2.122)$$

Using the fact that $(\mathbf{G}-\mathbf{G}')$ is a reciprocal lattice vector, $G_x - G'_x = m_x \times 2\pi/a$. Further, $L_x = N_x a$ and the components k_x and k'_x have the forms $n_x 2\pi/L_x$ and $n'_x 2\pi/L_x$ respectively, with N_x, m_x, n_x, n'_x whole numbers; the expression (2.122) then becomes

$$\prod_{x,y,z}\left[\frac{2\sin\pi(n_x - n'_x + m_x\, N_x)}{2\pi(n_x - n'_x + m_x\, N_x)}\right], \qquad (2.123)$$

so that the argument of the sine is always an integer multiple of π. We have factors of the form $\sin X/X$, which are 1 for $X = 0$, and identically zero for $X \neq 0$. We calculate the argument of the sine. If k and k' are in the first Brillouin zone, either they are equal and the above expression becomes $\Pi_{x,y,z}\sin(m_x\pi N_x)/(m_x\pi N_x) = \Pi_{x,y,z}\delta_{m_x}$, or they differ but are within the zone so that $|n_x - n'_x| < N_x$ and the expression vanishes. It follows that

$$\frac{1}{\Omega}\int \exp i(\mathbf{k}-\mathbf{k}'+\mathbf{G}-\mathbf{G}')\mathbf{r}\; d^3\mathbf{r} = \delta_{\mathbf{k},\mathbf{k}'}\;\delta_{\mathbf{G},\mathbf{G}'}, \qquad (2.124)$$

and the expression (2.121) , after replacement of the integral by Eq. (2.124) and carrying out the summation over $\mathbf{G}'$, becomes

$$< n'\;\mathbf{k}'|A|n\;\mathbf{k} > = \Omega\sum_{\mathbf{G}} U^{\prime *}_{n',\mathbf{k}',\mathbf{G}}\,,\; \mathcal{U}_{n,\mathbf{k},\mathbf{G}}\;\;\delta_{\mathbf{k},\mathbf{k}'} \qquad (2.125)$$

which shows that **a periodic operator has zero matrix elements between Bloch functions of different k.**

Appendix 2.2

Symmetries of the Band Structure

Transformation of the Wave Function under Symmetry Operations of the Direct Space

Consider a symmetry operation $\mathcal{R}$ and a state $\psi_{n,\mathbf{k}}$ corresponding to a point $\mathbf{k}$ in the first Brillouin zone. We will show that the application of $\mathcal{R}$ to the wave function rotates the wave vector in the reciprocal space in the same way that the point $\mathbf{r}$ is rotated by the operator R in the real space:

$$\mathcal{R}\ \psi_{n,\mathbf{k}}\ (\mathbf{r}) = \psi_{n,R\mathbf{k}}\ (\mathbf{r}). \tag{2.126}$$

To demonstrate this result we have to define the effect of a symmetry operation on a quantum state in general. Consider for example a rotation in ordinary space (here in two dimensions) taking the point A to the point B (Fig. 2.19(a)). Let R be the geometrical transformation $B = R(A)$ and a physical system, for example an electron, be in the state $\psi_1(\mathbf{r})$ localized around A. In Fig. 2.19(b) the hatched region represents the space where the probability density $\psi_1^*(\mathbf{r})\psi_1(\mathbf{r})$ is not negligible. Rotating the system amounts to bringing it into the state ψ_2 that we seek, and which is represented in Fig. 2.19(c). It is clear from the figure that

$$\psi_2(B) = \psi_1(A) = \psi_1(R^{-1}B). \tag{2.127}$$

This holds for any point $\mathbf{r}$, whence

$$\mathcal{R}\ \psi(\mathbf{r}) = \psi(R^{-1}\mathbf{r}). \tag{2.128}$$

This relation allows us to associate with a symmetry operation R in the ordinary space a (unitary) operation $\mathcal{R}$ in the state space. Applying this relation to $\psi_{n,\mathbf{k}}\ (\mathbf{r})$ we obtain

$$\begin{aligned}\mathcal{R}\ \psi_{n,\mathbf{k}}\ (\mathbf{r}) &= \exp i\mathbf{k}\cdot(R^{-1}\mathbf{r})\ \ u_{n,\mathbf{k}}\ (R^{-1}\mathbf{r}) \\ &= \exp i(R\ \mathbf{k})\cdot\mathbf{r}\ \ u_{n,\mathbf{k}}\ (R^{-1}\mathbf{r})\end{aligned} \tag{2.129}$$

which demonstrates Eq. (2.126). The symmetry operation has given us a new state which is to be associated with the vector $R\mathbf{k}$.

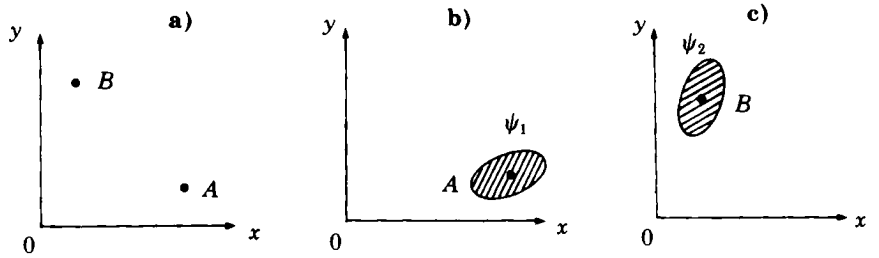


Fig. 2.19. (a) B is the transformation of A under the rotation R. The hatched region (c) is deduced from that of (b) by the same operation.

Symmetry of Constant Energy Surfaces

We show that $\mathcal{R}\psi_{n,\mathbf{k}}(\mathbf{r})$ is an eigenstate of the crystal Hamiltonian $\mathcal{H}$: let us apply the operator $\mathcal{R}$ to each side of the Schrödinger equation:

$$\mathcal{H}\,\psi_{n,\mathbf{k}}(\mathbf{r}) = E_{n\mathbf{k}}\,\psi_{n,\mathbf{k}}(\mathbf{r}), \tag{2.130}$$

$$\mathcal{R}\,\mathcal{H}\,\psi_{n,\mathbf{k}}(\mathbf{r}) = E_{n\mathbf{k}}\,\mathcal{R}\,\psi_{n,\mathbf{k}}(\mathbf{r}). \tag{2.131}$$

As $\mathcal{H}$ is invariant under the symmetry operation $\mathcal{R}$, it commutes with $\mathcal{R}$, so that

$$\mathcal{H}\,\mathcal{R}\,\psi_{n,\mathbf{k}}(\mathbf{r}) = E_{n\mathbf{k}}\,\mathcal{R}\,\psi_{n,\mathbf{k}}(\mathbf{r}) = E_{n\mathbf{k}}\,\psi_{n,R\mathbf{k}}(\mathbf{r}). \tag{2.132}$$

The state $\mathcal{R}\psi_{n,\mathbf{k}}$ deduced from $\psi_{n\mathbf{k}}$ by the symmetry operation corresponds to the same energy and is orthogonal to $\psi_{n,\mathbf{k}}$ since it corresponds to a different point of the zone. We can thus find for each symmetry operation a state degenerate with the original state and situated at $R\mathbf{k}$. This shows that **the constant energy surfaces have the symmetries of the crystal, making it sufficient to study only a fraction of the Brillouin zone.**

Appendix 2.3

Band Structure of Column IV Elements Calculated by the LCAO Method

We give here a first idea of the band structure of column IV elements, calculated by the tight binding method, for energies close to the band gap, and thus obtain a theory of the covalent bond in semiconductors. This calculation formally holds for the sequence diamond, silicon, germanium, and grey tin. (See G. Leman, Annales de Physique, Paris, **7** 1962 p. 505.)

The crystal is face-centered cubic with two atoms per cell, one at the origin, the other translated in the direction [1 1 1] of the quarter **b** of the principal diagonal of the cube (see Figs. 2.1 and 2.2). Each atom is at the center of a regular tetrahedron. The primitive cell, shown in Fig. 2.1, is rhombohedral with basis vectors:

$$\begin{aligned}\mathbf{a}_1 &= (a/2)(\mathbf{i}+\mathbf{j}),\\ \mathbf{a}_2 &= (a/2)(\mathbf{i}+\mathbf{k}),\\ \mathbf{a}_3 &= (a/2)(\mathbf{j}+\mathbf{k}).\end{aligned} \tag{2.133}$$

The electron configuration of column IV atoms is ns^2np^2 with the atomic Hamiltonian

$$\mathcal{H}_{\text{at}} = \frac{p^2}{2m} + V(\mathbf{r}), \qquad \text{or} \quad \mathcal{H}_{\text{at}} = \frac{p^2}{2m} + V(\mathbf{r}-\mathbf{b}), \tag{2.134}$$

and eigenstates ϕ_s, ϕ_p of energies E_s, E_p. We consider

$$\begin{aligned}\phi_0 &= (\phi_s + \phi_x + \phi_y + \phi_z)/2,\\ \phi_1 &= (\phi_s - \phi_x - \phi_y + \phi_z)/2,\\ \phi_2 &= (\phi_s - \phi_x + \phi_y - \phi_z)/2,\\ \phi_3 &= (\phi_s + \phi_x - \phi_y - \phi_z)/2.\end{aligned} \tag{2.135}$$

These functions, called hybrid orbitals sp_3, are not atomic eigenfunctions but form a basis for the tensor product $s \otimes p$. They have the essential property of "pointing," respectively, in the directions [1 1 1], [-1,-1,1],

[-1,1,-1], [1,-1,-1], i.e., toward the vertices of a tetrahedron in the observed direction of the covalent bond. The matrix elements of $\mathcal{H}_{\rm at}$ in this basis are

$$< \phi_i|\mathcal{H}_{\rm at}|\phi_i > = \frac{E_s + 3E_p}{4},$$
$$< \phi_i|\mathcal{H}_{\rm at}|\phi_l > = \frac{E_s - E_p}{4}, \tag{2.136}$$

where l denotes an orbital other than i localized at the same atom.

Similarly we consider hybrid orbitals pointing in the opposite directions:

$$\begin{aligned}
\phi_0' &= (\phi_s - \phi_x - \phi_y - \phi_z)/2 \quad \text{pointing toward } [-1,-1,-1],\\
\phi_1' &= (\phi_s + \phi_x + \phi_y - \phi_z)/2 \quad \text{pointing toward } [+1,+1,-1],\\
\phi_2' &= (\phi_s + \phi_x - \phi_y + \phi_z)/2 \quad \text{pointing toward } [+1,-1,+1],\\
\phi_3' &= (\phi_s - \phi_x + \phi_y + \phi_z)/2 \quad \text{pointing toward } [-1,+1,+1].
\end{aligned} \tag{2.137}$$

We seek a Bloch function solution for the crystal of the form

$$\psi_{n,\mathbf{k}} = C\sum_j \exp(i\,\mathbf{k}\cdot\mathbf{R}_j) \sum_{i=0,1,2,3} [A_i\ \phi_i(\mathbf{r}-\mathbf{R}_j) + A_i'\ \phi_i'(\mathbf{r}-\mathbf{R}_j-\mathbf{b})], \tag{2.138}$$

where the index j denotes the particular site in the lattice, the index i one of the four orbitals Eq. (2.135), and C is a normalization coefficient.

The crystal Hamiltonian is

$$\mathcal{H} = -\frac{\hbar^2}{2m}\Delta + \sum_j V(\mathbf{r}-\mathbf{R}_j) + V(\mathbf{r}-\mathbf{R}_j-\mathbf{b}), \tag{2.139}$$

whence the eigenstate problem:

$$\mathcal{H}\ \psi_{n,\mathbf{k}}\ (\mathbf{r}) = E_{n,\mathbf{k}}\ \psi_{n,\mathbf{k}}\ (\mathbf{r}). \tag{2.140}$$

Consider the functions ϕ_i and ϕ_i' centered, respectively, at the origin and at (1/4, 1/4, 1/4), and let us take their product with Eq. (2.140). We get

$$< \phi_i|\mathcal{H}|n,\mathbf{k} > = E_{n,\mathbf{k}} < \phi_i|n,\mathbf{k} >, \tag{2.141}$$

$$< \phi_i'|\mathcal{H}|n,\mathbf{k} > = E_{n,\mathbf{k}} < \phi_i'|n,\mathbf{k} > . \tag{2.142}$$

We can rewrite Eq. (2.139) in the form

$$\mathcal{H} = \mathcal{H}_{\rm at} + \sum_{j\neq 0} V(\mathbf{r}-\mathbf{R}_j) + \sum_j V(\mathbf{r}-\mathbf{R}_j-\mathbf{b}), \tag{2.143}$$

where $\mathcal{H}_{\rm at}$ is the electron Hamiltonian for the atom at the origin. One can thus see that the only contributions on the left in Eqs. (2.141) and (2.142) come either from the atomic terms or from the interaction term coupling ϕ_i or ϕ_i' to the only orbital from the nearest neighbor which points toward it (neglecting terms of the form $< \phi_i|V_{i\neq j}|\phi_i >$ and interactions between

orbitals centered on two neighboring atoms but not pointing to each other). The interaction or "transfer" integrals are all equal and negative since the potential is attractive and the functions ϕ and ϕ' have the same sign in the region where they overlap. We set

$$< \phi|V|\phi' > = -\lambda. \tag{2.144}$$

Let us consider the site at the origin: ϕ_0 points toward the site $\mathbf{b} = (a/4)(1,1,1)$ and is thus only coupled to the orbital ϕ'_0 centered at $\mathbf{b}$, with the coupling coefficient $\exp i\mathbf{k}\cdot\mathbf{a}_0 = 1$. ϕ_1, being directed along [-1,-1,1] is coupled to the orbital centered at the site $(a/4)(-1,-1,1) = \mathbf{b} - \mathbf{a}_1$, or $\phi'_1[\mathbf{r} - (-\mathbf{a}_1)]$ with the coefficient $\exp i\mathbf{k}\cdot\mathbf{R}_j = \exp(-i\mathbf{k}\cdot\mathbf{a}_1)$, etc. Finally Eqs. (2.141) and (2.142) can be written

$$\left[\frac{E_s + 3\,E_p}{4} - E_{n,\mathbf{k}}\right] A_i + \frac{E_s - E_p}{4}\sum_{l\neq i} A_l - \lambda\,\exp(-i\cdot\mathbf{k}\cdot\mathbf{a}_i)A'_i = 0, \tag{2.145}$$

$$-\lambda \exp i\,\mathbf{k}\cdot\mathbf{a}_i\; A_i + \left[\frac{E_s + 3\,E_p}{4} - E_{n,\mathbf{k}}\right] A'_i + \frac{E_s - E_p}{4}\sum_{l\neq i} A'_l = 0. \tag{2.146}$$

This system of homogeneous equations gives the eight coefficients A_i and A'_i which specify the wave function Eq. (2.138), provided the determinant of the coefficients vanishes. One thus obtains the following secular equation, with

$$x = E_{n,\mathbf{k}} - E_p, \qquad \delta = \frac{E_p - E_s}{4}, \qquad \text{and} \quad \alpha_n = \exp(-i\,\mathbf{k}\cdot\mathbf{a}_n), \tag{2.147}$$

$$\begin{vmatrix} x+\delta & \delta & \delta & \delta & \lambda\,\alpha_0 & 0 & 0 & 0 \\ \delta & x+\delta & \delta & \delta & 0 & \lambda\,\alpha_1 & 0 & 0 \\ \delta & \delta & x+\delta & \delta & 0 & 0 & \lambda\,\alpha_2 & 0 \\ \delta & \delta & \delta & x+\delta & 0 & 0 & 0 & \lambda\,\alpha_3 \\ \lambda\,\alpha_0^* & 0 & 0 & 0 & x+\delta & \delta & \delta & \delta \\ 0 & \lambda\,\alpha_1^* & 0 & 0 & \delta & x+\delta & \delta & \delta \\ 0 & 0 & \lambda\,\alpha_2^* & 0 & \delta & \delta & x+\delta & \delta \\ 0 & 0 & 0 & \lambda\,\alpha_3^* & \delta & \delta & \delta & x+\delta \end{vmatrix} = 0. \tag{2.148}$$

Setting $\Phi = |\frac{1}{4}\Sigma_n\alpha_n|$, a careful calculation shows that Eq. (2.148) can be written as

$$(x^2 - \lambda^2)^2(x^2 + 4\,\delta x - \lambda^2 + 4\,\delta\,\lambda\,\Phi),$$
$$(x^2 + 4\,\delta x - \lambda^2 - 4\,\delta\,\lambda\,\Phi) = 0, \tag{2.149}$$

with

$$\Phi(\mathbf{k}) = \frac{1}{2}\left[1 + \cos\frac{k_x\, a}{2}\cos\frac{k_y\, a}{2} + \cos\frac{k_x\, a}{2} \times \cos\frac{k_z\, a}{2} + \cos\frac{k_y\, a}{2}\cos\frac{k_z\, a}{2}\right]^{1/2} \tag{2.150}$$

Equation (2.149) gives $x(\Phi)$, i.e., the dispersion relation $E_n(\mathbf{k})$.

Resulting Band Structure

We obtain four flat bands (for which E_n does not depend on $\mathbf{k}$) which correspond to the doubly degenerate solutions $x_1 = \lambda$ and to the similarly doubly degenerate solutions $x_2 = -\lambda$.

The broad bands are associated with the other solutions of Eq. (2.149), i.e.,

$$\begin{aligned} x_3 &= -2\,\delta + \sqrt{4\,\delta^2 + \lambda^2 + 4\,\delta\,\lambda\,\Phi}, \\ x_4 &= -2\,\delta - \sqrt{4\,\delta^2 + \lambda^2 + 4\,\delta\,\lambda\,\Phi}, \\ x_5 &= -2\,\delta + \sqrt{4\,\delta^2 + \lambda^2 - 4\,\delta\,\lambda\,\Phi}, \\ x_6 &= -2\,\delta - \sqrt{4\,\delta^2 + \lambda^2 - 4\,\delta\,\lambda\,\Phi}. \end{aligned} \tag{2.151}$$

The band structure has a different shape for $\lambda > 2\delta$ or $\lambda < 2\delta$. It is shown in Fig. 2.20(a) for two directions of the vector $\mathbf{k}$ in the zone and in the special case $\lambda = \delta$. There are $8N$ electrons to be placed in these levels since the primitive cell contains two atoms each possessing four valence electrons. Taking account of the spin degeneracy of each atomic level we see that at zero temperature the bands x_4, x_6, and x_2 are filled. We are thus dealing here with a metal, as the band x_5 is empty and very near in energy to the filled levels.

By contrast if $\lambda > 2\delta$ we have a semiconductor. This is shown in Fig. 2.20(b) for the special case $\lambda = 3\delta$. The $8N$ electrons fill the x_4, x_6, and x_2 bands, and the material is an insulator at zero temperature. The width of the band gap is

$$E_g = x_{5,k=0} - x_2 = 2\,\lambda - 4\,\delta. \tag{2.152}$$

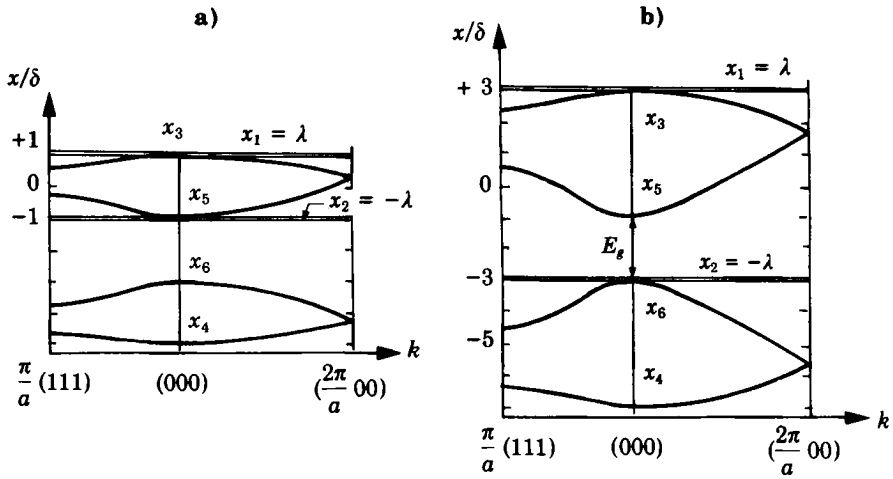


Fig. 2.20. Band structure of an element of column IV of the Periodic Table, calculated by the LCAO method. The transfer integral is $-\lambda$, and the distance between the atomic s and p levels is 4δ. The bands x_4, x_6, and x_2 are full. (a) $\lambda = \delta$; the material is a conductor; (b) $\lambda = 3\delta$; the material is an insulator.

Application: Binding Energy of Semiconductors of Column IV

This very simplified picture of band structure already gives an approximate understanding of several properties of these solids, for example the relation between binding energy and band gap width. The following values are measured for the binding energy E_C and the width of the band gap E_g, expressed in eV (P. Manca, Journal of Physics and Chemistry of Solids **20**, 268, 1961).

	C	Si	Ge	Sn
E_C	14.7	7.55	6.52	5.5
E_g	5.2	1.12	0.66	~ 0

These values are plotted in Fig. 2.21. The experimental points fall approximately on a line with the equation

$$E_C(\text{eV}) = 1.85\ E_g + 5.36. \tag{2.153}$$

In our model the binding energy is the difference between the energy $2E_s + 2E_p$ of an atom (2 s and 2 p electrons) and the energy per atom of the crystal. This energy E_X is, taking account of Eqs. (2.147) and (2.151),

$$E_X = 4\ E_p + 2\ x_2 + (\text{zone average of } x_6 + x_4). \tag{2.154}$$

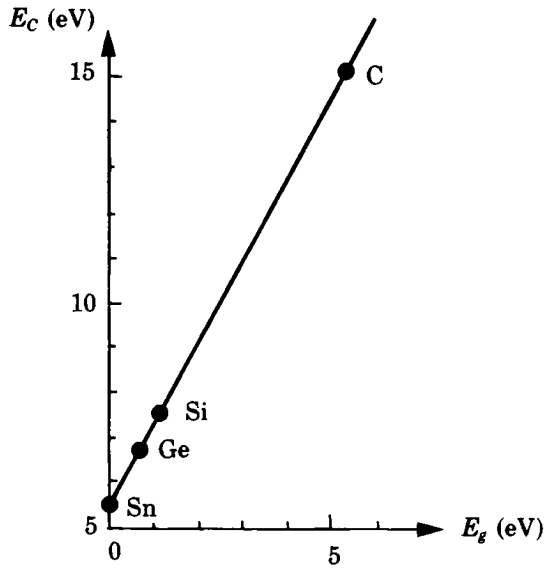


Fig. 2.21. The relation between cohesive energy and band gap width obtained by the LCAO method for the sequence of crystals of elements of column IV of the Periodic Table.

As a first approximation for the average we can take the value $(x_{6,k=0} + x_{4,k=0})$. We then find

$$E_X = 4\,E_p - 4\,\delta - 4\,\lambda, \tag{2.155}$$

and using Eqs. (2.152) and (2.155) the binding energy can be written as

$$E_C = 2(E_s + E_p) - E_X,$$
$$E_C = 2\,E_g + 4\,\delta. \tag{2.156}$$

This relation should be compared with the experimental relation Eq. (2.153). The energy $4\delta = E_p - E_s$ is the excitation energy for an atomic electron from level s to level p. This quantity, provided by atomic physics, is almost constant over column IV and is about 5 eV. The agreement with the experiment is very good. This description of the binding energy neglects the effect of the repulsive term that must exist if the crystal structure is not to collapse. If the repulsive interactions vary very rapidly with distance then the above description holds; this is for example true of the hard-sphere repulsive potential.

We can understand also, at least qualitatively, the influence of the temperature or pressure on E_g. When the temperature decreases or the pressure increases, the mean distance between atoms decreases, raising the value of the integral λ, and simultaneously the band gap.

However, the above calculation does not allow us to get more than the order of magnitude of these variations. The fundamental reason is that by limiting ourselves to combinations of a small number of atomic orbitals and only considering nearest neighbors we are working in a Hilbert space of too small a dimension. Nonetheless, this method gives an excellent approximation for the deep levels in solids.

Remark: for tin (the allotropic form called "grey tin") the gap is actually negative. This means that there is an overlap in energy between the top of the valence bands and the bottom of the conduction band. In this situation the highest energy electrons in the valence band populate the lowest energy states of the conduction band. There can thus exist a metallic-type conductivity due simultaneously to conduction-band electrons and to "holes" or empty places in the valence band, with equal numbers (see Chap. 3). We say that we are dealing with a "semimetal."

Appendix 2.4

The $\mathbf{k}\cdot\mathbf{p}$ Method

The $\mathbf{k}\cdot\mathbf{p}$ method is a semi-empirical method which uses quantities found from experiment in the theoretical calculation of the band structure. We start from the equation

$$\left[\frac{p^2}{2m} + V(\mathbf{r})\right] \psi_{n,\mathbf{k}}\,(\mathbf{r}) = E_{n,\mathbf{k}}\;\psi_{n,\mathbf{k}}\,(\mathbf{r}). \tag{2.157}$$

Replacing $\psi_{n,\mathbf{k}}$ by $u_{n,\mathbf{k}}(\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r})$ we note that

$$\begin{aligned} \mathbf{p}\;\psi_{n,\mathbf{k}} &= \exp(i\mathbf{k}\cdot\mathbf{r})\;\;(\mathbf{p}+\hbar\,\mathbf{k})\;\;u_{n,\mathbf{k}}\,(\mathbf{r}), \\ p^2\;\psi_{n,\mathbf{k}} &= \exp(i\mathbf{k}\cdot\mathbf{r})\;\;(\mathbf{p}+\hbar\,\mathbf{k})^2\;u_{n,\mathbf{k}}\,(\mathbf{r}), \end{aligned} \tag{2.158}$$

and thus

$$\left[\frac{(\mathbf{p}+\hbar\,\mathbf{k})^2}{2m} + V(\mathbf{r})\right] u_{n,\mathbf{k}}\,(\mathbf{r}) = E_{n,\mathbf{k}}\;\;u_{n,\mathbf{k}}\,(\mathbf{r}). \tag{2.159}$$

The periodic part of the Bloch function obeys an equation resembling the original equation apart from the vector $\hbar\mathbf{k}$. We can rewrite Eq. (2.159) in the form

$$\left[\frac{p^2}{2m} + \frac{\hbar\,\mathbf{k}\cdot\mathbf{p}}{m} + \frac{\hbar^2\,k^2}{2m} + V(\mathbf{r})\right] u_{n,\mathbf{k}}\,(\mathbf{r}) = E_{n,\mathbf{k}}\;u_{n,\mathbf{k}}\,(\mathbf{r}). \tag{2.160}$$

For a free electron in a box $V(\mathbf{r}) = 0$ an obvious solution is $u =$ constant or $E_k = \hbar^2k^2/2m$ and $\psi = \Omega^{-1/2}\exp i\mathbf{k}\cdot\mathbf{r}$. Equation (2.160) takes a particularly simple form for $\mathbf{k} = 0$:

$$\left[\frac{p^2}{2m} + V(\mathbf{r})\right] u_{n,0}(\mathbf{r}) = E_{n,0}\;u_{n,0}(\mathbf{r}). \tag{2.161}$$

We note that when the atoms are very far apart, the $E_{n,0}$ are the atomic levels and the $u_{n,0}(\mathbf{r})$ atomic eigenfunctions. We note also that the equation giving $u_{n,0}(\mathbf{r})$ has the symmetries of the crystal potential $V(\mathbf{r})$.

The $\mathbf{k} \cdot \mathbf{p}$ method assumes that we know the values $E_{n,0}$ either from theory or experiment. We then consider small values of $\mathbf{k}$ close to $\mathbf{k} = 0$ and treat the operator $(\hbar/m)\mathbf{k} \cdot \mathbf{p}$ as a perturbation in the Hamiltonian.

Although the $\mathbf{k} \cdot \mathbf{p}$ method is more general, we shall assume for simplicity that the crystal has a center of symmetry (in the diamond structure this center is halfway between two atoms) and we confine ourselves to the study of a given non-degenerate level n.

We first use the symmetry of Eq. (2.161). The Hamiltonian is invariant under inversion, the symmetry operation sending $\mathbf{r}$ to $-\mathbf{r}$. Then if $u_{n,0}(\mathbf{r})$ is an eigenfunction of energy $E_{n,0}$, $u_{n,0}(-\mathbf{r})$ is also an eigenfunction of the same energy, hence also $[u_{n,0}(\mathbf{r}) + u_{n,0}(-\mathbf{r})]$ and $[u_{n,0}(\mathbf{r}) - u_{n,0}(-\mathbf{r})]$. The eigenfunctions can thus be classified into even and odd functions.

In perturbation theory one starts by considering the first-order diagonal matrix elements of the perturbation Hamiltonian $\mathcal{H}_p$. If we consider a non-degenerate level $|n, 0>$ the first-order term $< n, 0|\mathbf{k} \cdot \mathbf{p}|n, 0 >$ vanishes:

$$\int u^*_{n,0}(\mathbf{r}) \frac{\partial}{\partial x} u_{n,0}(\mathbf{r})\, d^3\mathbf{r} = 0 \tag{2.162}$$

because $u_{n,0}$ is either odd or even. We note that $u_{n,0}(\mathbf{r})$ is an eigenfunction of Eq. (2.160) with the eigenvalue $E_{n,0} + \hbar^2 k^2/2m$. There remains only the second-order correction. To this order the energies are given by

$$E_{n,} = E_{n,0} + \frac{\hbar^2 k^2}{2m} + \hbar^2 \sum_{n' \neq n} \frac{< n', 0|\mathbf{k} \cdot \mathbf{p}/m|n, 0 >< n, 0|\mathbf{k} \cdot \mathbf{p}/m|n', 0 >}{E_{n,0} - E_{n',0}}, \tag{2.163}$$

and thus

$$E_{n,\mathbf{k}} = E_{n,0} + \sum_{\alpha,\beta} \frac{\hbar^2}{2m} \left(\frac{m}{m^*}\right)_{\alpha\beta} k_\alpha\, k_\beta, \tag{2.164}$$

with

$$\left(\frac{m}{m^*}\right)_{\alpha\beta} = \delta_{\alpha\beta} + \frac{2}{m} \sum_{n'} \frac{< n', 0|p_\alpha|n, 0 >< n, 0|p_\beta|n', 0 >}{E_{n,0} - E_{n',0}}. \tag{2.165}$$

This so-called $\mathbf{k} \cdot \mathbf{p}$ method allows us to find the effective masses directly either from the energy spectrum at $k = 0$ (the values of $E_{n,0}$) and from the parameters $| < n', 0|p_\alpha|n, 0 > |^2$. This is the most useful theoretical procedure for predicting and analyzing details of the band structure of semiconductors near the band extrema, and thus in the region of interest.

We note that if we know the $u_{n,0}(\mathbf{r})$ we can calculate the parameters directly. Generally these matrix elements are deduced from experiment. The energy differences which appear in the denominator are most usually found from optical absorption or reflectivity.

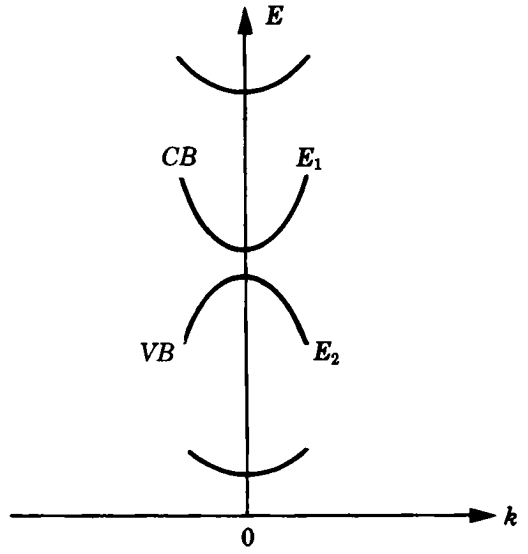


Fig. 2.22. The curvatures at $k = 0$ of the bands E_1 and E_2 are principally determined by their mutual interaction: the other bands are clearly further away, and the energy denominators make their interactions negligible (cf. Eq. (2.166)). In consequence the curvatures, and thus the effective masses of the bands E_1 (conduction) and E_2 (light holes) are practically opposite.

The order of magnitude of m^* is given by

$$\frac{m}{m^*} \sim 1 + 2 < \frac{p_x^2}{m} > \frac{1}{E_g}. \tag{2.166}$$

Here p_x^2/m is of the order of the ionization energy of an atom, for example 5 eV, thus for $E_g = 0.5$ eV, m/m^* will be of order 20.

If we are interested in two bands 1 and 2 close in energy with their extrema at $\mathbf{k} = 0$, we can confine ourselves to including just their interaction in the expression (2.165), the corresponding term being much bigger than all the others (Fig. 2.22).

Example: for a cubic crystal the constant energy surfaces are spheres around $k = 0$, and we have

$$\begin{aligned} \frac{1}{m_1^*} &\simeq \frac{1}{m} + \frac{2}{m^2} \frac{|<1|p_x|2>|^2}{E_1 - E_2}, \\ \frac{1}{m_2^*} &\simeq \frac{1}{m} + \frac{2}{m^2} \frac{|<1|p_x|2>|^2}{E_2 - E_1}, \end{aligned} \tag{2.167}$$

or

$$\frac{1}{m_1^*} = -\frac{1}{m_2^*} + \frac{2}{m}. \tag{2.168}$$

For germanium at $\mathbf{k} = 0$ the "light hole" band (see Chap. 3) has an effective mass $m_2^* = -0.042m$, while the conduction band has an effective mass m_1^* (at $k = 0$) of $0.036m$, in good agreement with the preceding expression: we sometimes say that in the region of small $\mathbf{k}$, these two bands are "mirror images" of each other, since $m_1^* \simeq -m_2^*$.

For the perturbation theory to provide a good approximation the matrix elements of the perturbation $(\hbar/m)\mathbf{k} \cdot \mathbf{p}$ between the functions $u_{n,0}$ and $u_{n',0}$ must be such that

$$\left| \frac{< n,0 \ |\hbar \, \mathbf{k} \cdot \mathbf{p}/m| \ n',0 >}{E_n - E_{n'}} \right| < 1, \tag{2.169}$$

or, for two close bands, from Eqs. (2.167) and (2.169):

$$\frac{1}{m_1^*} \sim \frac{2}{m^2} \frac{| < 1|p_x|2 > |^2}{E_1 - E_2},$$

$$\frac{\hbar \, k_x}{m} < \frac{E_1 - E_2}{| < 1|p_x|2 > |}.$$

Combining these two relations we get

$$\frac{\hbar^2 \, k^2}{2m_1^*} < E_1 - E_2. \tag{2.170}$$

The quantity $\hbar^2 k^2 / 2m_1^*$ is the kinetic energy in the E_1 band. For perturbation theory to give a good approximation, we require the energy in the band to be small compared with the band gap.

Example: $E_1 - E_2 = E_g = 1$ eV. For typical values of the electron energy in the band, of order the thermal energy $kT = 25$ meV at room temperature, the approximation easily holds.

3.

Excited States of a Pure Semiconductor and Quantum States of Impure Semiconductors

A semiconductor at zero temperature is an insulator. At room temperature we know that the system is not in its ground state. We thus have to consider the first excited states of a semiconductor. In these excited states a few electrons occupy the conduction band rather than the valence band, where they leave empty states. These empty states, called "holes," play a fundamental role in the conduction process.

Experiments show that the purity of semiconductors most often determines their behavior. The understanding of quantum states resulting from the presence of impurities is thus essential. Control of the concentration of selected impurities, "doping," is the main engineering tool for practical applications of semiconductors.

3.1 The Hole Concept

Up to now we have considered the ground state of a semiconductor at zero temperature. In this state the valence band is full and the conduction band empty. We are now interested in the first states accessible at non-zero temperature. The simplest excited state has one electron in the conduction band and one empty place in the valence band. Such a state can be obtained at low temperature by illuminating the crystal with electromagnetic radiation of energy greater than the width of the band gap. A photon can then be absorbed and excite an electron from the valence band into the conduction band.

Consider a solid in which we have created such a pair (electron in the conduction band + empty place in the valence band). The electron unbound from a covalent bond and placed in the conduction band (Fig. 3.1(a)) can

then carry a current in the presence of an electric field because there exist nearby empty energy levels. If we put this electron into a (non-stationary) state formed by a wave packet we can consider that there is a Si^- ion in the crystal and that this charge can move in the crystal by displacing the extra electron from one atom to another (Fig. 3.1(b)).

Now consider the valence band. There are normally four electrons per atom, the bond being shown schematically in Fig. 3.1(a). If we have formed a Si^- ion somewhere in the crystal, there remains the equivalent of a Si^+ ion at the place where the bond was broken. This represents the lack of an electron in a valence bond. It is clear that in the presence of an electric field the empty place can be filled by an electron from another bond which moves under the effect of the field. We then have the equivalent of the displacement of Si^+ in the direction of the electric field, hence the motion of a positive charge (Fig. 3.1(b)). **We see that we can speak of the motion of the lack of an electron as the displacement of a positive charge which we call a hole.**

We now describe the notion of a positive hole more precisely. We have seen in Sect. 2.3, Eqs. (2.42) and (2.43), that the total current of the electron states of a given band is zero. Let us consider a valence band with a single empty place at the state $\mathbf{k} = \mathbf{k}_e$. The current $\mathbf{j}_{\text{total}}$ corresponding to the sum of the states of this band can be decomposed as

$$\mathbf{j}_{\text{total}} = -e \sum_{\text{occupied } \mathbf{k}} \mathbf{v}(\mathbf{k}) - e\, \mathbf{v}(\mathbf{k}_e) = 0. \tag{3.1}$$

In the present case where $\mathbf{k}_{\text{empty}} = \mathbf{k}_e$, the current $\mathbf{j}$ transported by the occupied valence states is then (using the definition of $\mathbf{v}(\mathbf{k}_e)$)

$$\begin{aligned} \mathbf{j} &= e\, \mathbf{v}(\mathbf{k}_e) \\ &= \frac{e}{\hbar} \nabla_{\mathbf{k}}\, E_e(\mathbf{k}_e). \end{aligned} \tag{3.2}$$

$E_e(\mathbf{k}_e)$ is the dispersion relation of the valence electrons at the point $\mathbf{k}_e$.

We define a quasi-particle, the hole, by

hole = [valence band full except for one empty state].

The current carried by the hole is given by Eq. (3.2) above. If E is the total energy of the full valence band, the energy of the system [full band except for one empty state at the point $\mathbf{k}_e$] is $E - E_e(\mathbf{k}_e)$. The deeper the empty state lies in the valence band the larger the energy of the system (as $E_e(\mathbf{k}_e)$ is smaller). **The energy E of the quasi-particle we call the hole can thus be defined as**

$$E_h = -E \,(\textbf{electron missing in the state } \mathbf{k}_e) + \textbf{constant}. \tag{3.3}$$

We wish to attribute a positive charge to the hole. For the hole current to have the form (3.2), the wave vector $\mathbf{k}_h$ must be equal to $-\mathbf{k}_e$, i.e., opposite to that of the missing electron. Then we can write

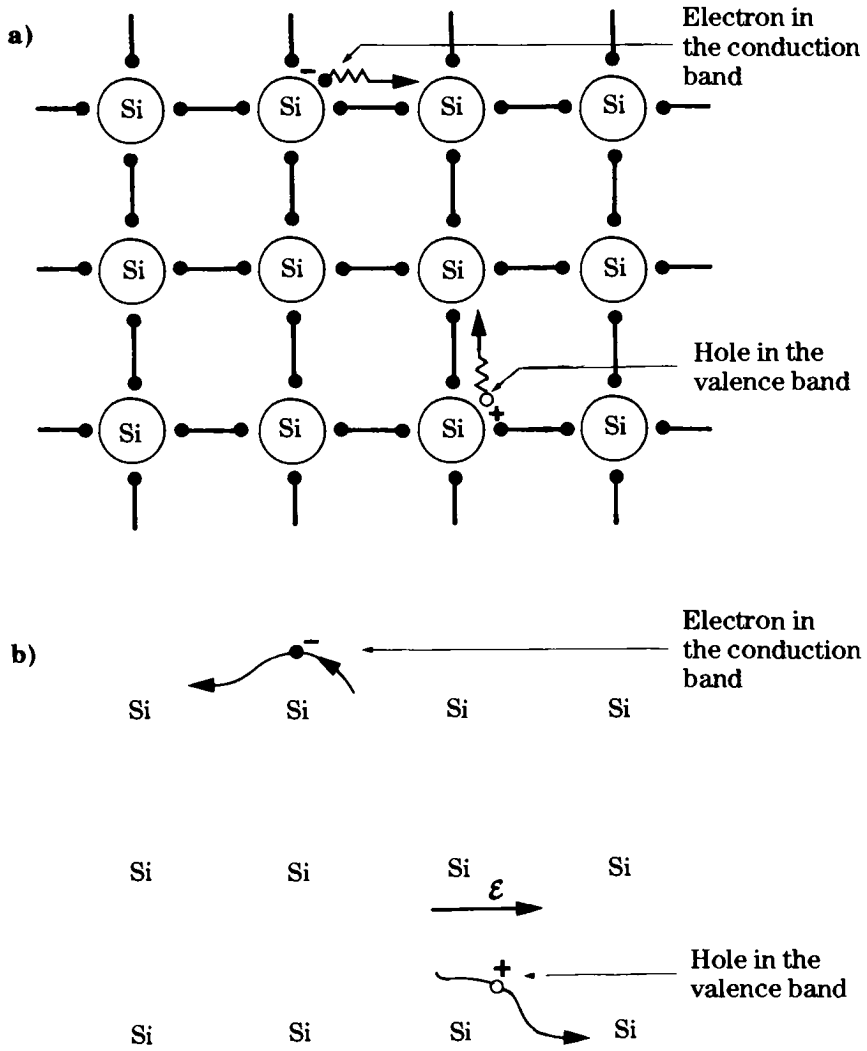


Fig. 3.1. (a) Formation of an electron-hole pair in silicon by absorption of a photon. A thick bar symbolizes a homopolar chemical bond, in which an electron pair with antiparallel spins is shared between two atoms; (b) displacement of these charges under the action of an applied electric field $\mathcal{E}$.

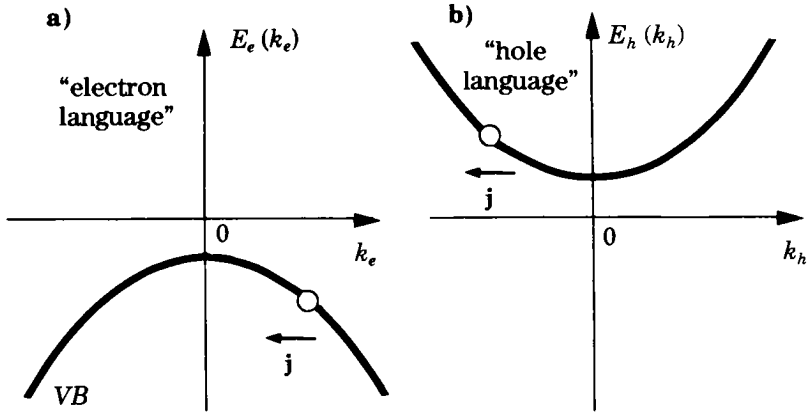


Fig. 3.2. Definition (a) in "electron language"; (b) in "hole language" of the wave vector, energy, and current of a hole in the valence band.

$$\mathbf{j} = \frac{e}{\hbar}\nabla_{\mathbf{k}e}\, E_e(\mathbf{k}_e) = \frac{e}{\hbar}\nabla_{\mathbf{k}h}\, E_h(\mathbf{k}_h). \tag{3.4}$$

Figure 3.2 illustrates these definitions. The hole velocity is

$$\mathbf{v}_h = \frac{1}{\hbar}\nabla_{\mathbf{k}h}\, E_h(\mathbf{k}_h). \tag{3.5}$$

This is the velocity of the electron missing at $\mathbf{k}_e$. The dynamical equation, i.e., the evolution of $\mathbf{k}_h$ and $\mathbf{v}_h$ in the presence of electric and magnetic fields, can be found from Eq. (2.33) with a change of sign, since $\mathbf{k}_h = -\mathbf{k}_e$:

$$\hbar\frac{d\,\mathbf{k}_h}{dt} = -\hbar\frac{d\,\mathbf{k}_e}{dt} = e\,(\boldsymbol{\mathcal{E}} + \mathbf{v}_h \times \boldsymbol{\mathcal{B}}). \tag{3.6}$$

This is the equation of motion of a positive charge of wave vector $\mathbf{k}_h$, moving with velocity $\mathbf{v}_h$, under fields $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$.

The dynamical equation in the real space results from Eq. (3.5) and the preceding equation. This allows us to define the effective mass of the hole:

$$\frac{dv_{h\alpha}}{dt} = \sum_{\beta}\left(\frac{1}{m_h^*}\right)_{\alpha\beta} F_\beta \text{ with } \left(\frac{1}{m_h^*}\right)_{\alpha\beta} = \frac{1}{\hbar^2}\frac{\partial^2 E_h}{\partial k_{h_\alpha}\,\partial k_{h_\beta}}. \tag{3.7}$$

The effective mass of the hole is positive in the region of the Brillouin zone where the function $E_e(\mathbf{k})$ has a negative second derivative, i.e., near maxima. This holds in particular for the top of the valence band of semiconductors.

In summary, the electron-hole correspondence is as follows:

$E_h = -E$ of the missing electron,
$\mathbf{k}_h = -\mathbf{k}$ of the missing electron,
hole velocity = velocity of missing electron,
hole charge = +e,
effective hole mass = − effective mass of missing electron.

The effective mass of the hole is thus positive for a negative curvature of $E(\mathbf{k})$, which holds near a maximum of $E(\mathbf{k})$, i.e., at a band maximum. We then have

$$\mathbf{v}_h = \frac{1}{\hbar} \nabla_{\mathbf{k}_h} E_h(\mathbf{k}_h), \tag{3.8}$$

$$\mathbf{j} = e\, \mathbf{v}_h, \tag{3.9}$$

$$\hbar \frac{d\, \mathbf{k}_h}{dt} = e(\boldsymbol{\mathcal{E}} + \mathbf{v}_h \times \boldsymbol{\mathcal{B}}), \tag{3.10}$$

$$\frac{d\, \mathbf{v}_{h,\alpha}}{dt} = \sum_{\beta} \left(\frac{1}{m_h}\right)_{\alpha\beta} e(\boldsymbol{\mathcal{E}} + \mathbf{v}_h \times \boldsymbol{\mathcal{B}})_\beta. \tag{3.11}$$

An experimental determination of the effective masses of electrons and holes in silicon by means of cyclotron resonance is described in Appendix 3.1 (cf Sect. 2.6d).

We can qualitatively illustrate these notions by the example of photoconductivity: optical excitation by direct transition of an electron in a state $\mathbf{k}_e$ of the valence band into a state of the same wave vector of the conduction band leaves a hole in the valence band of wave vector $\mathbf{k}_h = -\mathbf{k}_e$. Let us apply an electric field $\boldsymbol{\mathcal{E}}$ to this system; we show that the currents from the electron and hole add together. For a standard band scheme like that of Fig. 3.3 the effective masses m_e of the electrons and m_h of the holes are positive and isotropic.

Using the acceleration theorem in the real space [in the present case Eq. (3.38)], the electron motion in the presence of $\boldsymbol{\mathcal{E}}$ is given by

$$m_e \frac{d\mathbf{v}_e}{dt} = -e\, \boldsymbol{\mathcal{E}}, \tag{3.12}$$

giving the velocity and current after time Δt:

$$\Delta\, \mathbf{v}_e = -\frac{e\, \Delta t}{m_e} \boldsymbol{\mathcal{E}}, \tag{3.13}$$

$$\Delta\, \mathbf{j}_e = -e\, \Delta\, \mathbf{v}_e = \frac{e^2 \Delta t}{m_e} \boldsymbol{\mathcal{E}}. \tag{3.14}$$

For the hole,

$$m_h \frac{d\mathbf{v}_h}{dt} = e\, \boldsymbol{\mathcal{E}}, \tag{3.15}$$

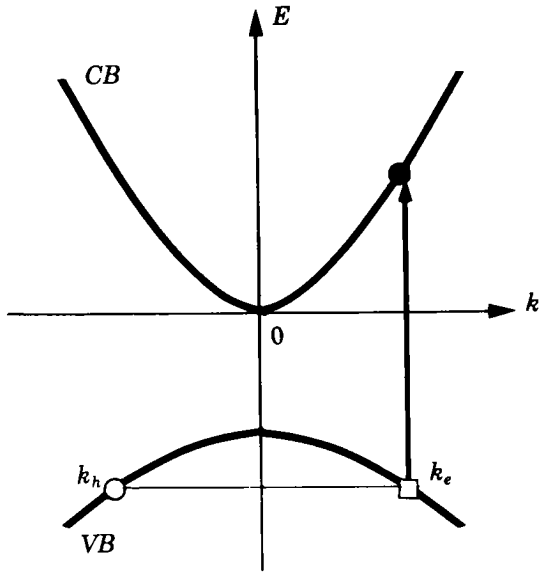


Fig. 3.3. Excitation of an electron-hole pair by direct optical transition.

$$\Delta \mathbf{v}_h = \frac{e\,\Delta t}{m_h}\boldsymbol{\mathcal{E}}, \tag{3.16}$$

$$\Delta \mathbf{j}_h = \frac{e^2 \Delta t}{m_h}\boldsymbol{\mathcal{E}}. \tag{3.17}$$

The electric drift currents for the electron and the hole have the same sign and add together. Figure 3.4 shows this result schematically. Transport problems, such as the calculation of conductivity, will be treated in detail in Chap. 5.

We return now to the real band structure of semiconductors (Sect. 2.4d). The valence band of Si or GaAs is actually degenerate in the neighborhood of its extremum at $k = 0$ (see Fig. 2.12). There are then two hole systems, and hence two effective hole masses. We distinguish heavy holes of mass m_{hh} and light holes of mass m_{lh}. In silicon, $m_{hh} = 0.49m$ and $m_{lh} = 0.16m$.

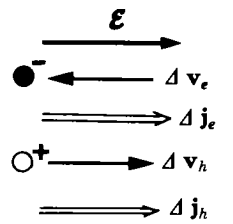


Fig. 3.4. Electric drift currents deriving for the electron ($\Delta \mathbf{j}_e$) and hole ($\Delta \mathbf{j}_h$) in the presence of an applied electric field $\boldsymbol{\mathcal{E}}$.

3.2 Impurities in Semiconductors

Until now we have only considered perfect crystalline solids, without defects or impurities. What can we say about imperfect crystals? There is no general answer to this question: it all depends on the nature of the defects or impurities. First, they act as scattering centers, scattering electrons because they break the periodicity of the crystal potential. Second, and even more important, impurities play the essential role of modifying the electron content in semiconductors: a pure silicon crystal at room temperature would "naturally" (i.e., in thermal equilibrium) contain of 10^{-12} free electrons per atom. In real crystals the majority of the free electrons will in fact come from impurities.

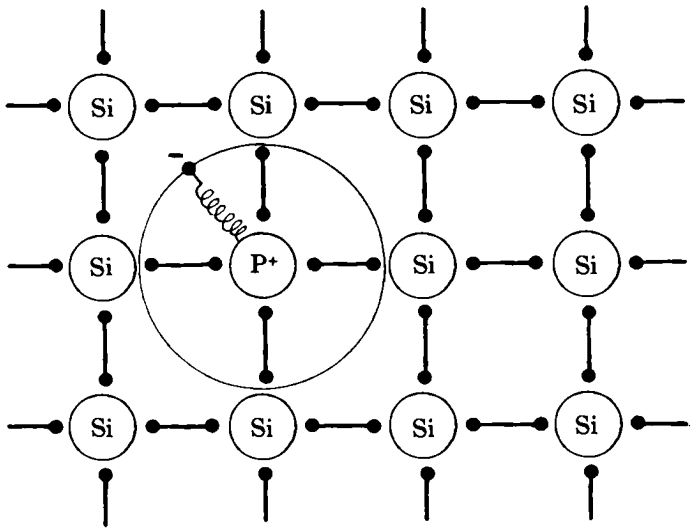


Fig. 3.5. Phosphorus atom substituted in a silicon lattice. Each black dot denotes an electron. At low temperature the extra electron is bound to the phosphorus nucleus. The radius of the orbit is in reality much larger than the interatomic distance. At high temperature the electron is released into the conduction band: phosphorus is a donor.

Consider for example a phosphorus atom replacing a silicon atom in a crystal (Fig. 3.5). This replacement can occur easily since the atoms have approximately the same size. We call this a substitutional impurity. The phosphorus atom has five valence electrons. We can to first approximation assume that four of these electrons, which fill states fairly similar to those of silicon, will participate in four covalent bonds with the four neighboring

atoms. The P–Si bonds differ little from the Si–Si bond. These four electrons thus form part of the valence band by replacing the four silicon electrons which have disappeared. But there remains an electron, normally bound to the phosphorus atom which has an additional nuclear charge $+e$ (Fig. 3.5).

The "internal" ionization of this system consists of the removal of the electron from its state bound to the phosphorus nucleus into the conduction band (Fig. 3.5). **Atoms which can give an extra electron to the crystal on ionization are called donors.** There is a "pseudo-atom" of hydrogen in the medium. Its internal "ionization" energy will at most be equal to the width of the band gap.

This result is important as it shows that impurities in a semiconductor will self-ionize at temperatures lower than those required for intrinsic ionization between the valence and conduction bands. The very reason for this is that the energy necessary for this process is not the ionization energy of a phosphorus atom in vacuum (about 10 eV) but the energy to ionize (i.e., unbind an electron from the positive charge) within the semiconductor. This requires at most an energy of order E_g, the width of the band gap, about one electron volt. The ionization energy for a hydrogen-like system is of the order of $e^2/8\pi\epsilon_0 a$, where a is the orbital radius. If the energy is reduced by a factor of 10, the size of the bound orbit should be of the order of $10a_1$, where a_1 is the radius of the first Bohr orbit of the hydrogen atom. As $a_1 = 0.53$ angstrom the size of the orbit is about 5 angstroms. But the reasoning above is no longer valid as the attracting force seen by an electron is no longer $-e^2/4\pi\epsilon_0 r^2$ but $-e^2/4\pi\epsilon_0\epsilon_r r^2$, where ϵ_r is the relative dielectric constant of the medium. Here ϵ_r is of the order of 10, and the potential is reduced by a factor of 10. Because of this the orbit is still bigger. Moreover the electron is then bound by a potential varying slowly over interatomic distances. We know that the response to a force of this type involves not the free electron mass but the effective electron mass in the crystal.

We may then regard this system as a pseudo-hydrogen atom, with the Hamiltonian

$$\mathcal{H} = \frac{p^2}{2\,m^*} - \frac{e^2}{4\,\pi\,\epsilon_0\,\epsilon_r\,r}. \tag{3.18}$$

The energy eigenvalues are given by quantum mechanics as

$$E_n = -\frac{e^4}{2\,(4\,\pi\,\epsilon_0)^2\,\hbar^2\,n^2} \times \frac{m^*}{\epsilon_r^2}. \tag{3.19}$$

The Bohr radius of the orbit of quantum number $n = 1$ is

$$a_1^* = \frac{4\,\pi\,\epsilon_0\,\epsilon_r\,\hbar^2}{e^2\,m^*} \tag{3.20}$$

or

$$a_1^* = \epsilon_r \left(\frac{m}{m^*}\right) a_1, \quad \text{with } a_1 = 0.53\text{Å}. \tag{3.21}$$

For a state with $n > 1$

$$E_n = \frac{m^*}{m} \frac{1}{\epsilon_r^2} \frac{E_{1H}}{n^2}, \quad \text{where } E_{1H} = -13.6 \text{ eV}. \tag{3.22}$$

and the wave function extension is of order $n^2 a_1^*$.

We see that the binding energy E_1 of the ground state is greatly reduced since $\epsilon_r = 16$ for germanium and 11.7 for silicon; $m^*/m \sim 0.2$ for germanium and 0.4 for silicon. We thus predict from this simple theory an ionization energy independent of the nature of the donor, of 0.01 eV for elements of column V (P, As, Pb, Bi) in germanium and an energy of 0.04 eV in silicon (see the excerpt from the Periodic Table). This is the energy that must be supplied to the electron to ionize the atom in the crystal. The donor occupies a level at a distance from the conduction band small compared with the gap; we call this a "shallow donor."

Excerpt from the Periodic Table

columns: IIb	III	IV	V	VI
	B	C	N	O
	Al	Si	P	S
Zn	Ga	Ge	As	Se
Cd	In	Sn	Sb	Te

The following table shows the very good agreement of the experimental values of E_1 in germanium and silicon, given in meV, with this theory:

	P	As	Sb	Bi
Ge	12	12.7	9.6	
Si	44	49	39	69

The radius of the ground state orbit is increased by the factor $\epsilon_r m/m^*$, which is about 50. This justifies the above approximations.

We can argue in a similar fashion for the substitutional impurities of group III of the Periodic Table (boron, aluminium, gallium, thallium). Group III atoms have only three valence electrons. Thus to fill all the valence states and realize for example four B–Si bonds, an electron has to be taken from a nearby Si–Si bond. For this reason **a substitutional element of column III in a semiconductor of column IV is an acceptor**. The electron taken from the Si–Si bond belongs to the valence band, and leaves a hole in its place. The binding energy of the captured electron is large, since this is the energy of the chemical valence bond. After the capture of this electron the ensemble (boron atom + captured electron) is negatively charged and attracts the free hole. This is shown in Fig. 3.6. The resulting bound state is the lowest energy state of the system. The acceptor is therefore neutral, since the electron captured to satisfy the bond and the hole localized near

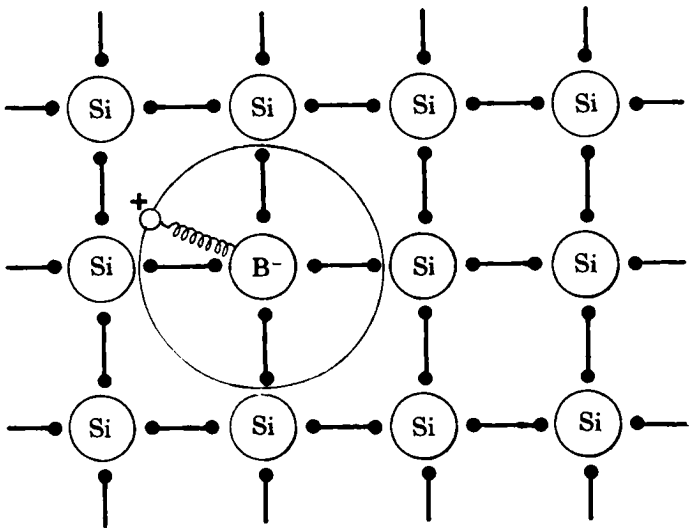


Fig. 3.6. Boron atom in a substitutional position in a silicon lattice. At low temperature the hole is bound to the boron nucleus in an orbit which covers many atomic sites. At high temperature the hole is released into the valence band; boron is an acceptor.

the boron atom form a system of zero net charge. To dissociate this neutral acceptor requires energy, but for the same reasons as for the donors the binding energy of the hole to the charged center of the impurity is weak. It is given by expression (3.22), where we have to replace the effective mass of the electron by the effective mass of the hole.

At very low temperature the hole will remain fixed in a hydrogen-like orbit and the crystal will not be a conductor, but **at room temperature the system will be ionized and the crystal will have a free hole** in the valence band for each element of column III present in the crystal. Then the center B will be negatively charged because of the captured electron.

The table below gives the experimental values in meV of the ionization energies of acceptors in germanium and silicon.

	B	Al	Ga	In	Tl
Ge	10.4	10.2	10.8	11.2	10
Si	46	57	65	160	260

The energy levels due to the presence of donors or acceptors are illustrated in Fig. 3.7.

The level E_d is below the minimum E_c of the conduction band (Fig. 3.7(a)). The fundamental level E_a of the hole should have a negative energy relative to the top of the valence band (in hole language), thus a positive

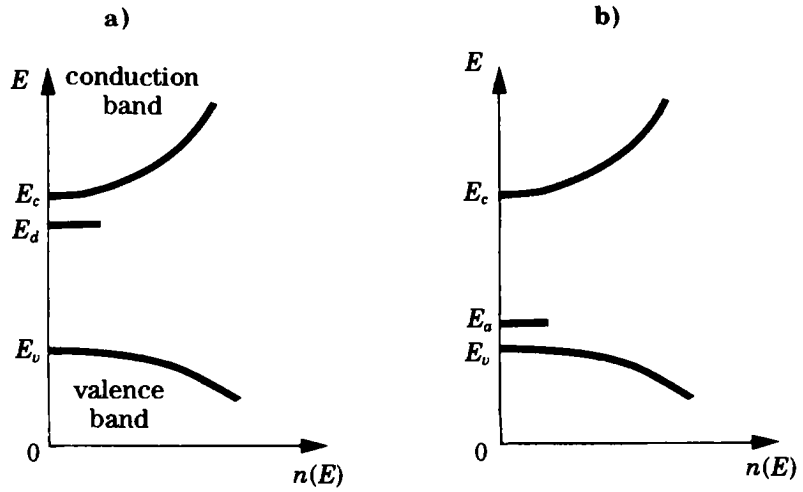


Fig. 3.7. Density of states curve in the vicinity of the band gap in the presence of (a) donors, (b) acceptors.

energy in electron language. This corresponds to the fact that we have to supply energy $E_a - E_v$ to an electron of the valence band to compensate the hole bound to the acceptor, thus ionizing the acceptor and creating a free hole (Fig. 3.7(b)).

The following remark is important for the statistics of semiconductors which we shall study in the next chapter. In the presence of N crystal sites of which N_a are acceptors, the N_a acceptors each capture an electron and create a hole, which at zero temperature remains trapped by the charged nucleus of the acceptor. The crystal is an insulator and its valence band is therefore full. Now, the number of electrons in the valence band is $4N - N_a$. Then in the presence of N_a acceptor sites out of a total of N atomic sites the number of places in the valence band is $4N - N_a$.

The above discussion of donor and acceptor ionization energies does not amount to a proof. In particular we have not shown under what conditions we can pass from the exact Schrödinger equation of the problem

$$\left[\frac{p^2}{2m} + V_p(\mathbf{r}) + V_I(\mathbf{r})\right] \psi = E\,\psi, \tag{3.23}$$

where m is the mass of the free electron, V_p the potential of the crystal, and $V_I(\mathbf{r})$ the potential of the impurity, to the equation

$$\left[\frac{p^2}{2m^*} + V_I(\mathbf{r})\right] \phi = (E - E_c)\phi, \tag{3.24}$$

where m^* is the effective mass, E_c the energy of the bottom of the conduction band, and ϕ a pseudo-wave function. It can be shown that the real wave function ψ can be written in the form $\psi(\mathbf{r}) = \phi(\mathbf{r})u_0(\mathbf{r})$, where ϕ plays the role of an envelope function ($u_0(\mathbf{r})$ is the periodic part of the Bloch function for $\mathbf{k} = 0$ in the case of a standard band). The theory taking us from Eq. (3.23) to Eq. (3.24) is called **effective mass theory**. A recent application of this theory to the study of "quantum wells" and "superlattices" is given in Appendix 3.2.

The fact that the experimental values of the ionization energies of the donors or acceptors vary from one impurity to another shows that some of our assumptions do not rigorously hold:

1 – The donor–Si bond differs from the Si–Si bond.

2 – The size of the atoms is not the same.

3 – The relative dielectric constant is not constant over all space. It varies from one very close to the impurity to ϵ_r over a few atomic layers.

Further, we should take into account the fact that the conduction band of Si or Ge has several minima (see W. Kohn, in Solid State Physics, Volume 5, Academic Press, New York, 1957).

If we consider impurities of column II or VI the simple picture of shallow levels no longer holds for several reasons. The internal levels are very different from those of column IV elements. The nuclear charge is stronger, the radius of the orbit is smaller, and the dielectric constant effect is reduced. This leads to "deep level" states, i.e., levels far from the band edges. These impurities, which exist in all semiconductors, are very important in the recombination of electron-hole pairs but they are relatively difficult to ionize thermally, as their energy distance from a band is much greater than the thermal energy kT. The case of amorphous semiconductors is more subtle, as geometrical defects play the role of chemical impurities. A qualitative description is given in Appendix 3.3.

3.3 Impurity Bands

Up to now we have discussed the effect of substitutional impurities, donors or acceptors, diluted within the semiconductor lattice. They give rise to discrete energy levels within the band gap. We can imagine that if the concentration, e.g., of donors becomes large enough in real space for the orbits to interact, the electrons from the donors will be delocalized within the crystal, even at low temperature. This happens for a donor concentration N_{do} such that $(4\pi/3)a_1^{*3} \cdot N_{do} \simeq 1$. The crystal is then a conductor at any temperature. An insulator–metal transition is thus observed as a function of concentration: for example, for $N_d < N_{do} = 3.7 \times 10^{24}$ m^{-3} silicon is an

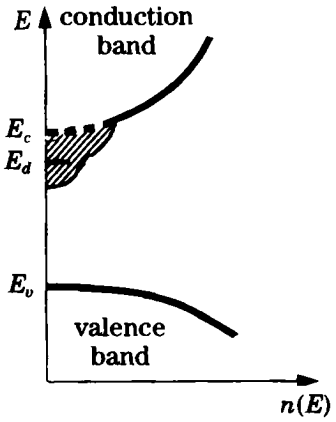


Fig. 3.8. Density of states in the presence of a strong donor concentration. An impurity band forms, shown hatched in the figure.

insulator at low temperature, while above it is a conductor. The presence of many donors leads to a broadening of the level E_d, which is then no longer separated from the conduction band (Fig. 3.8). An "impurity band" has formed.

Appendix 3.1

Problems on Cyclotron Resonance in Silicon

To study cyclotron resonance in a semiconductor one irradiates a single crystal with electromagnetic waves of frequency $\nu = \omega/2\pi$ from a waveguide. The crystal is placed between the poles of an electromagnet (Fig. 3.9). When the magnetic field is such that $\omega = \omega_c = eB/m_e$, where m_e is an effective mass, the crystal conductivity increases rapidly, leading to partial absorption of the waves, which can be observed by the power meter.

The experiment is performed at very low temperatures using a very pure single crystal, and the crystal is illuminated with light of energy greater than the band gap width. With a silicon crystal oriented with the **B** field in the bisecting plane of the cube $(1\bar{1}1)$ at an angle of 30° with the direction [001] (see Fig. 3.10), one observes the signal indicated in Fig. 3.11 as a function of magnetic field.

1. Why do we illuminate the sample to observe a signal?

2. Assume that in silicon there exists a minimum of the conduction band inside the first Brillouin zone at $k_z = k_{z,0}$; $k_x = k_y = 0$. Are there any others? If so, why and how many? $(x, y, z$ are the axes of the cubic cell of diamond.)

Show that the constant energy ellipsoids are ellipsoids of revolution and specify their axis of symmetry.

3. Consider first the ellipsoid of constant energy E centered at $k_{z,0} > 0$ described by the equation

$$E = \frac{\hbar^2}{2}\left[\frac{k_x^2}{m_T} + \frac{k_y^2}{m_T} + \frac{(k_z - k_{z,0})^2}{m_L}\right], \tag{3.25}$$

where m_T and m_L are the transverse and longitudinal masses. We call θ the angle between the magnetic field **B** and the axis of symmetry. We study the electron cyclotron resonance of electrons from this ellipsoid.

Show that the equation of motion of the electrons

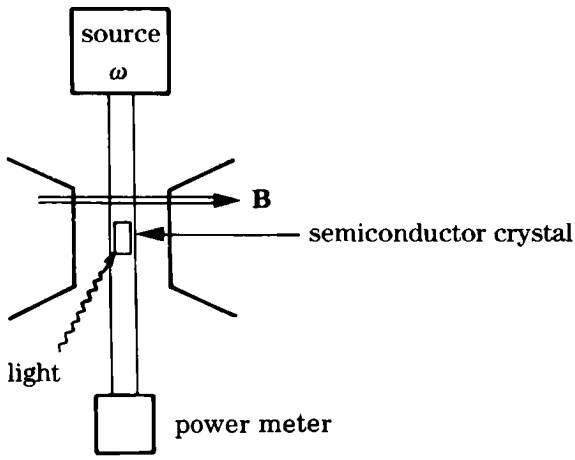


Fig. 3.9. Schematic view of a cyclotron resonance experiment.

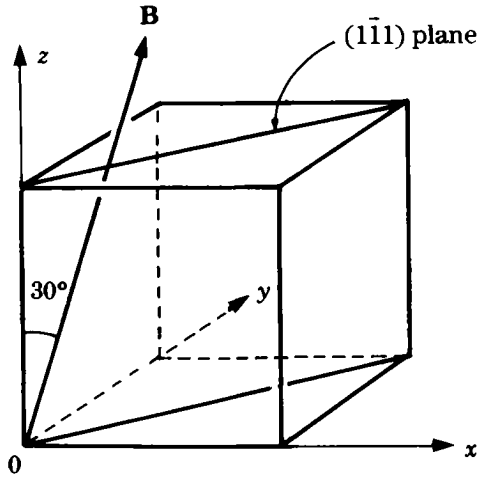


Fig. 3.10. Orientation of the magnetic field **B** the crystal axes of Si.

$$\frac{d\mathbf{v}}{dt} = -\left\|\frac{1}{M}\right\|(e\mathbf{v} \wedge \mathbf{B}) \tag{3.26}$$

has an oscillating solution:

$$\mathbf{v} = \mathbf{v}_0 \exp i\omega t \tag{3.27}$$

with $\omega = \omega_c = eB/m_e$ and

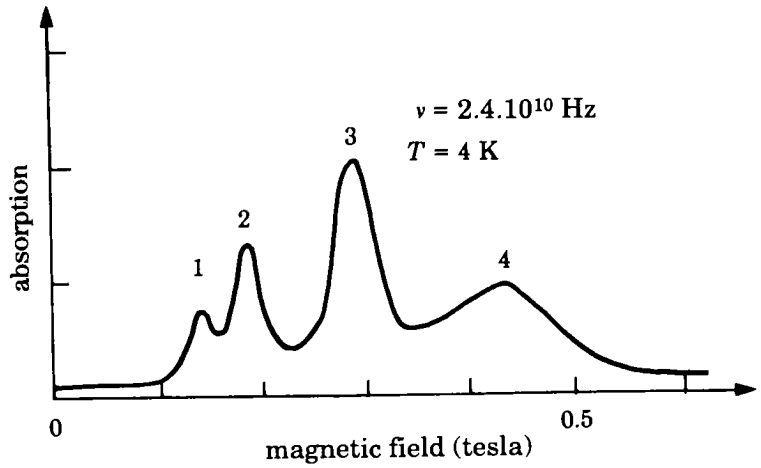


Fig. 3.11. Experimentally measured absorption.

$$m_e = \sqrt{\frac{m_T^2\, m_L}{m_T \sin^2\theta + m_L \cos^2\theta}}. \tag{3.28}$$

(The calculation is simplest if, for this question, the ellipsoid of revolution is referred to axes $k_{x'}$ and $k_{y'}$ such that **B** is in the plane $k_{x'}, 0, k_z$.)

4. Taking account of the conduction band structure found in the second question, how many electron cyclotron resonance lines should one observe for an arbitrary orientation of the magnetic field? Answer the same question for the particular direction described for the experiment.

5. When the magnetic field is rotated from the direction described for the experiment we observe that peaks 1 and 4 do not change while peaks 2 and 3 do move, the first sometimes splitting into two. What is the cause of resonances 2 and 3? What is the cause of resonances 1 and 4 which are independent of the field orientation?

6. Deduce from the experimental figure the effective masses of the electrons and holes in silicon. We recall that the field of the cyclotron resonance for a free electron is about 0.86 T for the frequency $\nu = 2.4 \times 10^{10}$ Hz.

7. What is the effect of collisions on this experiment? Why is it necessary to work at low temperature?

Solutions

1. At low temperature the electrons and holes are trapped and there are no free carriers. To observe the resonance requires free electrons or holes, which are created by photoexcitation.

2. The crystal is cubic and the constant energy surfaces must have the same symmetries as the cube. By symmetry with respect to the center of the zone we pass from $+k_{z,0}$ to $-k_{z,0}$. By rotating through $\pi/2$ we pass to $\pm k_{x,0}$ and $\pm k_{y,0}$, with $|k_{x,0}| = |k_{y,0}| = |k_{z,0}|$. Each ellipsoid must be rotationally symmetric around the axis of the cube on which it is centered (this is a result of the invariance under rotations through $\pi/2$ around this axis). The set of all ellipsoids has indeed the symmetry of the crystal (see Sect. 2.4 and Fig. 2.13).

3. Using the expression for the effective mass tensor deduced from Eq. (3.25) and replacing Eq. (3.27) in Eq. (3.26) we get

$$\begin{bmatrix} i\omega v_{0,x'} \\ i\omega v_{0,y'} \\ i\omega v_{0,z} \end{bmatrix} = \begin{Vmatrix} \frac{1}{m_T} & 0 & 0 \\ 0 & \frac{1}{m_T} & 0 \\ 0 & 0 & \frac{1}{m_L} \end{Vmatrix} \begin{bmatrix} -ev_{o,y'} B\cos\theta \\ ev_{o,x'} B\cos\theta - ev_{0,z} B\sin\theta \\ ev_{o,y'} B\sin\theta \end{bmatrix}$$

We have three homogeneous equations whose determinant vanishes if

$$\omega^2 = e^2\, B^2 \left(\frac{\sin^2\theta}{m_L m_T} + \frac{\cos^2\theta}{m_T^2} \right) \tag{3.29}$$

or

$$\omega = \frac{eB}{m_e} \tag{3.30}$$

with

$$m_e = \sqrt{\frac{m_T^2\, m_L}{m_T \sin^2\theta + m_L \cos^2\theta}}. \tag{3.31}$$

4. As the only relevant angle is between the magnetic field and the principal axis, and there are three principal axes, each common to two ellipsoids, we ought in general to observe three electron peaks. For the particular orientation of the experiment with B in a bisecting plane, the angle of B to the ellipsoids centered on $\pm k_{x0}$ and $\pm k_{y0}$ is the same and we expect two electron peaks.

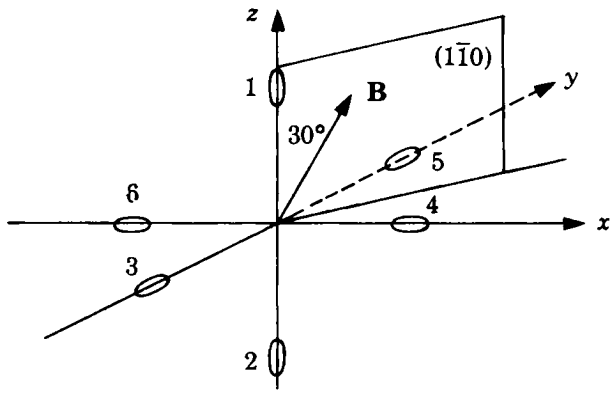


Fig. 3.12. Orientation of **B** relative to the ellipsoids of constant energy of the conduction band.

5. Peaks 1 and 4 are unmodified as they arise from holes, whose two bands are degenerate at $k = 0$ and are spherical. Thus $m_L = m_T$ and the resonance does not move when the magnetic field is rotated. Peaks 2 and 3 which move with the orientation are therefore the electron peaks.

6. For the ellipsoids 1 and 2 centered on Oz: $\theta = 30°$ and

$$m_{1e} = \sqrt{\frac{m_T^2\, m_L}{1/4m_T + 3/4m_L}} \tag{3.32}$$

for the electron resonance of ellipsoids 1 and 2.

The vector **B** has components $B(\sqrt{2}/4, \sqrt{2}/4, \sqrt{3}/2)$. The cosine of the angle between **B** and Ox is $\sqrt{2}/4$; similarly for the four ellipsoids in the plane and

$$m_{3e} = \sqrt{\frac{m_T^2\, m_L}{7/8m_T + 1/8m_L}} \tag{3.33}$$

for the electrons populating the ellipsoids 3, 4, 5, and 6.

We observe peaks 2 and 3 at about 0.18 T and 0.29 T. The strongest peak 3 must correspond to the four ellipsoids in the plane xOy. Comparing with the resonance field for a free electron of mass m at the same frequency we then have

$$\frac{8\, m_T^2\, m_L}{(7\, m_T + m_L)m^2} = \left(\frac{0.29}{0.86}\right)^2 \text{ and } \frac{4\, m_T^2\, m_L}{(m_T + 3\, m_L)m^2} = \left(\frac{0.18}{0.86}\right)^2 \tag{3.34}$$

from which we obtain $m_L/m_T = 4.75$ and $m_T \simeq 0.19\ m; m_L = 0.9\ m$.

Peak 1 appears at 0.13 T. It corresponds to an effective mass of $(0.13/0.86)m$, or about $0.15m$, and thus to light holes. Peak 4 at 0.44 T corresponds to heavy holes of mass $(0.44/0.86)m \simeq 0.5m$.

7. To observe the resonance, collisions must be rare enough that a circular or elliptical orbit can be completed between collisions. The effect of collisions is to broaden the resonance. One works at low temperature so that the time τ between collisions is long enough that $\omega_c \tau > 1$.

Appendix 3.2

Quantum Wells and Semiconducting Superlattic

The effective mass theory mentioned in Sect. 3.2 allows us to sta cussion of the physics of quantum wells and superlattices. These r stacks of alternating crystalline layers, possibly as small as a few monolayers, of semiconductors of differing chemical compositions. I perlattice the stacking is periodic, as shown in Fig. 3.13, which rep a GaAs–$Al_xGa_{1-x}As$ superlattice. We can only obtain such systems taxy, the crystalline growth of one semiconductor on another, is p This requires semiconductors of different chemical composition tl very close geometrically: the same type of crystalline lattice, and th size of elementary cell. The thickness of the layers may be of the o 50 angstroms, but this can be varied.

Such structures have a paradoxical physical property: we would the optical absorption threshold to be that of the material with the na band gap, or in other words that a quantum well would be less trans than any of its constituents. Experiment shows that this is not th and we shall see below that effective mass theory allows us to cle paradox.

As the two material constituents are different they do not have tl value of band gap. For example in GaAs, E_g is 1.42 eV while in $Al_{0.4}C$ the gap E_g is 2 eV. These objects are made by a technique called mo beam epitaxy (MBE), which gave its first promising results in the 19 this process the crystal grows in an ultra-high vacuum chamber in are placed crucibles containing gallium, aluminium, and arsenic (Fig The temperature of each crucible is controlled independently, so tl can control the speed of evaporation of each atomic species and thus of atoms of each species. In the ultra-high vacuum chamber the atom a GaAs substrate and one observes crystallization occurring atomic l atomic layer. It is possible to control the deposition to rates for exa one atomic layer per second. One can thus create tailored semicon structures on demand, in what is called "band gap engineering."

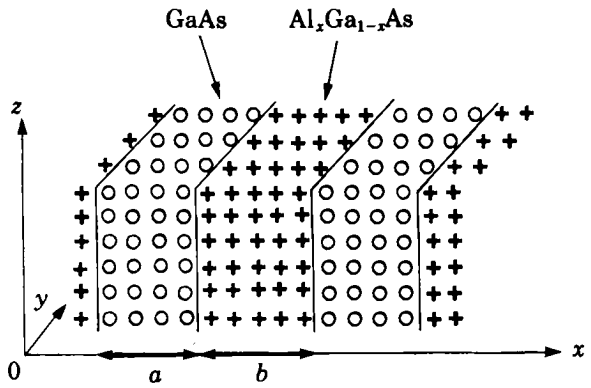


Fig. 3.13. Structure of a superlattice of period $a + b$ of $GaAs/Al_xGa_{1-x}As$.

Why do we have to work in ultra-high vacuum? We can determine from the kinetic theory of gases the rate at which a surface becomes polluted by an atomic monolayer by assuming that every atom which hits the surface sticks to it. For a pressure of 10^{-13} bar the pollution time of the order of an hour. To avoid reaching such low pressures, we could think of increasing the flux of the atoms making up the semiconductor layers, but we have to allow time for the atoms to reach their equilibrium positions on the surface.

The crystalline potentials acting on the electrons are different in the two materials, so we expect that the band structure itself, and particularly the

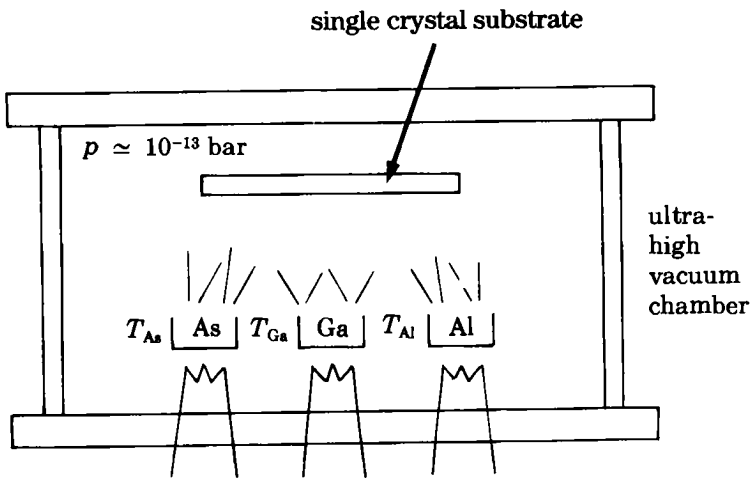


Fig. 3.14. Principle of molecular beam epitaxy.

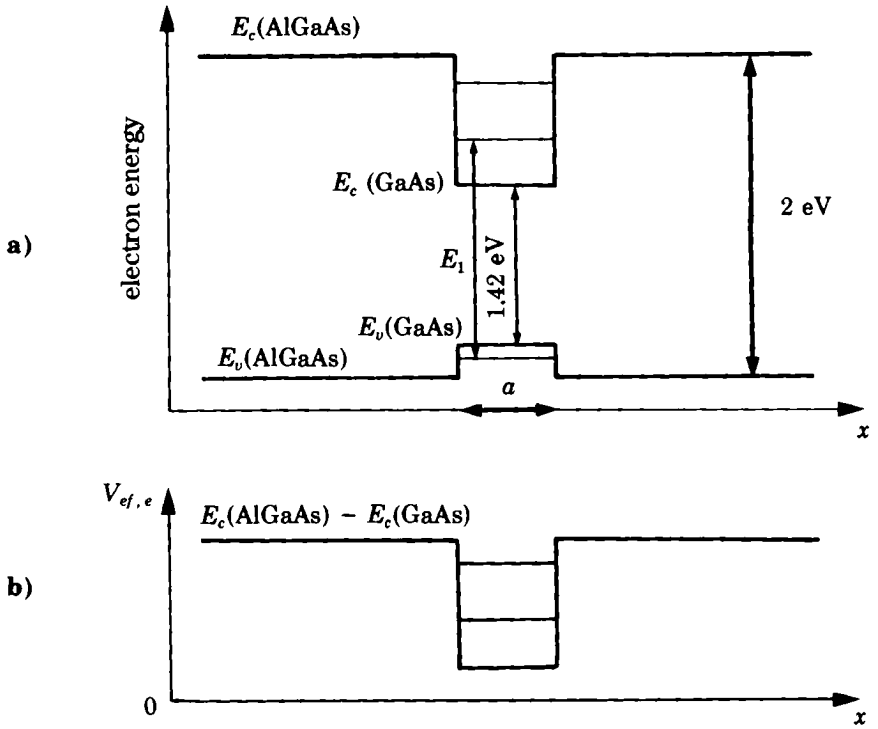


Fig. 3.15. Individual quantum well of GaAs in $Al_xGa_{1-x}As$. We show (a) the electron energy, and (b) the effective potential seen by the electrons, as functions of the distance perpendicular to the layers.

bottom of the conduction band and the top of the valence band would be at different energies. This is shown in Fig. 3.15(a) for the case of an isolated quantum well. Effective mass theory allows us to calculate the energy levels in this structure. We replace the real Schrödinger equation by an effective equation involving the effective Hamiltonian:

$$\left[\frac{p^2}{2\,m_e} + V_{ef,e}\,(x)\right]\phi_e(\mathbf{r}) = (E_e - E_c)\,\phi_e(\mathbf{r}). \tag{3.35}$$

Here m_e is the electron effective mass and $V_{ef,e}$ the effective potential seen by the electrons, which takes account of the different nature of the two materials. This potential is represented in Fig. 3.15(b). The functions ϕ are the "envelope" wave functions, which take the form

$$\phi_e(\mathbf{r}) = \phi_e(x)\exp(ik_y y)\ \cdot \exp(ik_z z), \tag{3.36}$$

where $\phi_e(x)$ is the solution of a one-dimensional Schrödinger equation in the potential of Fig. 3.15(b). We are led back to a problem in elementary quantum mechanics, that of a rectangular potential well in which one finds both

delocalized levels extending above the barrier and localized levels E_e^α, E_e^β with $\alpha, \beta = 1, 2, \ldots$ inside the well.

In a structure with a periodic stacking of layers, like that of Fig. 3.13, two situations can occur. If the material with the large gap is much thicker than the material with the small gap ($b >> a$), then the eigenstates are bound states in each individual quantum well (1), (2), ..., of energy $E_e^\alpha, E_e^\beta, \ldots$. (See also Appendix 6.3.) On the other hand if b is small enough we have to take account of the overlap between functions localized in neighboring wells and thus make a theory of "superbands" in the "superperiodic" potential of Fig. 3.16. This is what we call a "superlattice." The levels E_e^α, E_e^β widen into bands $E_e^{\alpha'}, E_e^{\beta'}, \ldots$ called "minibands."

Similarly we can find the hole states by using the hole effective potential and the effective mass of the hole in an equation analogous to Eq. (3.35):

$$\left[\frac{p^2}{2\,m_h} + V_{ef,h}(x)\right] \phi_h(\mathbf{r}) = (E_h - E_v)\phi_h(\mathbf{r}). \tag{3.37}$$

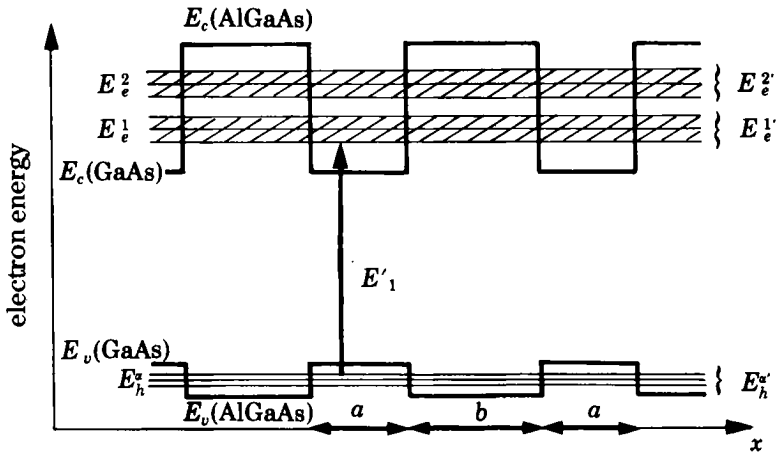


Fig. 3.16. Band profile of a superlattice $GaAs/Al_xGa_{1-x}As$, of period $a + b$.

We obtain in this way hole states $E_h^\alpha \ldots$ or bands $E_h^{\alpha'} \ldots$. Placing all these electron energy levels in Fig. 3.15(a) or 3.16, we see that the new absorption threshold is E_1 or E_1' above the band gap of GaAs. For energies less than E_1 the structure is transparent and our paradox is resolved. The physical effect at the basis of this phenomenon is the increase of energy of a quantum state when it is confined, a general consequence of the Heisenberg principle.

Note that the price of an MBE machine is about one million dollars.

Appendix 3.3

Amorphous Semiconductors

An amorphous solid is one without the property of periodicity in space. Most of the properties of solids mentioned up to now were based on invariance under translations, which gives rise to Bloch's theorem. Should we conclude that a solid of given composition would change all its properties depending on the existence or not of long-range order? To answer this question it is useful to return to the chemical approach (cf. Sect. 1.2). In an amorphous as well as in a crystalline solid, there will be a "chemical bond" formed by the hybridization of the orbitals. This bond is a short-range property, while periodicity is a long-range one.

Let us take the example of amorphous silicon. The nearest neighbors of a silicon atom will still be at the vertices of a tetrahedron, but there will be distortions of the bond angles at the second, third neighbors and so on. This is shown in Fig. 3.17. We will arrive at a distribution of the atoms where some of the chemical bonds cannot be established for geometrical reasons. We say that we have "dangling bonds."

The energy levels corresponding to the dangling bonds are in the band gap of the crystal, as it is precisely the hybridization which creates the band gap. This is shown schematically in Figs. 3.18(a) and 18(b), where we compare the density of states in two semiconductors of the same chemical composition, one crystalline and the other amorphous.

An amorphous semiconductor, even though chemically pure, is thus "electronically" impure, since it has many states in the band gap. However, if we can replace the broken bonds with a strong chemical bond, each dangling bond we suppress will remove a state from the gap, the corresponding energy state now lying inside the valence band. This occurs for example when amorphous silicon is exposed to hydrogen at high temperatures. The hydrogen reacts and gives Si–H bonds, the corresponding states disappearing from the band gap (Fig. 3.19). Paradoxically, hydrogenated amorphous silicon is then more "electronically" pure than pure amorphous silicon. The density of states approaches that of crystalline silicon.

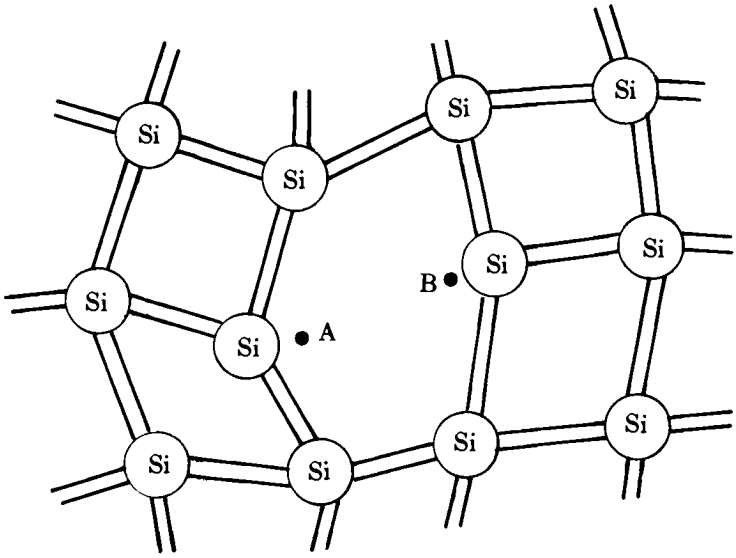


Fig. 3.17. Schematic representation of a disordered lattice of amorphous silicon. Most of the atoms establish four bonds, but because of the deformations situations exist in which atoms A and B are too far apart to establish a bond.

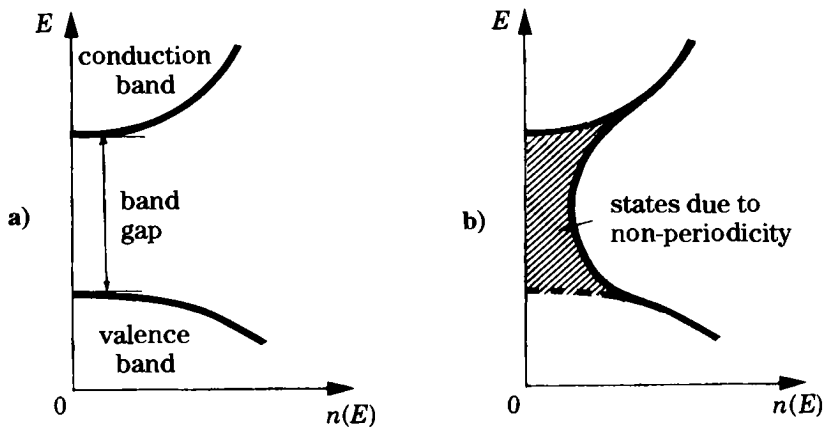


Fig. 3.18. Density of states in semiconductors (a) crystalline, and (b) amorphous, of the same chemical composition.

In practice hydrogenated amorphous silicon is made directly by decomposing silane (SiH_4) in a radiofrequency plasma. This is a very cheap semiconductor which is used particularly in the manufacture of pocket calculator photocells.

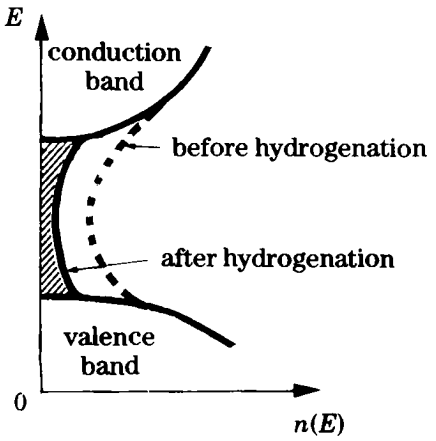


Fig. 3.19. Reduction of the density of states in the band gap by hydrogenation of amorphous silicon. It is reduced from about 10^{20} $eV^{-1} \cdot cm^{-3}$ to about 10^{15} $eV^{-1} \cdot cm^{-3}$.

The present price for amorphous silicon cells is between one-tenth and one-hundredth of the current crystalline panels.

4.

Statistics of Homogeneous Semiconductors

4.1 Occupation of the Electron Levels

Knowing the band structure, at least in the region close to the band gap, as well as the localized quantum states caused by the presence of shallow impurities, calculation of the electrical conductivity of the semiconductor now requires to find the number of mobile charges and their nature, electrons or holes, at thermal equilibrium. For this we must calculate at temperature T the occupation probabilities of the accessible energy levels.

The electrons have half-integer spin and are fermions, so the system state only contains one electron per single-particle quantum state. For a given value of $\mathbf{k}$ there are two quantum states with different spins which can be simultaneously occupied. We recall that in such non-interacting Fermi–Dirac gases we define the Fermi factor f as the mean value of the operator measuring the occupation number of an electron state of energy E. If the chemical potential is E_F, hereafter called the Fermi level, we can show (see Appendix 4.1) that

$$f = \frac{1}{1 + \exp(E - E_F)/kT}, \tag{4.1}$$

where k is the Boltzmann constant.

We note that if $(E - E_F)$ is large compared with kT, i.e., if E exceeds E_F by several times the thermal energy, the Fermi factor of the electron can be approximated by a Maxwell–Boltzmann expression:

$$f \simeq \exp\left(\frac{E_F}{kT}\right) \exp\left(-\frac{E}{kT}\right). \tag{4.2}$$

This means that if f is small compared with 1 the effect of exclusion is negligible: then from a statistical point of view the electron behaves like a classical particle.

In semiconductors if we consider the mean occupation rate of donor levels which are **localized** states, Eq. (4.1) does not apply. In fact there are

two degenerate spin states corresponding to each donor, but the two states cannot be occupied simultaneously because of the Coulomb repulsion between localized electrons (the corresponding state has a very high energy). In this case we can show (Appendix 4.1) that the Fermi function must be replaced, if we limit ourselves to the ground state of the donor, by

$$f' = \frac{1}{1 + (1/2)\exp(E - E_F/kT)}. \tag{4.3}$$

4.2 Hole Occupation

If the occupation probability of a level is given by Eq. (4.1), the probability that this level will not be occupied by an electron, i.e., the probability of occupation by a hole, is

$$f_h(E) = 1 - f(E) = \frac{1}{\exp(E_F - E/kT) + 1}. \tag{4.4}$$

If we measure the energy of a hole in the opposite sense from the electron energy (cf. Sect. 3.1) we have

$$f_h(E_h) = \frac{1}{1 + \exp(E_h - E_{Fh}/kT)} \tag{4.5}$$

which behaves similarly to the Fermi function Eq. (4.1).

As for an electron, these expressions can be approximated once the Fermi level is at few kT from the hole energy ($E_h - E_{Fh} >> kT$):

$$f_h \simeq \exp\left(-\frac{E_h - E_{Fh}}{kT}\right) = \exp\left(\frac{E_{Fh}}{kT}\right)\exp\left(-\frac{E_h}{kT}\right). \tag{4.6}$$

We recover the exponential law of classical particles: the exclusion effect is negligible, as the average occupancy is much smaller than 1.

4.3 Determination of the Chemical Potential

Because a semiconductor is by definition an insulator at zero temperature we know that **the Fermi level must be within the band gap at T = 0**. However if we consider a real, and therefore impure, crystal, its properties, including the electron chemical potential, will be determined by its nature and impurity content. We will regard the electron gas occupying various energy states (valence band, impurity levels, conduction band) as a statistical canonical ensemble and determine its chemical potential. A given crystal will in general simultaneously contain "donor" and "acceptor" impurities.

Consider a crystal of silicon. Let N be the number of sites of the crystal, N_d the number of substitutional sites occupied by donors, e.g., phosphorus P, and N_a the number of substitutional sites occupied by acceptors, e.g., boron B. Leaving aside the electrons in deep atomic levels, the total number N_T of electrons present in the crystal is

$$\begin{aligned} N_T &= 4N - 4N_d + 5N_d - 4N_a + 3N_a \\ &= 4N + N_d - N_a. \end{aligned} \tag{4.7}$$

At finite temperature the electrons can be in states of four types: in the valence band forming Si–Si bonds, in acceptor levels to form the four bonds of B–Si type (the acceptor is assumed ionized), in donor levels which are then neutral, or in the conduction band (cf. Sect. 3.2). We call p the concentration of holes, n_a^0 the concentration of neutral acceptors, n_a^- the concentration of ionized acceptors, n_d^0 the concentration of neutral donors, n_d^+ the concentration of ionized donors, and n the concentration of electrons.

Let us think in "electron language." We know (see Sect. 3.2) that when the valence band is full it contains $4N - N_a$ electrons. Writing N_T as the sum of the number of electrons in the valence band $4N - N_a - p$, the number of ionized acceptors n_a^-, the number of neutral donors n_d^0, and the n conduction electrons, we have

$$N_T = 4N - N_a - p + n_a^- + n_d^0 + n. \tag{4.8}$$

On the other hand,

$$\begin{aligned} N_a &= n_a^0 + n_a^-, \\ N_d &= n_d^0 + n_d^+. \end{aligned} \tag{4.9}$$

From Eqs. (4.7), (4.8), and (4.9) we deduce

$$p + n_d^+ = n + n_a^-. \tag{4.10}$$

Writing Eq. (4.8) amounts to writing

$$\int f(E)n(E)\, dE = N_T, \tag{4.11}$$

where $n(E)$ is the density of states of energy E. The states are either delocalized in the valence and conduction bands, or localized for the discrete acceptor and donor states.

This equation expresses the conservation of electron number in the system at a given temperature (canonical ensemble) and therefore must give the chemical potential or Fermi level through the use of Eq. (4.10). We note that since the crystal remains neutral when we substitute donors or acceptors, the appearance of positive charges in the form of free holes and ionized donors must be compensated by the appearance of negative charges: conduction electrons and ionized acceptors.

The conservation of electron number in a homogeneous semiconductor is thus equivalent to electrical neutrality. We shall see later that in inhomogeneous semiconductors local electrical neutrality does not hold.

4.4 Statistics of Pure or Intrinsic Semiconductors

We describe as intrinsic a (hypothetical) semiconductor without any impurity or defect. We shall see that some of the properties demonstrated below hold for real semiconductors in certain temperature ranges.

We call $n_v(E)$ and $n_c(E)$ the densities of states for valence and conduction electrons (Fig. 4.1). We have seen in Sect. 2.3, Eq. (2.52), that near the band edges, at E_c and E_v, the curve for the density of states is parabolic.

The total number n of electrons in the conduction band is

$$n = \int_{CB} n_c(E)\, f(E)\, dE. \tag{4.12}$$

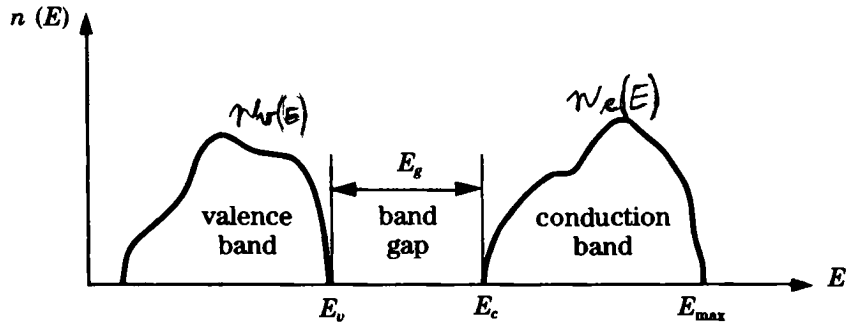


Fig. 4.1. Density of states of the bands near the gap.

If the constant energy surfaces are spheres (that is, if the effective mass is isotropic) and if there is only one minimum, at $k = 0$, the value of $n_c(E)$ for the two spin orientations and unit volume is given by Eq. (2.52) as

$$n_c(E) = 4\pi(2m_e)^{3/2}\frac{1}{h^3}(E - E_c)^{1/2}. \tag{4.13}$$

Let us assume that the Fermi level is several times kT below E_c. Then we can use Eqs. (4.2) and (4.12):

$$n = \exp\frac{E_F - E_c}{kT}\int_{E_c}^{\infty} dE\; n_c(E)\exp[-(E - E_c)/kT]. \tag{4.14}$$

Because of the exponential factor in the integral we were able to extend the upper limit to infinity and use Eq. (4.13) which holds in principle only at the bottom of the band. Setting $x = (E - E_c)/kT$ and using the result

$$\int_0^\infty x^{1/2} e^{-x} dx = \frac{\sqrt{\pi}}{2}$$

we obtain

$$n = N_c \exp \frac{E_F - E_c}{kT}, \tag{4.15}$$

where

$$N_c = 2\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2}. \tag{4.16}$$

We see that N_c plays the role of the degeneracy in a single level of energy E_c. We call it the effective or equivalent density of states of the conduction band. We note however that N_c is a function of temperature.

We can make an exactly similar calculation for the holes in the valence band. The number of empty electron places is given by Eq. (4.4), which tends to $\exp(E - E_F)/kT$ for $(E_F - E) >> kT$ (electron population close to 1, hole population low). We find the hole number p by an analogous equation:

$$p = N_v \exp \frac{E_v - E_F}{kT}, \tag{4.17}$$

where N_v is the equivalent density of states of the valence band

$$N_v = 2\left(\frac{2\pi m_h kT}{h^2}\right)^{3/2} \tag{4.18}$$

with m_h the hole mass.

In the case where, as in silicon or germanium, the bottom of the conduction band comprises several ellipsoids centered at various points of the Brillouin zone, we must take account on the one hand of the number of minima, by multiplying N_c by this number, and on the other hand of the anisotropy of the effective mass. The constant energy surfaces around these minima are ellipsoids with equations:

$$E = \frac{\hbar^2}{2} \cdot \left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z}\right), \tag{4.19}$$

where the origin of the wave vector is taken at the energy minimum considered. The volume of an ellipsoid of semiaxes a, b, c is $(4/3)\pi abc$. The volume inside a surface of energy E is thus

$$\frac{4}{3}\pi\left(\frac{2m_xE}{\hbar^2}\right)^{1/2}\left(\frac{2m_yE}{\hbar^2}\right)^{1/2}\left(\frac{2m_zE}{\hbar^2}\right)^{1/2} \qquad (4.20)$$

and the density of states per unit volume for the two spin orientations per ellipsoid is

$$n(E) = \pi(2)^{3/2}\frac{(m_xm_ym_z)^{1/2}}{h^3}E^{1/2}. \qquad (4.21)$$

For silicon we have to replace the factor $m_e^{3/2}$ which appears in expression (4.16) by $(m_Lm_T^2)^{1/2} \times 6$ to take account of the six ellipsoids. We get finally

$$N_c = 2.8\ 10^{25}\ \mathrm{m}^{-3} \text{ at } 300\ \mathrm{K}.$$

If the valence band is degenerate at $k = 0$ as is the case in Si, Ge, GaAs, we have in first approximation two spherical bands each with an isotropic mass $m_{h,h}$ and $m_{h,l}$. We simply have to sum over the two corresponding densities of states. The factor $m_h^{3/2}$ in Eq. (4.18) must be replaced by $m_{h,h}^{3/2} + m_{h,l}^{3/2}$. With this change N_v is $10^{25}\ \mathrm{m}^{-3}$ in silicon at 300 K.

The relations (4.15) and (4.17) hold **even if the crystal is not pure** on the condition that the Fermi level is several times kT from the band edges. Taking the product we thus always have

$$np = N_cN_v\exp\left(-\frac{E_g}{kT}\right) = n_i^2. \qquad (4.22)$$

The quantity n_i^2 plays the role of the constant in the law of mass action for the reaction

$$\text{electron} + \text{hole} \rightleftarrows \emptyset + \text{energy}.$$

If the crystal is pure (intrinsic) Eq. (4.10) is

$$n = p = n_i = p_i \qquad (4.23)$$

or, taking account of Eqs. (4.15) and (4.17),

$$N_c\exp\frac{E_F - E_c}{kT} = N_v\exp\frac{E_v - E_F}{kT}.$$

We thus find the position of the Fermi level

$$E_F - E_c = -\frac{1}{2}E_g + \frac{kT}{2}\operatorname{Ln}\frac{N_v}{N_c}. \qquad (4.24)$$

For "standard" bands (Sect. 2.5)

$$E_F - E_c = -\frac{1}{2}E_g + \frac{3}{4}kT\operatorname{Ln}\frac{m_h}{m_e}. \qquad (4.25)$$

We see that for an intrinsic semiconductor the Fermi level lies close to the middle of the band gap whatever the temperature. For a band structure exactly symmetrical with respect to the band gap the Fermi level is independent of the temperature. Substituting Eq. (4.25) into Eq. (4.15) or Eq. (4.17) we obtain the intrinsic concentrations

$$n_i = p_i = 2\left(\frac{2\pi}{h^2}\right)^{3/2} m_e^{3/4} m_h^{3/4} (kT)^{3/2} \exp\left(-\frac{E_g}{2kT}\right). \tag{4.26}$$

The case of multiple minima, or of degeneracy at $k = 0$, have to be taken into account in the expression for N_c and N_v but in all cases the relation (4.22) will hold.

As we have seen in Sect. 3.1b, the electrical conductivity of a crystal is the sum of the conductivities due to the electrons and the holes, σ_e and σ_h:

$$\sigma = \sigma_e + \sigma_h \tag{4.27}$$

or

$$\sigma = n\mu_e e + p\mu_h e. \tag{4.28}$$

In this expression μ_e is the electron mobility and μ_h that of the holes.

We see from this formula that if the mobilities do not vary too rapidly with temperature (Sect. 5.4) the variation of σ will be very rapid, as $\exp(-E_g/2kT)$. This explains the very strong increase in conductivity with temperature and provides a method of measuring E_g. In contrast to metals, where the number of carriers is constant, the conductivity of semiconductors increases with temperature mainly through the increased number of carriers. Figure 4.2 shows the variation of intrinsic concentration with temperature in several semiconductors. We note the small concentration at room temperature: for silicon

$$n_i \simeq 1.6 \times 10^{16} \text{ m}^{-3} \text{ at } 300 \text{ K}.$$

The concentration of silicon atoms in silicon is of the order of 10^{28} m^{-3}. At room temperature the intrinsic band-to-band ionization is thus very low, of the order of 10^{-12} in relative concentration. We then understand how shallow impurity levels may dominate the properties of silicon at room temperature, even at very great purity, as their internal ionization energy is very small.

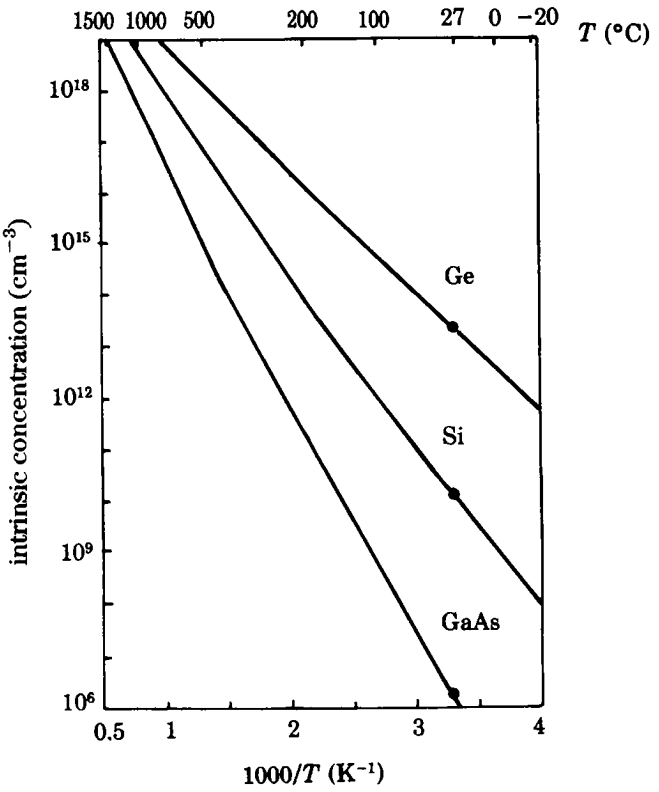


Fig. 4.2. Concentration of intrinsic charge carriers as a function of temperature in Si, Ge, GaAs. In this Arrhenius plot, the slopes of the straight lines are proportional to $E_g/2$, as long as the influence of the $T^{3/2}$ term is not significant.

4.5 Statistics of a Semiconductor Containing Impurities: The Notion of Majority and Minority Carriers

The relations (4.15), (4.17), and (4.22) actually only use the property of the Fermi level of being several times kT distant from the band edges, so that we can neglect the 1 in the denominator of the Fermi function. In this case we say that the semiconductor is **non-degenerate**, a **degenerate** semiconductor corresponding to the case where the Fermi level falls close to or even within a band. The latter situation occurs in heavily doped semiconductors, for which the donor (or acceptor) level is broadened up to merging into the conduction (or valence) band. We shall consider here

only **impure** but **non-degenerate** semiconductors where the relations (4.15), (4.17), and (4.22) remain valid.

We note that the word degenerate recurs several times in semiconductor physics in referring to different things. This is the normal usage, but it is important that this restricted vocabulary should not cause confusion. For this reason we recall below the various meanings of the word that may be encountered:

1 – In any crystal the states $\mathbf{k}$ and $-\mathbf{k}$ are degenerate; this is a consequence of the fact that the Hamiltonian is real.

2 – If there are symmetries of the crystal potential (rotation, symmetry with respect to a point...) the surfaces of constant energy reflect this symmetry, and a state deduced from $\psi_{n,\mathbf{k}}(\mathbf{r})$ by one of the symmetry transformations is degenerate with $\psi_{n,\mathbf{k}}(\mathbf{r})$. There is thus a degeneracy associated with the crystal symmetry: for example the minima of the six conduction bands of silicon. At the center of the zone this shows itself through the possible existence of degenerate levels.

3 – Finally there is the meaning "degenerate electron gas" in a solid when the population of the states is not small compared with 1.

Sentences 1 and 2 mean that the states have the same energy; sentence 3 means that Fermi–Dirac statistics cannot be approximated by Maxwell–Boltzmann statistics.

n-type Semiconductor

Consider a semiconductor containing a concentration N_d of donors whose ionization energy is E_d but which is completely free of acceptors ($N_a = 0$). At very low temperature the electrons are in the lowest energy states and thus bound to the donor centers and in the valence band.

At higher temperatures the donors will progressively ionize. Let us assume that the Fermi level is below E_d so that all the donors are ionized. Then the number of conduction electrons is

$$n = N_d \tag{4.29}$$

and the relation (4.15) determines the Fermi level:

$$E_F - E_c = kT \log \frac{N_d}{N_c}. \tag{4.30}$$

We see that if N_d is of order 10^{20} m^{-3} the quantity $(E_F - E_c)$ is of order 25 meV $\cdot$ (Ln 10^{-5}) ~ -0.25 eV at room temperature. The Fermi level is well below E_d for a wide range of temperatures. When the temperature is increased the Fermi level moves away from the conduction band and the assumption of complete ionization of the donors remains valid.

However beyond a certain temperature, which depends only on the donor concentration, intrinsic ionization is no longer negligible. Then we have to write Eq. (4.10) in the form

$$n = n_d^+ + p \quad \text{with} \quad n_d^+ = N_d. \tag{4.31}$$

This relation expresses the fact that the electrons populating the conduction band originate from the donors and the valence band. Using Eqs. (4.22) and (4.31) we obtain

$$n = \frac{1}{2}N_d + \left(\frac{N_d^2}{4} + n_i^2\right)^{1/2}. \tag{4.32}$$

In the preceding limit of intermediate temperature ($n_i << N_d$),

$$n \approx N_d + \frac{n_i^2}{N_d} \approx N_d, \tag{4.33}$$

$$p \approx \frac{n_i^2}{N_d}. \tag{4.34}$$

The temperature range for which the electron number remains equal to N_d, independent of the temperature, is called the **saturation regime**. The number of holes p is $n_i \cdot (n_i/N_d) << n_i$. The ratio of the number of electrons to the number of holes is

$$\frac{n}{p} = \frac{N_d^2}{n_i^2}. \tag{4.35}$$

This ratio is very large: for $N_d = 10^{23}$ m^{-3}, $n/p \sim 10^{14}$. **For this reason the electrons are called majority carriers and the holes minority carriers. Then a silicon crystal containing donors is called n-type silicon, as the electric current is carried by electrons, which greatly outweigh the holes. It is also called n-doped.** Doping of a crystal may be the result of precise manufacturing techniques or occur by accident.

We obtain the remarkable result that **the conductivity of such a crystal only depends on its concentration of impurities in this temperature range.** We say that **the conductivity is n-type extrinsic.**

At **high temperature** we can have $n_i >> N_d$; then Eqs. (4.32) and (4.22) can be approximated as

$$\begin{aligned} n &\approx n_i + \frac{1}{2}N_d, \\ p &\approx n_i - \frac{1}{2}N_d, \end{aligned} \tag{4.36}$$

so that we recover the intrinsic regime.

p-doped Semiconductor

Let us consider a crystal where instead of "donor" impurities we have only a concentration N_a of "acceptors," for example boron in silicon. We can discuss the situation either in terms of the electrons in the electron energy diagram or in terms of the holes in the hole energy diagram. In the electron energy diagram let us assume that at room temperature the Fermi level is above the acceptor level, and thus a few times kT from the valence band. The acceptor levels will thus be populated by electrons from the valence band which leave holes. We then have

$$p = N_a \tag{4.37}$$

and from Eq. (4.17),

$$E_F - E_v = kT \operatorname{Ln} \frac{N_v}{N_a}. \tag{4.38}$$

At higher temperature we have to take account of intrinsic ionization by writing Eq. (4.10) in the form

$$n + N_a = p, \tag{4.39}$$

which takes into account the fact that free holes can be created in the valence band, either by trapping of an electron in an acceptor level or by excitation of a free electron into the conduction band. Taking account of Eq. (4.22) we then have

$$p = \frac{1}{2} N_a + \left(\frac{N_a^2}{4} + n_i^2 \right)^{1/2} \tag{4.40}$$

and if $N_a >> n_i$,

$$p \approx N_a + \frac{n_i^2}{N_a} \sim N_a. \tag{4.41}$$

In this case $n \sim (n_i^2/N_a) \sim p \cdot (n_i^2/N_a^2)$. The holes have concentration N_a and the electrons have the very low concentration $n_i \cdot (n_i/N_a)$. In this case the electrons are **minority carriers**. The holes are the **majority carriers**. In this p-doped crystal electric current is carried by the **majority hole carriers**. We say that the conductivity is **p-type extrinsic**.

This explains why certain semiconducting crystals behave under the Hall effect for example as if the current were transported by "positive" electrons. There is a deep symmetry between the properties of n-doped and p-doped crystals.

We could have reasoned entirely in terms of the statistics of holes. Then Fig. 3.7(b) becomes Fig. 4.3 (inverting energies): the holes are trapped by the acceptors at very low temperatures, but if E_{Fh} is situated as in Fig. 4.3 the acceptors are ionized and

$$p = N_a = N_v \exp \frac{E_{Fh} - E_h}{kT}. \tag{4.42}$$

For p-type semiconductors we also have a saturation regime, with the number of charge carriers equal to N_a and independent of temperature.

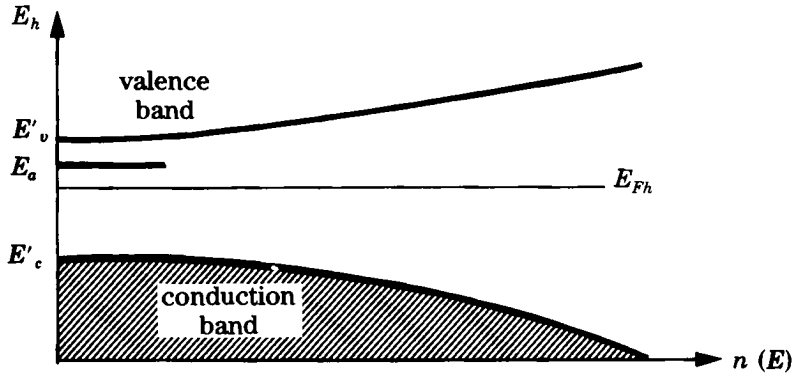


Fig. 4.3. Density of states $n(E)$ for the different hole energies E_h and Fermi level position E_{Fh} near room temperature.

At **high temperature** where $n_i > N_a$ we recover an intrinsic regime in which

$$p = n_i + \frac{1}{2}N_a,$$
$$n = n_i - \frac{1}{2}N_a. \tag{4.43}$$

Figure 4.4 summarizes the results for a semiconductor at ambient temperature, depending on the type of doping.

4.6 Compensated Semiconductor at Intermediate Temperature

Of course in any real crystal, even the purest one can make, there are various kinds of impurities. We must now consider the statistics of electrons in such a crystal.

Let us suppose we are at an intermediate temperature in the saturation regime of an n-type semiconductor containing N_d donors per unit volume. The Fermi level is several kT below the conduction band. Let us introduce some acceptors into the crystal at concentration N_a. For each acceptor there is an energy level situated close to the valence band which will be populated

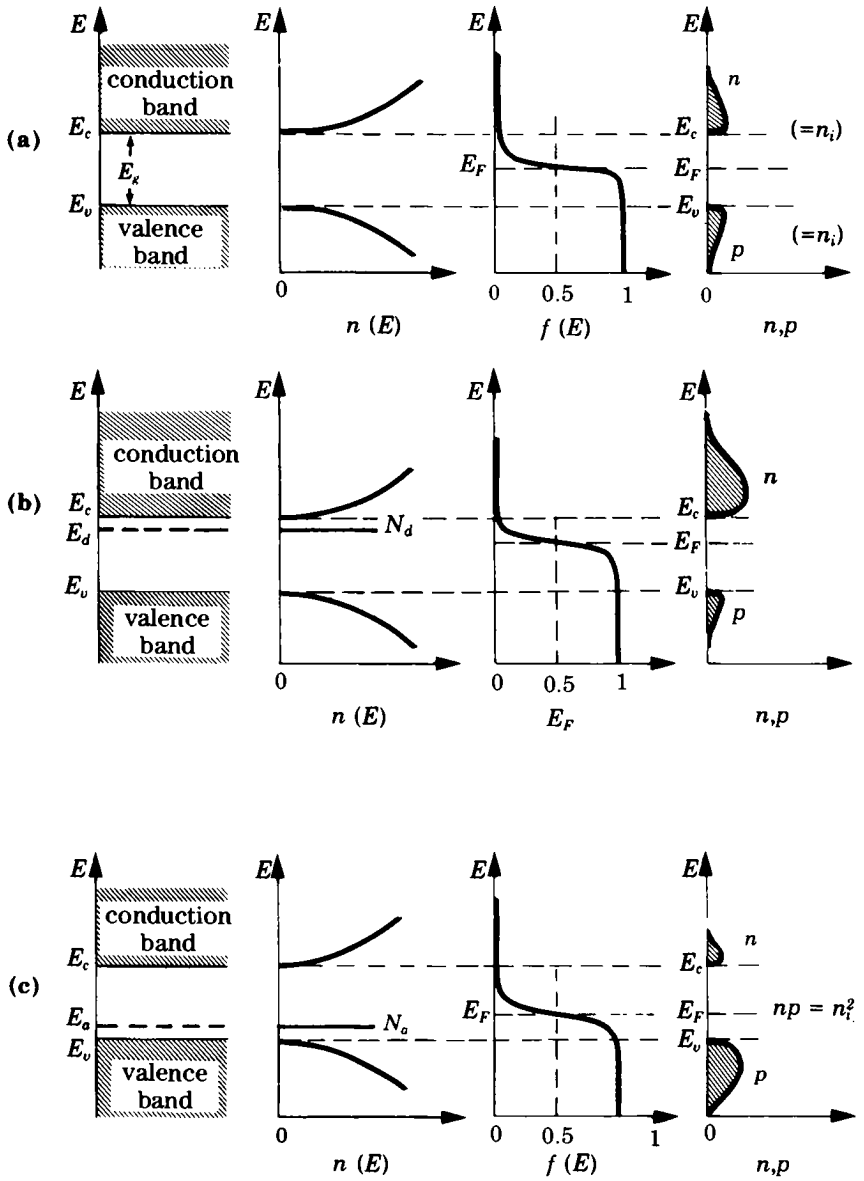


Fig. 4.4. From left to right, the band diagram, the density of states, the Fermi function, and the carrier concentration as functions of energy. (a) Intrinsic case; (b) n-type semiconductor; (c) p-type semiconductor. We note that $np = n_i^2$ in all three cases. (After Sze, Semiconductor devices, J. Wiley, 1985.)

by an electron, since this level is well below the Fermi level. The acceptor has "trapped" an electron. All the acceptors ($<< N_d$) will thus be ionized and the number of free electrons will be decreased accordingly. Thus

$$n = N_d - N_a \tag{4.44}$$

and from Eq. (4.13),

$$E_c - E_F = kT \operatorname{Ln} \frac{N_c}{N_d - N_a}. \tag{4.45}$$

We say that the material is **compensated**. Here we have a partially compensated n-type material. The majority carriers are still the electrons; the number of holes is still very small as the Fermi level is high in the band gap. We note however that as the compensation is increased by raising N_a, the Fermi level decreases. If $N_a = N_d$ the Fermi level will take up the intrinsic position given by Eq. (4.24).

If the number of acceptors becomes larger than the number of donors all the donors will be ionized by trapping an electron in an acceptor level. We will then have $N_a - N_d$ effective acceptors and the concentration of free holes will be

$$p = N_a - N_d, \tag{4.46}$$

where the Fermi level is given by Eq. (4.17) as

$$E_F - E_v = kT \operatorname{Ln} \frac{N_v}{N_a - N_d}. \tag{4.47}$$

E_F will be quite low in the band gap, and the electron number, given by Eq. (4.22) with $p = N_a - N_d$, will remain very small. We will then have a compensated p-type semiconductor.

These ideas allow us to understand the role of purity in the properties of semiconductors at intermediate temperatures. The band structure of the material determines three characteristic concentrations N_c, N_v, and $n_i = (N_c N_v)^{1/2} \exp(-E_g/kT)$.

If a material contains N_d shallow donors, they will ionize and release $n = N_d$ electrons into the conduction band. Let us consider the simultaneous presence of deep donors. If their concentration is not too high ($\lesssim N_c$) they will not play any role because their energy is such that they will not be ionized. Consider further the presence of deep acceptors; these will capture electrons coming from the donors and partially compensate the semiconductor (Appendix 4.2). As long as the concentration of acceptors remains small compared with that of the donors we will still have an n-type semiconductor.

The arguments above show why pure crystals are important in semiconductor applications: until it was possible to manufacture crystals with impurity concentrations less than N_c the phenomena described above were

not understood. Section 4.9 deals with the fabrication of pure semiconductors.

4.7 Semiconductor at Low Temperatures

At **low temperatures** we have to consider **partial ionization** of the donors or acceptors (the Fermi level is close to the donor or acceptor levels). We shall confine ourselves for simplicity to uncompensated n-type semiconductors. Then we always have $n_a^- = N_a = 0$. Electrical neutrality (4.10) becomes

$$p + n_d^+ = n. \tag{4.48}$$

Using Eqs. (4.3) and (4.9) we get

$$n_d^+ = N_d - n_d^0, \tag{4.49}$$

$$n_d^0 = N_d \frac{1}{1 + \frac{1}{2}\exp(E_d - E_F)/kT}, \tag{4.50}$$

$$(n - p) = N_d \frac{\frac{1}{2}\exp(E_d - E_F)/kT}{1 + \frac{1}{2}\exp(E_d - E_F)/kT}. \tag{4.51}$$

Substituting E_F from Eq. (4.15) into Eq. (4.51) we get

$$(n - p)n = \frac{N_c}{2}(N_d - n + p)\exp\left(\frac{E_d - E_c}{kT}\right). \tag{4.52}$$

At **zero temperature** $n = 0; p = 0$; the semiconductor is an insulator. The Fermi level must lie above the donor levels.

At **very low temperature** the ionization of the donors is weak and the hole concentration is negligible, as the Fermi level is very high in the band. Neglecting p and n compared with N_d, Eq. (4.52) becomes

$$n = \left(\frac{N_c N_d}{2}\right)^{1/2} \exp\left(\frac{E_d - E_c}{2kT}\right). \tag{4.53}$$

The electron number increases, with an activation energy equal to half the binding energy of the donor. The Fermi level may be obtained by replacing n by its expression (4.15):

$$E_F - E_c = \frac{E_d - E_c}{2} + \frac{kT}{2}\operatorname{Ln}\frac{N_d}{2N_c}. \tag{4.54}$$

At **intermediate temperature** the exponential is of order 1 and the hole number is still negligible; we can write Eq. (4.52) in the form

$$(N_d - n) \sim \frac{2n^2}{N_c \exp(E_d - E_c/kT)}. \tag{4.55}$$

For $N_d << N_c$ the solution $n = N_d$ is a very good approximation, and we recover the saturation regime.

At **high temperature** we recover the intrinsic regime (4.36): we can replace $\exp[(E_d - E_c)/kT]$ by 1 and must now retain p in Eq. (4.52). Further, we use Eqs. (4.52) and (4.22) in the case where n_i and N_d are both small compared with N_c.

For a partially compensated crystal such that

$$(N_c/2)\exp[-(E_c - E_d)/kT]N_a << N_d$$

one obtains at very low temperature a regime where the electron number varies as $\exp[(E_d - E_c)/kT]$.

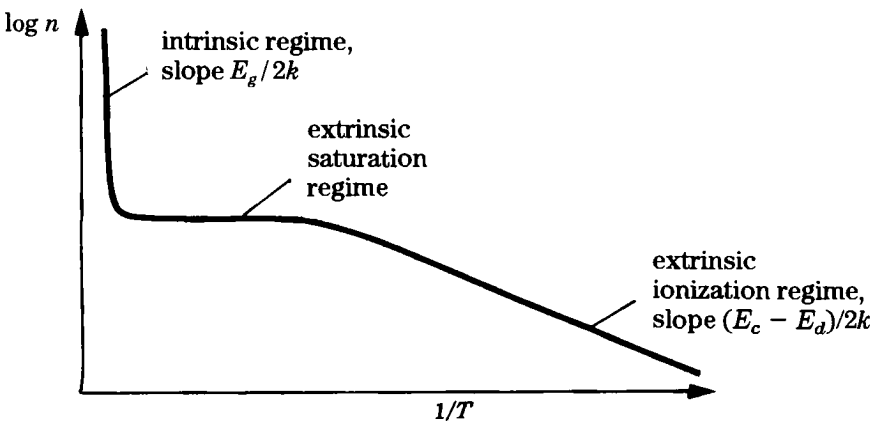


Fig. 4.5. Variation of the logarithm of the concentration as a function of the inverse temperature for an n-type semiconductor. In practice room temperature is in the saturation regime. (From Smith "Semiconductors," Cambridge University Press, 1968.)

Figure 4.5 shows the behavior of the concentration as a function of the inverse temperature. At zero temperature the concentration tends to zero. Similar results hold for a p-type semiconductor.

Exercise: using the data of Chaps. 3 and 4 (and a pocket calculator), show that in silicon doped with 10^{21} *atoms of phosphorus per* m^3, *the saturation regime where the number of carriers is constant extends from about* 90 K, *where* $n \sim 0.9N_d$, *to* 500 K, *where* $n \sim 1.1N_d$.

Figure 4.6 shows the variation of Fermi level with temperature for an n-type semiconductor (upper figure) and a p-type semiconductor (lower figure). At zero temperature in the n-type semiconductor the Fermi level is between the donor level and the conduction band.

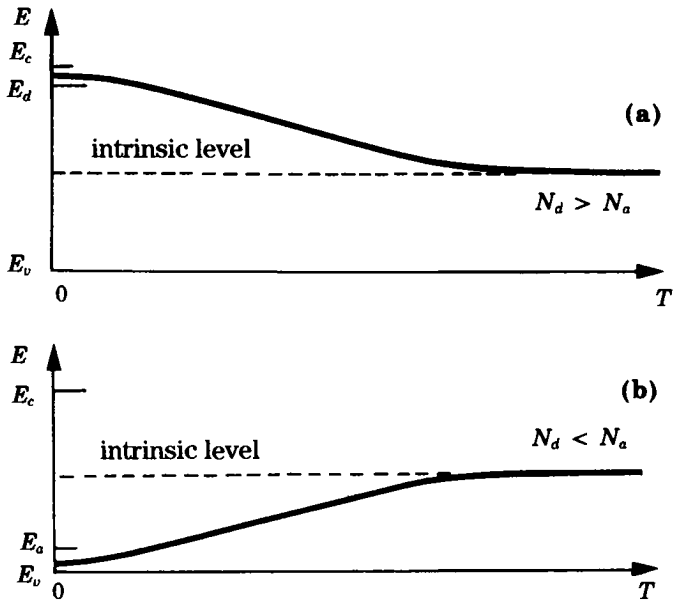


Fig. 4.6. Position of the Fermi level as a function of temperature in the cases of (a) n doping, and (b) p doping. Note the different units of the abscissa compared with Fig. 4.5. In the saturation regime the number of carriers is fixed but the Fermi level changes. (From Smith, "Semiconductors," Cambridge University Press, 1968.)

The problem in Appendix 4.2 shows how a semiconductor of given resistivity is manufactured in practice by choosing the type of dopants (deep or shallow) and their concentration.

4.8 Application: The Semiconducting Thermometer

These results obviously lead to the principle of low-temperature thermometers which measure the resistance of an extrinsic semiconductor. At and above room temperature one generally uses metal resistors as thermometers (such as the etalon thermometer using a platinum resistor), but this is not possible at low temperature because metal resistivity does not depend strongly on the temperature in this range.

The reason for this is that in a metal the number of carriers is constant and the collision time appearing in the resistivity is $\tau(E_F)$, where E_F is the Fermi energy of the crystal. What appears in the resistivity, then, is the collision time of electrons with the Fermi velocity v_F, independent of the temperature. This time varies with temperature at high temperatures, as the thermal vibrations of the lattice are the main source of collisions, and their amplitude increases with temperature. By contrast at low tem-

perature this collision process becomes negligible, and the mean free path is correspondingly infinite. The mean free path, and thus $\tau(E_F)$ is then determined by collisions of electrons of velocity v_F with impurities, so that the metal resistivity is then independent of the temperature.

On the other hand, for a semiconductor in the extrinsic regime the conductivity, which varies with the number of free carriers, remains a function of the temperature at low temperature as we have seen in Fig. 4.5. Arsenic-doped germanium is used currently as a thermometer between 0.1 and 100 K (Fig. 4.7).

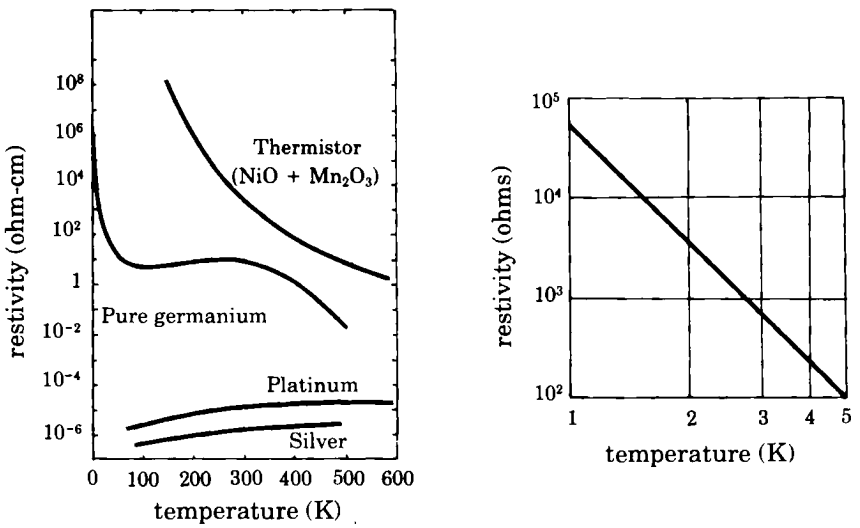


Fig. 4.7. (a) Variation of the resistivity of pure germanium as a function of temperature. Note the rapid rise of the resistivity at very low temperature. (b) Calibration curve for a germanium thermometer. Note the log scale.

4.9 Growth of Pure Crystals

We have seen how important it is to have pure single crystals so as to obtain properly controlled resistivities. Silicon is the basic material for almost all products involving semiconductor components throughout the world. The use of these materials implies a degree of purity out of all comparison with what is normally required in other domains. In many applications the

presence of deep recombination centers must be avoided and purities of the order of 10^{-9} have to be reached for some elements.

These single crystals are almost always obtained using a molten bath with a composition close to the desired one. This removes the complications associated with non-uniform materials. To reach sufficient purity one has to start with fairly pure polycrystalline material, itself obtained by hydrogen reduction of $SiCl_4$, which is liquid and can be distilled several times for purification.

One of the main problems for the growth of single crystals of large size is the control of the temperature gradients which must exist for solidification to occur. When the crystal forms from the liquid, the enthalpy of fusion ΔH_f is released and must be removed from the system while preserving its homogeneity. Usually one uses the extraction or Czochralsky method illustrated in Fig. 4.8. A small seed crystal is mounted on a shaft and brought into contact with the surface of the bath. The temperature gradients are controlled so that the crystal grows at the surface of the seed.

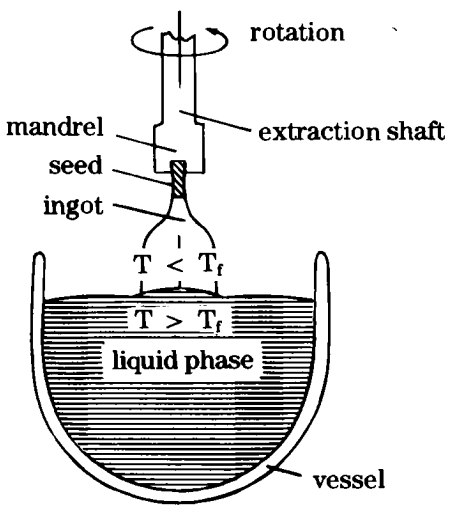


Fig. 4.8. Growth of a single crystal by the extraction method. From Leturcq and Rey, "Physique des Composants Actifs à Semi-Conducteurs," Dunod, 1978.

The crystal is then slowly pulled while being rotated. Crystals up to 50 cm in diameter have been obtained by this method; present standard diameters for microelectronic applications are from 4 to 6 inches (10 to 15 cm).

However in this method the vessel remains as a source of contamination. There is almost always pollution by the vessel which may be slightly

attacked at the bath temperature (1418°C for silicon). It is therefore preferable to eliminate the vessel. This is possible using the ingenious melting zone technique shown in Fig. 4.9. A polycrystalline ingot is heated locally by induction until a narrow melting zone forms. The slow displacement of the heating coil moves the liquid region as the polycrystalline material fuses and solidifies to a single crystal.

Once the crystal is obtained it is possible to purify it further by zone refining, a technique developed by William Pfann at the Bell Laboratories. This procedure has improved the attainable purity limits by several orders of magnitude and one can now obtain silicon crystals with a purity of 10^{-10}. It is true to say that without this technique a large part of present electronics and information technology would not exist. Like many discoveries, zone refining seems obvious once seen, but required a stroke of genius to think of it.

Let us consider a solid in equilibrium with the liquid of the melting zone. We define a coefficient of liquid–solid segregation:

$$K = \frac{C_L}{C_S}, \tag{4.56}$$

where C_L and C_S are the concentrations in the liquid and the solid. In general K is much greater than 1, as the impurities are much more soluble in the liquid than the solid. One starts at $z = 0$ with a crystal having a uniform impurity concentration C_0. The first melting zone has concentration C_0. By contrast the first solidified layer obtained as the zone rises has a concentration $C_0' = C_0/K$ and is therefore purer. As the melting zone moves, it gains

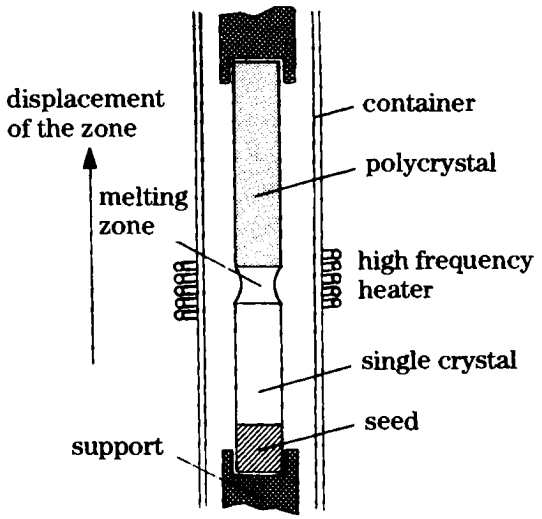


Fig. 4.9. Purification of a single crystal by the melting zone method. From Leturcq and Rey, "Physique des Composants Actifs à Semi-Conducteurs," Dunod, 1978.

in impurities and leaves behind a purified crystal until the liquid reaches a concentration KC_0. Subsequent motions of the zone do not purify but make the crystal uniform. After a passage along the crystal, it is allowed to cool and one starts again at $z = 0$, heating again from the concentration $C_0' = C_0/K$. The first crystallized zone after the second passage thus has a concentration $C_0'' = C_0/K^2$. One thus makes several passages, beginning at the origin, obtaining the remarkable purities we have mentioned.

The cost of electronic grade single crystals of silicon is about \$100/kg, a remarkably low price for such a technological achievement.

Appendix 4.1

Occupation Number of a Donor Level

We first recall the definition of the Fermi factor. We consider the grand canonical ensemble where the density matrix D is written

$$D = \frac{\exp[-\beta(\mathcal{H} - E_F N)]}{\mathrm{Tr}\exp[-\beta(\mathcal{H} - E_F N)]}, \tag{4.57}$$

where $\beta = 1/kT$, $\mathcal{H}$ is the Hamiltonian and N the number of particle operators. The operators $\mathcal{H}$ and N have the eigenvalues $\Sigma_{\text{states}} n_i E_i$ and $\Sigma_{\text{states}} n_i$, respectively.

Here the states i are completely defined (energy, orbital quantum number, and spin determined), and the eigenvalues of n_i are 0 and 1 for each state i. Then for each state

$$< n_i > = \mathrm{Tr}\ n_i D, \tag{4.58}$$

$$\begin{aligned} &= \frac{\mathrm{Tr}\ n_i \exp[-\beta(n_i E_i - E_F n_i)] \times \mathrm{Tr}_{j\neq i} \exp[\Sigma_j - \beta(n_j E_j - E_F n_j)]}{\mathrm{Tr}\exp[-\beta(n_i E_i - E_F n_i)] \times \mathrm{Tr}_{j\neq i} \exp[\Sigma_j - \beta(n_j E_j - E_F n_j)]} \\ &= \frac{\mathrm{Tr}\ n_i \exp[-\beta(n_i E_i - E_F n_i)]}{\mathrm{Tr}\exp[-\beta(n_i E_i - E_F n_i)]} = \frac{1}{1 + \exp\beta(E_i - E_F)}. \end{aligned} \tag{4.59}$$

For a donor level we wish to know if an electron can occupy any of the localized levels (Fig. 4.10), whatever its spin.

Consider first the donor ground state, a hydrogen-like $1s$ state. In fact, taking account of the spin, we have two states, $1s\uparrow$ and $1s\downarrow$. Because of electron repulsion, there is only one electron to place, and there are thus three possible states: empty, one electron in $1s\uparrow$, and one electron in $1s\downarrow$.

The occupation number $< n_{1s} >$ of the $1s$ level can be expressed by using Eq. (4.59) for the levels $1s\uparrow$ and $1s\downarrow$:

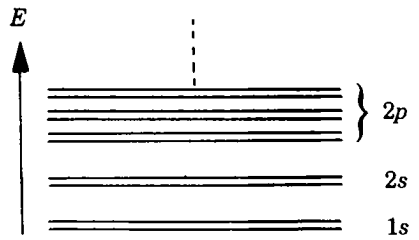


Fig. 4.10. Schematic donor levels including possible orbital and spin degeneracies.

$$< n_{1s} > = < n_{1s\uparrow} > + < n_{1s\downarrow} >, \tag{4.60}$$

$$= \sum_{i=\emptyset,s\uparrow,s\downarrow} \frac{\mathrm{Tr}\, n_i \exp[-\beta(n_i E_i - E_F n_i)]}{\mathrm{Tr} \exp[-\beta(n_i E_i - E_F n_i)]}, \tag{4.61}$$

$$= \frac{2\exp[-\beta(E_{1s} - E_F)]}{1 + 2\exp[-\beta(E_{1s} - E_F)]},$$

$$< n_{1s} > = \frac{1}{1 + (1/2)\exp[\beta(E_{1s} - E_F)]}. \tag{4.62}$$

This is expression (4.3), which we use to calculate the occupation number of the donor levels.

A fully rigorous treatment takes account of all the localized donor states. For the $2p$ level for example, there are six sublevels of the same energy ($l = 1, m = -1, 0, 1$, and $s = \uparrow$ or $\downarrow$). Each of these sublevels may be occupied by zero or one electron. When there is one electron the $2p_{-1\uparrow}$ term in the occupation number is $\exp[-\beta(E_{2p} - E_F)]$, and the six sub-levels contribute in the same way. For a level α of degeneracy $g_\alpha = 2(2l + 1)$, where l is the orbital quantum number, the term $\exp[-\beta(E_\alpha - E_F)]$ occurs g_α times. Further, as the localized donor levels can be occupied by one electron at most, there is in all only a single empty level. Hence the mean occupation number $< n >$ of the ensemble of localized donor levels can be written as

$$< n > = \Sigma_\alpha < n_\alpha >$$

$$= \frac{2\exp[-\beta(E_{1s} - E_F)] + \ldots + g_\alpha \exp[-\beta(E_\alpha - E_F)]}{1 + 2\exp[-\beta(E_{1s} - E_F) + \ldots + g_\alpha \exp[-\beta(E_\alpha - E_F)]}, \tag{4.63}$$

$$< n > = \frac{1}{1 + \{\Sigma_\alpha g_\alpha \exp[-\beta(E_\alpha - E_F)]\}^{-1}}. \tag{4.64}$$

In fact the number and degeneracy of the localized excited states is such that the sum $\Sigma_\alpha g_\alpha \exp[-\beta(E_\alpha - E_F)]$ diverges. This means that if we consider **all** the localized excited states we will always find $< n >= 1$,

i.e., the electron will never be in the conduction band. This paradox, well known for the hydrogen atom, can be resolved. The Bohr radius grows with α, and we should count in the sum only **those states which do not overlap**; if this condition fails the electrostatic potential will no longer go as $1/r$ and the calculation of the donor states is no longer correct.

In all practical cases the numerical values deduced from formulas (4.62) and (4.64) differ very little, and we shall use Eq. (4.62) in this book.

Appendix 4.2 ×

Problem: Substrates for Microelectronics

The microelectronics industry makes great use of epitaxy, the crystalline growth of semiconductors, or semiconductor devices, on a substrate of similar type. The substrate acts as the mechanical and thermal support, but must not short-circuit the components. Thus it must have a very high resistance.

The objective of this problem is showing how to use appropriate doping by deep impurities to render insulating a single crystal of impure gallium arsenide (containing residual impurities). We recall that the electrical resistivity ρ is related to the conduction electron and hole number densities n, p by the relation

$$\rho = (ne\mu_e + pe\mu_h)^{-1},$$

where μ_e and μ_h denote the mobilities of these carriers, and e the electron charge.

The questions in this problem aim at numerical estimates of the resistivity. We simplify the problem as much as possible by adopting adequate approximations (for the Fermi factors in particular).

We give the following characteristics of gallium arsenide: relative dielectric constant $\epsilon_r = 13$; band gap $E_c - E_v = 1.4$ eV; effective electron mass $m_e = 7 \times 10^{-2} m$; electron mobility $\mu_e = 8500$ cm$^2 \cdot$ V$^{-1} \cdot$ s^{-1}; effective hole mass $m_h = 0.5m$; hole mobility $\mu_h = 400$ cm$^2 \cdot$ V$^{-1} \cdot$ s^{-1}; here m is the electron mass in vacuo. It is convenient to use the following numerical values:

$$N_0 = \frac{2}{h^3}(2\pi mkT)^{3/2} = 2.10^{25} \text{ m}^{-3} \text{ for } T = 300 \text{ K},$$

$$E_H = \frac{e^4 m}{2(4\pi\epsilon_0\hbar)^2} = 13.6 \text{ eV}.$$

E_H is the binding energy of the hydrogen ground state.

First Question

Assume the semiconductor is **completely** pure. What is its resistivity at $T = 300$ K?

Second Question

Because of the different affinities of *Ga* and *As* atoms, it is impossible *in practice* to produce *GaAs* crystals with less than 10^{20} impurities per m^3. We assume that these impurities are shallow donors. They then introduce energy levels close to the conduction band which we can calculate by likening the donor to a modified hydrogen atom. For this atom we take account of the screening effect through the dielectric constant, and use the fact that the conduction electrons are quasi-particles of effective mass m_e, as shown in the book. To calculate the statistical distribution of the electrons we assume that each impurity has the effect of introducing into the band gap *exactly one state* (without spin degeneracy) of energy E_d whose position with respect to the bottom of the conduction band is that of the ground level of this pseudo-hydrogen atom.

What is the position of this ground level E_d and how does the binding energy of the electron in this level compare with kT for $T = 300$ K?

What is the position of the chemical potential at $T = 300$ K in the purest material obtainable in practice? What is the corresponding resistivity?

Third Question

We can *deliberately* introduce some chromium atoms into the semiconductor. We assume that their number density N_{Cr} is 10^{23} m^{-3}. Each chromium atom introduced into the crystal brings an electron with it. The earlier approximation of an equivalent hydrogen atom does not work for these atoms. We assume that to each chromium atom there corresponds in the band gap an energy level E_{Cr}, this time doubly degenerate, situated 0.7 eV below the bottom of the conduction band.

Write down the conservation of the electron number in this system. Deduce that the Fermi level is very close to E_{Cr}, and that the introduction of chromium increases the resistivity. What is the maximum resistivity one can reach at $T = 300$ K?

Solutions

First Question

We assume that the chemical potential is in the band gap, i.e., we set $\beta = 1/kT, \beta(E_c - E_F) >> 1$, and $\beta(E_F - E_v) >> 1$ and thus approximate by neglecting the $+1$ term in the denominator of the Fermi factor

$$f(E) = \frac{1}{\exp\beta(E - E_F) + 1} \simeq \exp[-\beta(E - E_F)]. \tag{4.65}$$

Under these conditions the quasi-particles obey ideal gas statistics, and the electron and hole number densities are given by

$$n = \frac{2}{h^3}(2m_e\pi kT)^{3/2}\exp\beta(E_F - E_c), \tag{4.66}$$

$$p = \frac{2}{h^3}(2m_h\pi kT)^{3/2}\exp\beta(E_v - E_F), \tag{4.67}$$

so that $n = (m_e/m)^{3/2}N_0\exp\beta(E_F - E_c)$ and $p = (m_h/m)^{3/2}N_0 \times \exp\beta(E_v - E_F)$.

We eliminate the chemical potential by taking the product of these two expressions:

$$np = N_0^2\left(\frac{m_e}{m}\right)^{3/2}\left(\frac{m_h}{m}\right)^{3/2}\exp[-\beta(E_c - E_v)] = n_i^2. \tag{4.68}$$

For a completely pure semiconductor the condition of electrical neutrality, expressing the conservation of the total number of electrons, can be written

$$n = p.$$

We get

$$n = p = N_0\left(\frac{m_e m_h}{m^2}\right)^{3/4}\exp[-\beta(E_c - E_v)/2]. \tag{4.69}$$

Numerically we have $n = p = n_i \simeq 1.1 \cdot 10^{12}$ m^{-3} at $T = 300$ K. The chemical potential E_F is given by

$$\begin{aligned}\beta(E_F - E_c) &= \mathrm{Ln}\,[(n/N_0)(m/m_e)^{3/2}] \\ &= \mathrm{Ln}\,[(p/N_0)(m/m_h)^{3/2}].\end{aligned} \tag{4.70}$$

Given $n = p << N_0$, we justify *a posteriori* the original assumption. The resistivity ρ is $\rho = [ne(\mu_e + \mu_h)]^{-1}$. We find $\rho \sim 6.10^6\Omega \cdot$ m.

Second Question

The binding energy of the ground level of the pseudo-hydrogen atom is

$$E_c - E_d = E_H \times \frac{m_e}{m} \times \frac{1}{\epsilon_r^2}. \tag{4.71}$$

We have now taken account of the effective mass m_e of the quasi-particles; the screening effect caused by all the electrons of the solid leads to an attractive potential of the form $e^2/4\pi\epsilon_0\epsilon_r r$ instead of $e^2/4\pi\epsilon_0 r$.

Numerically we have $E_c - E_d \sim 8.10^{-3}$ eV. This is less than the energy kT for $T = 300$ K, which is 25×10^{-3} eV. The donor is neutral when the level E_d is occupied; it is positively charged when it is empty. The neutrality condition for a material containing N_d donors per m^3 is

$$n = p + [1 - f(E_d)]N_d. \tag{4.72}$$

The chemical potential is now closer to the bottom of the conduction band (E_c). Assuming that p is negligible, we can then see that:
• if $1 - f(E_d) << 1$, then the chemical potential lies above E_d, and is thus close to the conduction band and p is indeed negligible;
• if $1 - f(E_d) \sim 1$ then $n \gtrsim N_d$ from Eq. (4.72). If $N_d > n_i$, then from Eq. (4.68) $p < n$, i.e., $p < 1.1 \times 10^{12}$ m^3, while $N_d \geq 10^{20}$ m^{-3}, so that $p << N_d$.

Now, given that at 300 K, $\beta(E_v - E_d)$ is less than 1, the Fermi factor $f(E)$ has the same order of magnitude for $E = E_c$ and $E = E_d$:

$$f(E_c) \sim f(E_d).$$

For $N_d = 10^{20}$ m^{-3} and $N_0 = 2 \times 10^{25}$ m^{-3}:

$$f(E_d)N_d \ll f(E_c)\left(\frac{m_e}{m}\right)^{3/2} N_0 = n.$$

Expression (4.72) thus reduces to

$$n \simeq N_d$$

and we can clearly neglect p by comparison with N_d (the second assumption above).

The chemical potential is given by

$$\beta(E_c - E_F) = \mathrm{Ln}[(N_0/N_d)(m_e/m)^{3/2}], \tag{4.73}$$

and we find $E_c - E_F \simeq 0.2$ eV and thus $E_F - E_v \simeq 1.2$ eV.

The chemical potential is below E_d, closer to the conduction band than the valence band, justifying the approximation $n >> p$, but it is far enough from E_c for the approximation of the Fermi factor made in the first question to hold.

The corresponding resistivity is then

$$\rho = \frac{1}{N_d e \mu_e},$$

where $N_d = 10^{20}\text{m}^{-3}$ so that $\rho \simeq 0.07\,\Omega \cdot \text{m}$.

Third Question

The chromium levels are clearly deeper in the band gap than the donor levels; thus the donor electrons are trapped in these levels and E_F is decreased. The energy E_{Cr} is lower than the value of the chemical potential calculated in the preceding question. The electron number n will thus decrease and the resistivity increase. The electrons come either from the valence band (p in number) or the chromium levels (N_{Cr} electrons) or the donors (N_d). At equilibrium they are either delocalized in the conduction band (n) or localized in the chromium sites $[2f(E_{\text{Cr}})N_{\text{Cr}}]$ or donor sites $[f(E_d)N_d]$. Hence the balance:

$$p + N_{\text{Cr}} + N_d = n + 2f(E_{\text{Cr}})N_{\text{Cr}} + f(E_d)N_d. \tag{4.74}$$

Since $N_{\text{Cr}} \sim 10^{23}\ \text{m}^{-3}$ is much larger than all the concentrations N_d, n, p under consideration, this equation requires

$$N_{\text{Cr}} \simeq 2f(E_{\text{Cr}})N_{\text{Cr}}, \tag{4.75}$$

that is, $f(E_{\text{Cr}}) \simeq 1/2$, and thus that E_F coincides with E_{Cr} to within a few kT. We say that the chemical potential is **pinned** at the level E_{Cr} in the center of the band. We deduce

$$n = \left(\frac{m_e}{m_0}\right)^{3/2} N_0 \exp\beta(E_{\text{Cr}} - E_c), \tag{4.76}$$

$$p = \left(\frac{m_h}{m_0}\right)^{3/2} N_0 \exp\beta(E_v - E_{\text{Cr}}). \tag{4.77}$$

With $E_c - E_{\text{Cr}} = E_{\text{Cr}} - E_v = 0.7$ eV we obtain $n = 2.5 \times 10^{11}\ \text{m}^{-3}$ and $p \sim 5 \times 10^{12}\ \text{m}^{-3}$. Thus $\rho \simeq 1.5 \cdot 10^7\,\Omega \cdot \text{m}$.

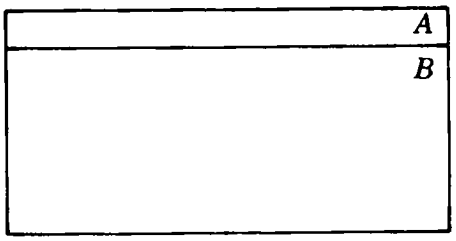


Fig. 4.11. Epitaxial structure: A: active layer; B: semi-insulating substrate.

The introduction of chromium allowed us to obtain a larger resistivity than that of the perfectly intrinsic material. There are a few less very mobile electrons and a few more fairly immobile holes than in the pure material. It is thus a method for obtaining a material of resistivity comparable to that of the pure material, starting from a material which necessarily contains impurities. This kind of material is called a **semi-insulating** semiconductor. We thus have a method allowing us to transform an impure substrate B into an insulator by strongly doping it with deep impurities. We can then grow by epitaxy a high quality crystalline layer A, whose electrical properties can be controlled by selective doping with shallow impurities (Fig. 4.11).

5. Transport Phenomena in Semiconductors

5.1 Introduction

Whenever an electric field, a concentration gradient, or temperature gradient is present in a semiconductor we observe charge transport (electric current) mass transport (diffusion of carriers), or energy transport (heat conduction). Further, when a semiconductor is subject to photoexcitation there occurs what we may call an internal photochemical reaction: the creation of electron-hole pairs through the excitation of an electron from the valence band into the conduction band.

For each of these different types of excitation there exist mechanisms restoring the system to equilibrium, with characteristic rates which are, however, different. We shall see that the average time between collisions, measureable from the mobility $\mu = e\tau/m^*$ or from the width of the cyclotron resonance, is of the order of 10^{-13} s. This is very short compared with the lifetime of an electron-hole pair created by light. The latter time can for example be found from the decay of "photoconductivity," the variation of conductivity connected with the presence of additional electrons and holes, following a very short flash of light (cf. Chap. 6). In this way we measure lifetimes between 10^{-3} and 10^{-9} seconds.

It is important to understand that even in the absence of any applied excitation, there is ceaseless creation and destruction of electron-hole pairs by thermal motion within the semiconductor. We shall see that the characteristic time of this spontaneous process is exactly the lifetime, and thus very long compared with the time between collisions. **Because these times are very different we can study these processes separately, i.e., determine the transport properties such as conductivity, by regarding the electrons and holes as non-interacting gases.** This is the approach of the present chapter. As a second stage we can study the much slower effects of creations and annihilations (recombinations) of electron-hole pairs by light or through other causes: this is the subject of Chap. 6.

We first recall the elementary transport model constructed by Drude in 1900, which relates the finite value of the electrical conductivity to microscopic electron collisions (Sect. 5.2). The predictions of this model will be more rigorously justified in Sect. 5.3 by use of the Boltzmann transport equation. The main results obtained for semiconductors, and the orders of magnitude of the characteristic times and lengths, are given in Sec. (5.4). On a first reading the latter section can be studied after reading only Sects. 5.2 and 5.3d.

5.2 Drude's Model of Conductivity and Diffusion

We have seen in the preceding chapter that the occupation probability of conduction band states of energy E can be approximated by a function proportional to a Boltzmann factor $\exp[-E/kT]$. Moreover we saw in Chap. 2 that electrons respond to an external force in accord with classical dynamics (2.38), provided we replace their mass by the effective mass m_e. This is also true of the holes. We may therefore try to use a classical treatment to understand transport properties. This is the basic idea in the application of Drude's model to semiconductors.

We thus consider a perfect electron gas, obeying classical mechanics, and confined within a solid by the Coulomb interaction with the ions, which appear only as a potential well of the size of the crystal. Like classical gas particles, the electrons are subject to random collisions. We know (Sect. 2.2) that these collisions only occur with imperfections in the crystal and the boundaries of the potential well.

Let us consider the example of Coulomb scattering of an electron by ionized impurities that deflect its path. If the impurities are randomly distributed in the solid, the scattering probability for an electron by impurities during the time interval $\mathrm{d}t$ is independent of the observation time t, and proportional to $\mathrm{d}t$. We write it as $\mathrm{d}t/\tau$, where the characteristic time τ depends on the impurity concentration. The electron will be more sensitive to an impurity if it "feels" it for longer, and thus if its kinetic energy E is lower: this must be reflected in the dependence of τ on E.

In Drude's model, whatever the collision mechanism, we take the scattering probability for an electron in the interval $\mathrm{d}t$ to be $\mathrm{d}t/\tau$, where the collision time is assumed to be independent of E. We call $p(t)$ the probability that an electron suffers no collision between $t = 0$ and the time t. The probability of surviving until $t + \mathrm{d}t$ without collisions is the product of the probability of reaching t without collision, and that of suffering no collision between t and $t + \mathrm{d}t$, so that

$$p(t + dt) = p(t)(1 - dt/\tau). \tag{5.1}$$

We get

$$dp = -p\, dt/\tau, \tag{5.2}$$

$$p(t) = \exp(-t/\tau) \tag{5.3}$$

The probability density $P(t)\mathrm{d}t$ that an electron has suffered no collisions between 0 and t but then collides between t and $t + \mathrm{d}t$ is

$$P(t)\, dt = \exp(-t/\tau)\, dt/\tau. \tag{5.4}$$

The probability density $P(t)$, an exponential distribution, gives the mean time $< t >$ between collisions:

$$< t > = \int_0^\infty tP(t)\, dt = \tau. \tag{5.5}$$

The average distance λ_c between collisions, called the mean free path, is the product of τ and the average velocity. We will see in Sect. 5.4 that τ is of the order of 10^{-13} s. For an effective mass $m_e \sim 0.1m$, where m is the free electron mass, and a thermal speed v at 300 K of several 10^5 m/s, $\lambda_c = v\tau$ is about 20 nm.

Using the exponential distribution we can calculate the electrical conductivity and diffusivity.

5.2a Electrical Conductivity ×

In the absence of an electric field, collisions make all directions of motion equally probable, each one on average occurring without memory of the preceding electron velocity. The electron motion thus produces no global electric current.

In the presence of an electric field $\mathcal{E}$ the electrons feel a constant force:

$$\mathbf{F} = -e\boldsymbol{\mathcal{E}} = m_e \frac{d\mathbf{v}}{dt}. \tag{5.6}$$

From the argument of Sect. 2.2 the mass involved here is the effective mass m_e of the electron in the solid. The electrons thus acquire an additional velocity component v in the direction of $\boldsymbol{\mathcal{E}}$ whose magnitude grows linearly with time until the following collision (Fig. 5.1): this component grows proportionally with the time for an interval Δt_1, is abruptly annihilated, and then begins to grow again during the interval Δt_2, and so on.

The mean value of the velocity in this process is, from Eq. (5.5),

$$< \mathbf{v} > = -\frac{e < t >}{m_e}\boldsymbol{\mathcal{E}}, \tag{5.7}$$

$$< \mathbf{v} > = \mathbf{v}_e = -\frac{e\tau}{m_e}\boldsymbol{\mathcal{E}} = -\mu_e\boldsymbol{\mathcal{E}}, \tag{5.8}$$

which is the drift velocity $\mathbf{v}_e$. This quantity is proportional to the electric field $\boldsymbol{\mathcal{E}}$. Here we have introduced the electron mobility μ_e:

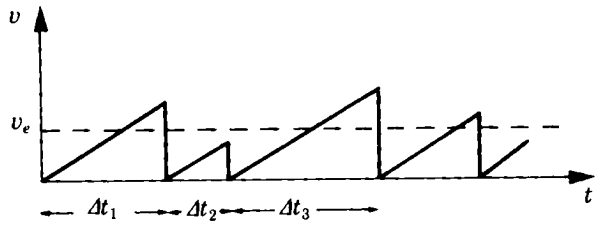


Fig. 5.1. Increase of the velocity component of a free electron along the direction of the applied electric field as a function of time in the Drude model (the resulting velocity is antiparallel to the field). The drift velocity is v_e.

$$\mu_e = \frac{e\tau}{m_e}. \tag{5.9}$$

A convenient way of deriving the results of the Drude model is to introduce into the dynamical equation (5.6) a viscous friction force $-m_e\mathbf{v}/\tau$, where τ is the collision time:

$$m_e\frac{d\mathbf{v}}{dt} = -e\boldsymbol{\mathcal{E}} - m_e\frac{\mathbf{v}}{\tau}. \tag{5.10}$$

The stationary solution ($d\mathbf{v}/dt = 0$) gives again the result (5.8). Equation (5.10) can be extended to more complex cases: sinusoidal driving terms and applied electric and magnetic fields (cf. Appendices 2.6b and 5.1).

The current density $\mathbf{J}_e$ is

$$\mathbf{J}_e = -ne\mathbf{v}_e, \tag{5.11}$$

where n is the carrier density; $\mathbf{J}_e$ is thus proportional to $\boldsymbol{\mathcal{E}}$, allowing us to define the conductivity σ_e:

$$\mathbf{J}_e = \sigma_e\boldsymbol{\mathcal{E}} \tag{5.12}$$

with

$$\sigma_e = ne^2\frac{\tau}{m_e} = ne\mu_e. \tag{5.13}$$

A solid is in the most general case anisotropic, so that the conductivity is a tensor (cf. Appendix 5.1).

Similarly for a semiconductor possessing free holes of effective mass m_h and concentration p, with collision time τ_h, a discussion analogous to that for the electrons gives

$$\mathbf{v}_h = \frac{e\tau_h}{m_h}\boldsymbol{\mathcal{E}} = \mu_h\boldsymbol{\mathcal{E}}, \tag{5.14}$$

where $\mathbf{v}_h$ is the hole drift velocity in the electric field $\boldsymbol{\mathcal{E}}$ and μ_h their mobility, defined by

$$\mu_h = \frac{e\tau_h}{m_h}. \tag{5.15}$$

The hole current $\mathbf{J}_h$ is proportional to $\boldsymbol{\mathcal{E}}$, and we define the hole conductivity σ_h by

$$\mathbf{J}_h = \sigma_h \boldsymbol{\mathcal{E}} \tag{5.16}$$

with

$$\sigma_h = \frac{p\, e^2 \tau_h}{m_h}. \tag{5.17}$$

We can verify (cf. Sect. 3.1) that **the drift velocities of the electrons and holes are in opposite directions, but the current densities $\mathbf{J}_h$ and $\mathbf{J}_e$ are both in the same direction as $\boldsymbol{\mathcal{E}}$ and therefore add.** The total conductivity is thus

$$\sigma = \sigma_e + \sigma_h,$$
$$\sigma = ne\mu_e + pe\mu_h. \tag{5.18}$$

5.2b Diffusion

Let us consider a set of identical particles (e.g., gas molecules) with total concentration n. Assume that some of these particles differ from the others, for example that they are radioactive. Let n_1 be the concentration of these particles. In equilibrium the particles are uniformly distributed in the allowed volume so that the concentrations n and n_1 are independent of position. Let us assume now that the distribution of n_1 is not uniform, but depends on position, e.g., $n_1 = n_1(x)$, while n remains constant. This is not an equilibrium situation. There must therefore be particle motions tending to increase the entropy, i.e., trying to make the concentration n_1 uniform, although there is no net motion of matter. Let us call the particle current $\mathbf{J}_N$. If n_1 is not constant we expect as a first approximation that the current will be proportional to the concentration gradient, here chosen parallel to the x axis:

$$J_{Nx} = -D\frac{\partial n_1}{\partial x}. \tag{5.19}$$

The coefficient of proportionality D is called the diffusion coefficient. For a positive diffusion coefficient, a positive gradient $\partial n_1/\partial x$ gives a negative flux tending to equalize the concentration. This equation, called Fick's law, describes diffusion in very many cases.

If we combine Eq. (5.19) and the equation of particle number conservation:

$$\frac{\partial n_1}{\partial t} = -\nabla \cdot \mathbf{J}_N, \tag{5.20}$$

we obtain the **diffusion equation**

$$\frac{\partial n_1}{\partial t} = D\frac{\partial^2 n_1}{\partial x^2}. \tag{5.21}$$

We immediately see from this equation that if the total particle number remains constant, diffusion moves particles from dense regions, where $\partial^2 n_1/\partial x^2$ is negative, to dilute ones, where it is positive.

5.2c Diffusion in the Drude Model

We consider a classical gas of non-uniform concentration in the Drude model, and attempt to demonstrate Fick's law (5.19) in this case. We assume that the concentration gradient is along x and seek the particle flux across unit surface perpendicular to the x axis at time t. This flux consists of all the particles directed towards this surface from the left since their last collision, from which we must subtract all the particles similarly coming from the right. As the current is proportional to the particle velocity, the total current can be written as

$$\begin{aligned} J_{Nx} = &\int_{-\infty}^{t} n_1[x - v_x(t-t')]v_x \exp[-(t-t')/\tau]\frac{dt'}{\tau} \\ &- \int_{-\infty}^{t} n_1[x + v_x(t-t')]v_x \exp[-(t-t')/\tau]\frac{dt'}{\tau}. \end{aligned} \tag{5.22}$$

This simply states that for a particle to cross the surface at time t with speed v_x requires that it should not suffer a further collision after its last collision at time t' [using the exponential distribution, cf. Eq. (5.4)]. At this time the particle had abscissa $x - v_x(t-t')$, and the number of such particles is $n_1[x - v_x(t-t')]$. Expanding to first order, which is valid only if the gradient is approximately constant over a mean free path, we have

$$J_{Nx} = -\frac{\partial n_1}{\partial x}\int_{-\infty}^{t} 2v_x^2(t-t')\exp[-(t-t')/\tau]\frac{dt'}{\tau}, \tag{5.23}$$

$$J_{Nx} = -\frac{\partial n_1}{\partial x}2v_x^2\tau. \tag{5.24}$$

We now have to average over velocities, but counting only positive v_x, since the particles on the left with velocities to the left do not contribute to the current (Fig. 5.2). For an ideal gas the mean value is

$$< v_x^2 >=< v_y^2 >=< v_z^2 >= kT/m \tag{5.25}$$

and here we have to take one-half of this value. We thus have

$$J_{Nx} = -\frac{\partial n_1}{\partial x}\frac{kT}{m}\tau. \tag{5.26}$$

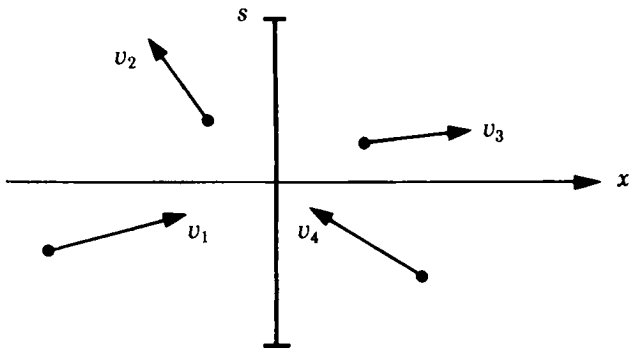


Fig. 5.2. Drude model for diffusion. Electrons with velocities like v_1 cross the surface s at time t and contribute positively to the diffusion current J_x. We have to subtract the contribution of electrons with velocities such as v_4.

We have thus established Fick's equation (5.19) within the framework of the Drude model, and obtained a theoretical value for the diffusion coefficient

$$D = \frac{kT}{m}\tau. \tag{5.27}$$

If we now consider the conduction-band electrons we can give a completely analogous discussion to the extent that the electrons obey classical statistics (Chap. 4). Here, the mean free path is determined by collisions with lattice defects, e.g., impurities or deformations caused by thermal vibrations. The only difference is that we have to replace the mass by the effective mass on taking the mean of v_x^2. We thus get for electrons

$$D_e = \frac{kT}{m_e}\tau_e, \tag{5.28}$$

and for holes

$$D_h = \frac{kT}{m_h}\tau_h. \tag{5.29}$$

If we compare the relations (5.9), (5.28) and (5.15), (5.29) we see that we can write a relation of the form

$$D = \frac{kT}{e}\mu \tag{5.30}$$

for each type of carrier.

This relation has a more general applicability than the restricted cases for which we have derived it, and is called the **Einstein relation**.

Altogether, if there is simultaneously an electric field and a concentration gradient, the two currents, drift and diffusion, add together and we have for the charge current the expression

$$\mathbf{J}_e = en\mu_e \boldsymbol{\mathcal{E}} + eD_e \nabla n. \tag{5.31}$$

For positively charged holes the total current (drift + diffusion) is

$$\mathbf{J}_h = pe\mu_h \boldsymbol{\mathcal{E}} - eD_h \nabla p. \tag{5.32}$$

Note the sign difference between Eqs. (5.31) and (5.32) which reflects the fact that the diffusion of holes or electrons is always opposite to the concentration gradient, whereas this is not true of the associated electric current, because of the sign of the charge carried by each particle.

5.2d Limitations of the Drude Model

This simple model is based on an exponential distribution in which a unique collision time τ appears, and allows one to describe a conductivity, i.e., Ohm's law, and the diffusion constant appearing in Fick's law. It assumes that the electrons obey classical statistics, although this is not always true in semiconductors. We can check its validity by analyzing the behavior of the mobility, as measured, e.g., by the Hall effect: one observes that the mobility varies with temperature (see Sect. 5.4), which implies that τ is a function of electron energy. This is true of collisions with ionized impurities, among others. We thus need a more rigorous formalism based on the Boltzmann transport equation, which we shall give in Sect. 5.3a. We shall recover an expression for the conductivity (Sect. 5.3b) similar to that given by the Drude model, with an interpretation of τ as a certain average of the collision times $\tau(E)$ over the energy E. We shall also be interested in the phenomenon of particle diffusion linked to the presence of a concentration gradient (Sect. 5.3c) and we will show that the Einstein relation is very general.

5.3 Semiclassical Treatment of Transport Processes

The theory of transport processes deals with the relation between currents and the forces which produce them. The formulations used for the calculation of transport properties are based essentially on classical mechanics in the sense that one regards the electrons and holes as particles with well-defined positions and crystal momenta, except for the duration of a collision, which is assumed negligible compared to the time between collisions: this system is completely analogous to an ideal classical gas. However, a part of the discussion requires quantum mechanics, namely the treatment

of the collision process itself. It is in this sense that the treatment is called semiclassical.

The physical conditions for a semiclassical treatment are first, the distribution of energy levels must be continuous, or at least the distance between energy levels much smaller than kT. Second, interactions or correlations between particles must be weak. Next, the time that a particle spends in a state of kinetic energy E, i.e., the mean time between collisions, must be large compared with $\hbar/E$, so that the energy of the state is well-defined. Also, the spatial variation of the applied fields must be small over a mean free path. Finally, the time variation of these fields must be small during a collision event.

These conditions are satisfied in semiconductors for a very wide range of temperatures and fields. We can thus apply the classical treatment to the electron and hole gas in semiconductors. We recall here the main lines of this treatment.

In (a) we give the evolution equation for the probability that an electron has position $\mathbf{r}$ and momentum $\mathbf{k}$ at time t. Conductivity is introduced in (b), and diffusion in (c). The main results are summarized in (d), and may be assumed if it is desired to skip Sects. 5.3a,b,c in a first reading.

Let us first make the terminology precise. The treatment below was conceived by Boltzmann to describe the transport properties of gases. In solid state physics the term "Boltzmann equation" is always used to describe the semiclassical transport equation, even for metals, where the electrons obey Fermi–Dirac statistics.

5.3a The Boltzmann Equation for a Semiconductor

Let $f(\mathbf{k}, \mathbf{r}, t)$ be the phase density (or distribution function), such that at time t the number of particles with momentum within $\mathrm{d}^3\mathbf{k}$ of $\mathbf{k}$ and position within $\mathrm{d}^3\mathbf{r}$ of $\mathbf{r}$ is $f(\mathbf{k}, \mathbf{r}, t)\mathrm{d}^3\mathbf{k}\,\mathrm{d}^3\mathbf{r}$: this represents the number of particles in the volume element of the one particle phase-space at time t. The distribution function may be either Maxwell–Boltzmann or Fermi–Dirac, according to the system under consideration. To clarify these ideas, Fig. 5.3 shows a section of phase-space along the (x, k_x) plane. The Boltzmann transport equation which governs the evolution of $f(\mathbf{k}, \mathbf{r}, t)$ can be obtained as follows. Between the times t and $t + \mathrm{d}t$ the points representing the particles move smoothly towards the volume element $\mathrm{d}^3\mathbf{k}'\mathrm{d}^3\mathbf{r}'$, equal to $\mathrm{d}^3\mathbf{k}\mathrm{d}^3\mathbf{r}$ to second order, under the action of external forces and diffusion.

However because of collisions some particles (**a**) are "lost," while others (**b**) are "gained." The conservation of particle number imposes only the time independent integral relation

$$\int f(\mathbf{k}, \mathbf{r}, t)\; d^3\mathbf{k}\; d^3\mathbf{r} = \text{constant}. \tag{5.33}$$

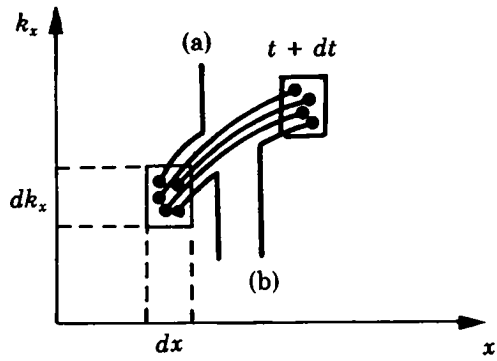


Fig. 5.3. Phase-space with two dimensions x and k_x. Time evolution of particles in the range $\mathrm{d}x\ \mathrm{d}k_x$.

Comparison of the number of particles within $\mathrm{d}^3\mathbf{k}\mathrm{d}^3\mathbf{r}$ of $(\mathbf{k},\mathbf{r})$ at time t, and the number within $\mathrm{d}^3\mathbf{k}\mathrm{d}^3\mathbf{r}$ of $(\mathbf{k}+\mathrm{d}\mathbf{k},\mathbf{r}+\mathrm{d}\mathbf{r})$ at $t+\mathrm{d}t$ requires that

$$f(\mathbf{r}+\mathbf{dr},\mathbf{k}+\mathbf{dk},t+dt) = f(\mathbf{r},\mathbf{k},t) + \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} dt, \tag{5.34}$$

where $(\partial f/\partial t)_{\text{coll}}$ is the variation of particle number caused by collisions. This gives to first order

$$\frac{\partial f}{\partial \mathbf{r}}\frac{d\mathbf{r}}{dt} + \frac{\partial f}{\partial \mathbf{k}}\frac{d\mathbf{k}}{dt} + \frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \tag{5.35}$$

which is the **Boltzmann equation**. Using Eq. (2.33), the equation of motion in the reciprocal space, this can be rewritten as

$$\mathbf{v}\cdot\nabla_{\mathbf{r}} f + \frac{1}{\hbar}q(\boldsymbol{\mathcal{E}} + \mathbf{v}\times\boldsymbol{\mathcal{B}})\cdot\nabla_{\mathbf{k}}\, f + \frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}}. \tag{5.36}$$

The first term represents the density variation in phase-space when the point under consideration moves in a spatially inhomogeneous system. This term is related to the diffusion. The second term shows the variation of f under the effect of an electromagnetic force applied to a charge q. The third term gives the explicit time dependence, for example in the case of sinusoidal driving force. This term vanishes in a steady state. The collision term is difficult to calculate exactly. One often uses the so-called **relaxation time approximation**, in which

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = -\frac{f - f_0}{\tau(\mathbf{k})}, \tag{5.37}$$

that is, if we impose a distribution function f differing from the equilibrium distribution f_0, the system will return to equilibrium in a characteristic time

$\tau(\mathbf{k})$. This hypothesis gives a good description of most collision processes in semiconductors. Thus the steady-state distribution function f is the solution of

$$\mathbf{v} \cdot \nabla_{\mathbf{r}} f + \frac{q}{\hbar}(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \boldsymbol{\mathcal{B}}) \cdot \nabla_{\mathbf{k}} f = -\frac{f - f_0}{\tau(\mathbf{k})}. \tag{5.38}$$

5.3b Conductivity

This is the effect produced by an electric field in a homogeneous semiconductor at constant temperature in the absence of a magnetic field. If there is no temperature gradient we have $\nabla_{\mathbf{r}} f = 0$ and Eq. (5.38) gives for the stationary state

$$f = f_0 - \tau \hbar^{-1} q \, \boldsymbol{\mathcal{E}} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}). \tag{5.39}$$

We assume that the deviation of f from the equilibrium distribution function f_0 is small, so that in the last term of Eq. (5.39) we can replace f by f_0, which depends on k only through the energy. Then

$$f = f_0 - \tau \hbar^{-1} \frac{\partial f_0}{\partial E} \, q\boldsymbol{\mathcal{E}} \cdot \nabla_{\mathbf{k}} E_{\mathbf{k}} \tag{5.40}$$

so that for an electron, with the field taken along the x direction,

$$f = f_0 + \tau e \mathcal{E} v_x \frac{\partial f_0}{\partial E}. \tag{5.41}$$

The electric current density J_x corresponding to a state $\mathbf{k}$ is $-2ev_x$ (the factor 2 comes from the spin). Since the density in $\mathbf{k}$-space is $(1/2\pi)^3$ the total current is

$$J_x = \frac{-e^2 \mathcal{E}}{4\pi^3} \int d^3\mathbf{k} \, v_x^2 \, \frac{\partial f_0}{\partial E} \tau. \tag{5.42}$$

Noting that the total electron number N_T in the volume Ω is

$$N_T = n\Omega = \frac{1}{4\pi^3} \int f d^3\mathbf{k} \int_{\Omega} d^3\mathbf{r} = \frac{\Omega}{4\pi^3} \int f_0 d^3k, \tag{5.43}$$

we get

$$J_x = -e^2 \mathcal{E} n \frac{\int d^3\mathbf{k} \, v_x^2 \, (\partial f_0/\partial E) \, \tau}{\int d^3\mathbf{k} \, f_0}. \tag{5.44}$$

The current J_x is proportional to the electric field $\boldsymbol{\mathcal{E}}$, and we can deduce the electric conductivity (cf. Eq. (5.12)) or the mobility μ. The expression for μ can be found from Eq. (5.44), and is positive by definition.

As we have seen in Chap. 4, the Fermi level is usually sufficiently far from the the conduction and valence bands that we can approximate the distribution function f_0 by a Maxwell–Boltzmann law:

$$f_0 = \text{const. } \exp(-E/kT), \tag{5.45}$$

$$\frac{\partial f_0}{\partial E} = -\frac{1}{kT} f_0. \tag{5.46}$$

The **semiconductor** is thus **non-degenerate**. The mobility μ can then be written, using Eq. (5.44), as

$$\mu = +J_x/ne\mathcal{E} = \frac{e}{kT} \frac{\int d^3\mathbf{k}\, v_x^2\, \tau(E) f_0}{\int d^3\mathbf{k} f_0}. \tag{5.47}$$

If the effective mass is a scalar, i.e.,

$$E = \frac{\hbar^2 k^2}{2m_e} = \frac{1}{2} m_e \mathbf{v}^2, \tag{5.48}$$

we have the relations

$$\int v_x^2 \phi(E) d^3\mathbf{k} = \int v_y^2 \phi(E) d^3\mathbf{k} = \int v_z^2 \phi(E) d^3\mathbf{k} = \frac{1}{3} \int \mathbf{v}^2 \phi(E) d^3\mathbf{k}$$
$$= \frac{2}{3m_e} \int E\phi(E) d^3\mathbf{k}. \tag{5.49}$$

Further $d^3\mathbf{k} = \text{const. } E^{1/2} dE$, so that Eq. (5.47) can be written

$$\mu = \frac{2}{3m_e} \frac{1}{kT} \frac{\int_0^\infty E^{3/2} \tau(E) f_0(E)\, dE}{\int_0^\infty E^{1/2} f_0(E)\, dE}. \tag{5.50}$$

Noting that

$$\int_0^\infty E^{3/2} e^{-E/kT} dE = \frac{3kT}{2} \int_0^\infty E^{1/2} \exp(-E/kT)\, dE \tag{5.51}$$

we finally get

$$J_x = ne^2 \mathcal{E} \frac{< \tau >}{m_e} = \sigma \mathcal{E} \tag{5.52}$$

with

$$< \tau >= \frac{\int_0^\infty \tau(E) E^{3/2} \exp(-E/kT)\, dE}{\int_0^\infty E^{3/2} \exp(-E/kT)\, dE}. \tag{5.53}$$

We see that the conductivity depends on the effective mass and the mean relaxation time defined by Eq. (5.53).

For electrons the mobility is given by

$$\mu_e = e \frac{< \tau_e >}{m_e} \tag{5.54}$$

and similarly for the holes by

$$\mu_h = e\frac{<\tau_h>}{m_h}. \tag{5.55}$$

Expressions (5.54) and (5.55) for the mobilities resemble those of the Drude model (definitions (5.9) and (5.15)). However the collision times are averages which use the non-trivial weighting function $E^{3/2}\exp(-E/kT)\mathrm{d}E$. Starting from the observed temperature dependence of these averages, we can deduce $\tau(E)$, and thus identify the microscopic collision mechanisms. This will be done in Sect. 5.4, where we shall also give the orders of magnitude of the collision times.

The Boltzmann equation method can be extended to the calculation of magnetic effects like the Hall effect, magnetoresistance (variation of the conductivity because of the bending of the current lines by the Lorentz force), or thermoelectric effects. One finds that the important quantities involve not just $<\tau>$ but also $<\tau^{-1}>, <\tau^2> / <\tau>^2, \ldots$, where the mean is defined as in Eq. (5.53), i.e., using the weighting function $E^{3/2}\exp(-E/kT)\mathrm{d}E$. Some phenomena such as magnetoresistance for a semiconductor with a single carrier disappear if the relaxation time is independent of the energy.

For the case of a **degenerate semiconductor**, for which f_0 is a Fermi–Dirac distribution function, which cannot be approximated as Eq. (5.45), we have, to the same order in the field as before

$$\int_0^\infty dE\ \tau(E)\frac{\partial f_0}{\partial E}E^{3/2} = -\tau(E_F)E_F^{3/2},$$

$$\int_0^\infty dE\ E^{1/2}f_0 = \frac{2}{3}E_F^{3/2} \tag{5.56}$$

so that

$$J_x = ne^2\mathcal{E}\frac{\tau(E_F)}{m_e}. \tag{5.57}$$

The conductivity and the mobility thus involve the collision time of electrons with the Fermi energy, which itself lies in an allowed band in this case. These results are intuitively understandable: from the Pauli principle, motion of carriers requires accessible quantum states, which can only be found near E_F. (See Sect. 4.8.)

5.3c Diffusion

Equation (5.38) gives the form of the distribution function f for the case of a small concentration gradient: f thus differs little from f_0, and depends on $\mathbf{r}$. In the absence of external forces

$$f = f_0 - \tau\hbar^{-1}\nabla_{\mathbf{r}} f \cdot \nabla_{\mathbf{k}} E_{\mathbf{k}} \,, \tag{5.58}$$

where f_0 is constant in space, and the velocity has been replaced by its expression (2.29) as a function of $E_{\mathbf{k}}$.

We will calculate the particle current density $\mathbf{J}_N$. The current corresponding to a state $\mathbf{k}$ is $2\mathbf{v}(\mathbf{k})$ (2 for the spin); taking into account the density of states $(1/2\pi)^3$ the total current density is

$$\mathbf{J}_N = \frac{1}{4\pi^3}\int f\mathbf{v}(\mathbf{k})d^3\mathbf{k}, \tag{5.59}$$

$$\mathbf{J}_N = \frac{1}{4\pi^3}\int f_0\mathbf{v}(\mathbf{k})d^3\mathbf{k} - \frac{1}{4\pi^3}\int \mathbf{v}(\mathbf{k})\tau v_x\frac{\partial f}{\partial x}d^3\mathbf{k}. \tag{5.60}$$

The first term vanishes as it involves a symmetrical integral of an odd function: f_0 depends on $\mathbf{k}$ only through the energy, which is an even function of $\mathbf{k}$. In any case, $\mathbf{J}$ must vanish in equilibrium. Similarly, in the second term only the velocity component v_x enters, and the (diffusion) current is in the x direction. It thus has the form

$$J_{N,x} = -\frac{1}{4\pi^3}\int v_x^2\tau\frac{\partial f}{\partial x}d^3\mathbf{k}. \tag{5.61}$$

In the case considered, where there is a concentration gradient along x allowing a **local** equilibrium, the phase-space density can be written as

$$f = A(x)\exp(-E/kT) \tag{5.62}$$

so that the density at the point $\mathbf{r}$ is

$$n(x) = \frac{A(x)}{4\pi^3}\int \exp(-E/kT)d^3\mathbf{k}. \tag{5.63}$$

Comparing Eqs. (5.58) and (5.62), we see that $\tau v_x(\partial A/\partial x)/A(x)$ must be small compared with unity: the assumption of a small concentration gradient is equivalent to taking $A(x)$ as slowly varying over a mean free path τv_x.

The current becomes

$$J_{Nx} = -\frac{1}{4\pi^3}\frac{\partial A}{\partial x}\int v_x^2\tau\exp(-E/kT)d^3\mathbf{k}. \tag{5.64}$$

From Eq. (5.63),

$$\frac{1}{4\pi^3}\frac{\partial A(x)}{\partial x} = \frac{\partial n/\partial x}{\int\exp(-E/kT)d^3\mathbf{k}}. \tag{5.65}$$

We thus have

$$J_{Nx} = -\frac{\partial n}{\partial x} \frac{\int v_x^2 \tau \exp(-E/kT) d^3\mathbf{k}}{\int \exp(-E/kT) d^3\mathbf{k}} \tag{5.66}$$

which defines the diffusion coefficient in Fick's law:

$$D = \frac{\int v_x^2 \tau(E) \exp(-E/kT) d^3\mathbf{k}}{\int \exp(-E/kT) d^3\mathbf{k}}. \tag{5.67}$$

Comparing with expression (5.45) for the mobility we deduce the Einstein relation for electrons:

$$D = \frac{kT}{e}\mu. \tag{5.68}$$

We have thus generalized Fick's law and the Einstein relation, which we had previously only demonstrated in the context of the ultra-simplified Drude model. Obviously we get a similar result for holes. All that enters in the relation for D is the hole mobility.

For a **degenerate** system we find

$$D = \frac{\tau(E_F) v_F^2}{3} \tag{5.69}$$

so that

$$D = \frac{2}{3}\frac{E_F}{e}\mu. \tag{5.70}$$

We note that the existence of a relation between the mobility and the diffusion coefficient is extremely general: it shows that in fact there exists only a single transport coefficient relating the current of charged particles, the concentration gradient and the electrostatic potential gradient. This is a general property of transport of independent particles, whose proof is beyond the scope of the present treatment. The two gradients are the two components of the gradient of the electrochemical potential.

5.3d Summary

The semiclassical Boltzmann model allows a precise interpretation of the relaxation time appearing in the expressions for the mobility and diffusion coefficient in the Drude model (expressions (5.9) and (5.28)). What appears is a mean $< \tau >$ of the collision times $\tau(E)$, defined by

$$< \tau >= \frac{\int_0^\infty \tau(E) E^{3/2} \exp(-E/kT)\ dE}{\int_0^\infty E^{3/2} \exp(-E/kT)\ dE}. \tag{5.53}$$

We should remember that other averages over $\tau(E)$ appear for other properties such as the Hall effect and magnetoresistance, and the Drude model cannot be straightforwardly applied.

We shall now discuss several of the collision mechanisms determining $< \tau >$, and hence μ, in semiconductors.

5.4 The Mobility of Semiconductors

5.4a Collision Mechanisms

Any disruption of the crystal periodic structure is a source of collisions. In a given crystal there may be various types of collisions with characteristic times $\tau_1, \tau_2 \ldots$. We thus have to calculate an equivalent time as

$$\frac{1}{\tau} = \frac{1}{\tau_1} + \frac{1}{\tau_2} + \ldots, \tag{5.71}$$

since the probabilities of collisions caused by different mechanisms add if the events have small probability. For metals this is called Matthiesen's rule.

The main collisions mechanisms are:

(α) Scattering by **crystal vibrations (phonons)**. This is the main source of collisions at intermediate and high temperatures. The amplitude of the vibrations increases with temperature and we expect the collision probability $1/\tau$ to increase with T. In fact one can show that

$$\tau = aE^{-1/2}T^{-1} \tag{5.72}$$

implying mobilities which decrease as the temperature rises. For a non-degenerate semiconductor we get

$$\mu = \mu_0 \left(\frac{T_0}{T}\right)^{3/2}. \tag{5.73}$$

Exercise: Show that Eqs. (5.53), (5.54), and (5.72) give a law of the form (5.73).

(β) Collisions with **ionized impurities** — the effect of the Coulomb field of the impurity. As mentioned above, a rapid particle feels less the potential of an impurity, so $1/\tau$ decreases as E grows. One can show that

$$\tau = aE^{3/2} \tag{5.74}$$

which, via Eqs. (5.53) and (5.54), leads to mobilities varying as $T^{3/2}$. The mobility increases with temperature because electrons with larger velocities are less sensitive to the Coulomb fields of the ionized impurities.

Exercise: Show that Eq. (5.74) implies mobilities varying as $T^{3/2}$.

(γ) Neutral impurities and dislocations can also contribute to collisions.

Figure 5.4 shows the dependence of mobility on temperature for variously doped samples of silicon. The law $T^{-3/2}$ (Eq. (5.73)) does not hold

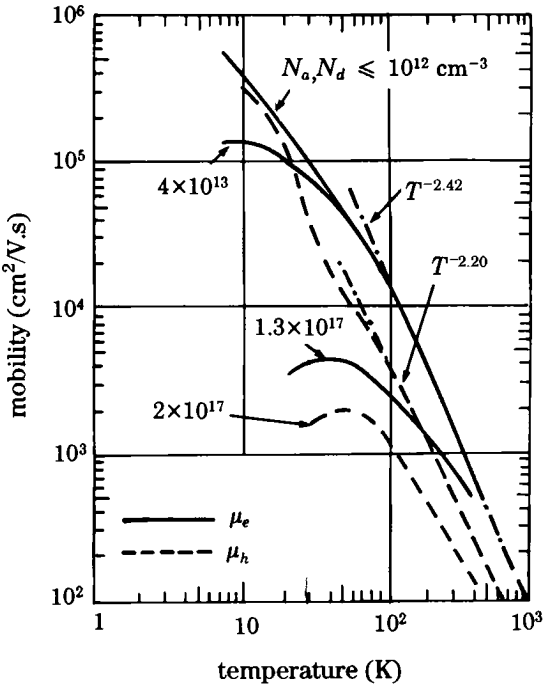


Fig. 5.4. Temperature dependence of the electron and hole mobilities μ_e, μ_h for silicon samples with different doping levels. The electron mobility is shown as continuous curves and the hole mobility as dashed. The dash–dot curves are the best fits to the experimental results.

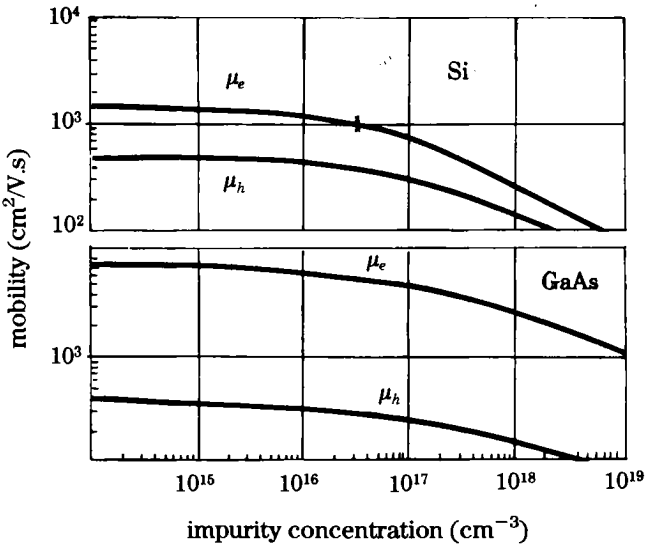


Fig. 5.5. Electron and hole mobility in silicon and gallium arsenide at room temperature, as a function of the impurity concentration.

very well. At low temperature crystal vibrations have very small amplitudes ($\tau_{\text{vibration}}$ increases) and collisions on impurities become dominant (compare Eqs. (5.72) and (5.74)). The larger the impurity concentration, the higher the temperature at which they control the mobility, the weaker the maximum mobility.

At high temperatures, on the other hand, mobilities depend little on the concentration, as shown by the measurements at 300 K given in Fig. 5.5. Using mobility data and effective masses, found, e.g., by cyclotron resonance, we can deduce collision times. We note that the variation of mobility at high temperature is much smaller than the concentration variation caused by intrinsic ionization. This is why the conductivity increases very strongly at high temperature despite the decrease of mobility.

Exercise: Use the data of Fig. 5.5 and the effective masses of Sect. 2.4 to show that the collision time is of order 10^{-13} *s in Si and GaAs, both for electrons and holes. Assume that the mobility of the conduction electrons of silicon is essentially determined by the transverse effective mass.*

The Einstein relation gives the orders of magnitude of the diffusion coefficients. We have $D = \mu(kT/e)$. At room temperature kT/e is 25meV; if $\mu = 0.1$ m^2/V.s, the diffusion coefficient is $25 \cdot 10^{-4}$ m^2/s.

5.4b Selective Doping of Superlattices ✕

We shall see at the end of this book that the operation speed of certain semiconductor devices improves with mobility. One might think to operate at low temperature, noting the data of Sect. 5.4a, but we are limited by collisions with ionized impurities. One might try to purify the semiconductors, but the shallow impurities cannot be removed without also removing the conduction electrons. The ideal would be to have electrons without impurities. What appeared to be fantasy has been realized following an idea by Störmer (1980).

We saw in Sect. 3.3 that present modern growth techniques by molecular beam epitaxy allow us to construct superlattices in composition. Let us assume that some donor atoms are added during the controlled growth, but only in the semiconductor with the largest gap, $Al_x\ Ga_{1-x}\ As$ (Fig. 5.6). Then the ground state for the electrons is not localized at the donors, in the $Al_x\ Ga_{1-x}\ As$ region. The electrons fill the free states in the GaAs quantum wells. But these are states of conduction parallel to the GaAs layers, with very little penetration into the large-gap material containing the impurities. We can say that under these conditions the electrons are localized in zones where the impurities are absent and so their mobility should be increased. This is indeed observed. The present record (1993) is $\mu_e = 1.5 \times 10^2$ m$^2 \cdot$ V$^{-1} \cdot$ s^{-1} at low temperature, while the best mobility measured in a single crystal is of the order of 10 m^2.V^{-1}. s^{-1}. These are huge values. For comparison we recall that the mobility of the best room temperature conductor, metallic silver, is of the order of 10^{-1} m$^2 \cdot$ V$^{-1} \cdot$ s^{-1}.

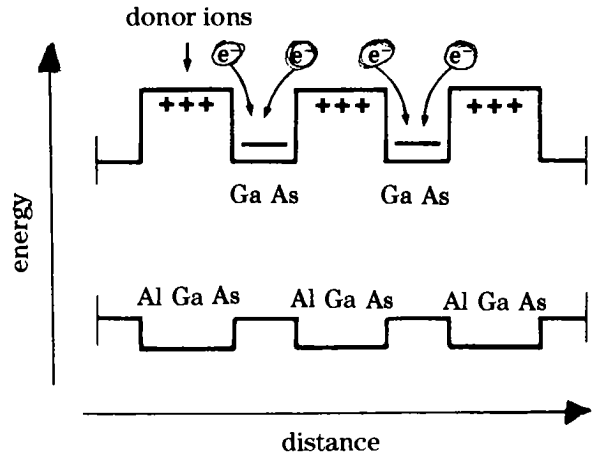


Fig. 5.6. Selective doping of a superlattice.

In the GaAs layer the collision time is increased by the same factor as the mobility, and the mean free path reaches a fraction of a micrometer. If one can manufacture sufficiently small structures, one has electrons that suffer no collisions in the plane of the layer. Such electrons are called "ballistic." The mobility is not increased in the direction perpendicular to the layers, so the mean free path remains of order 10 nm, precisely of the order of the thickness of the GaAs layer constituting the quantum well. Devices are in development (ballistic electron transistors) which use this property.

Appendix 5.1

Problems on the Hall Effect and Magnetoresistance of Semiconductors in the Drude Model

In these problems we study the conductivity originating from carriers which may have different charges (negative electrons or positive holes), in the presence of a magnetic field. We consider the transport equation determining the *mean* velocity $\mathbf{v}$ of a given type of carrier in the simplest form of the Drude model:

$$\frac{d\mathbf{v}}{dt} + \frac{\mathbf{v}}{\tau} = \frac{\mathbf{F}}{m^*}. \tag{5.75}$$

Thus in a steady state (which we assume throughout these problems)

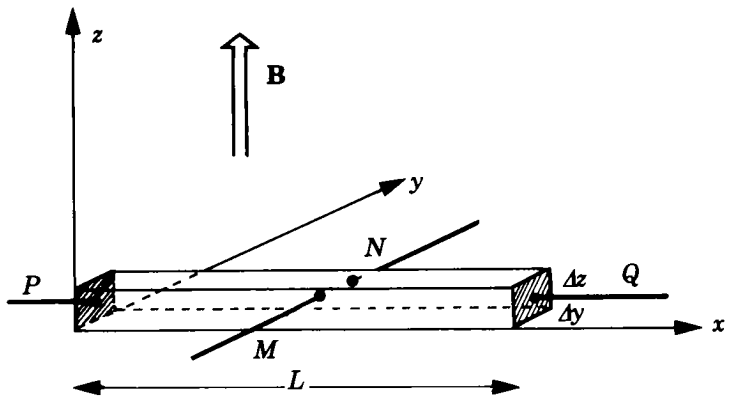
$$\mathbf{v} = \frac{\tau}{m^*}\mathbf{F}, \tag{5.76}$$

where τ and m^* are positive scalars (independent of the magnitude and direction of $\mathbf{v}$) and $\mathbf{F}$ is the external force applied to the carriers. This formulation, although extremely simple, does not alter the essential physics of the problems.

First Part

(**1**) We consider a crystal with only one kind of carrier, of charge q ($= \pm 1.6 \times 10^{-19}$ C) and concentration n. We apply a fixed electric field $\boldsymbol{\mathcal{E}}$. Derive the relations between $\mathbf{v}$ (mean velocity), $\mathbf{J}$ (current density), and $\boldsymbol{\mathcal{E}}$ in terms of the mobility μ and the conductivity σ. Pay particular attention to signs (by convention the mobility μ is a positive quantity).
(**2**) We now apply a fixed magnetic field along Oz ($B_x = 0, B_y = 0, B_z = B$). Show that Ohm's law generalizes to $\boldsymbol{\mathcal{E}} = \bar{\bar{\rho}}\mathbf{J}$, or equivalently $\mathbf{J} = \bar{\bar{\sigma}}\boldsymbol{\mathcal{E}}$, and show explicitly the resistivity and conductivity tensors $\bar{\bar{\rho}}, \bar{\bar{\sigma}}$ referred to axes $Oxyz$. In the literature the angle θ, with $|\tan\theta| = \mu B$, is called the "Hall angle." Does this suggest to you a geometrical interpretation?

(**3**) Slender-bar geometry: We consider a bar of length L along Ox with small cross sections Δy, Δz, such that the current density $\mathbf{J}$ lies along Ox. (a) Give the relation between J_x and $\mathcal{E}_x$ (longitudinal conductivity). (b) Although the current is purely longitudinal, we note that a transverse electric field (the "Hall field") appears; give an expression for it.
(**4**) Practical application: in a silicon bar with dimensions $L = 2$ cm, $\Delta z = 0.2$ cm, $\Delta y = 0.2$ cm, with magnetic field $B = 0.1$ tesla, we pass a total current of $I_x = 10$ mA. We measure a voltage $V_Q - V_P = 4.15$ V between the ends of the bar, and a voltage $V_N - V_M = 0.21 \times 10^{-3}$ V across the faces MN of the bar. Calculate the characteristics of the silicon sample.



Remark: for a voltmeter to perturb as little as possible the potential difference it measures, its internal resistance must be large. In practice every voltmeter will have various sensitivities; by construction, its internal resistance will be smaller for greater sensitivity. Thus a voltmeter with an internal resistance of 1000 Ω in the range 0–1 V will only have internal resistance 10 Ω in the range 0–1 mV. But the cost of a voltmeter increases as its internal resistance per volt. Thus a cheap ($\simeq$ \$50) modern transportable voltmeter has an internal resistance of 10 kΩ/V, a good instrument with 10 MΩ/V costs around \$180, while for greater internal resistances one has to use an electronic voltmeter ($\simeq$ \$1000).

What types of voltmeters are required in practice for these measurements? (The contacts M and N are welds of area about 0.1 cm × 0.1 cm.)

Second Part

(**1**) We assume now that there are two types of carriers (1) and (2), with charges, densities, and mobilities q_1, n_1, μ_1 and q_2, n_2, μ_2, respectively (for example, two types of electrons or electrons and holes). In all cases $|q_1| = |q_2| = e$ = electron charge. Write down the conductivity tensor.
(**2**) Consider again the slender bar. Calculate the longitudinal conductivity $\sigma_{\text{long}} = J_x/\mathcal{E}_x$. Calculate the Hall constant $R_H = \mathcal{E}_y/J_x B$.

Note: the exact calculation is laborious. Use the fact that in general $\mu_1 B$ and $\mu_2 B$ are smaller than 1 and calculate to second order in these quantities. Do this explicitly for the case of a semimetal or intrinsic semiconductor ($q_1 = -q_2$ and $n_1 = -n_2$).
(**3**) Discuss what happens at the surface of the sample in the case of a single carrier type, and in the case of equal electron and hole densities.

Solutions

First Part

(**1**) Under an applied field $\boldsymbol{\mathcal{E}}$ the force is $\mathbf{F} = q\boldsymbol{\mathcal{E}}$ and the mean velocity is

$$\mathbf{v} = \frac{q\tau}{m^*}\boldsymbol{\mathcal{E}}. \tag{5.77}$$

The mobility μ is defined in Eqs. (5.9) and (5.15) as the ratio of $\mathbf{v}$ and $\boldsymbol{\mathcal{E}}$. As indicated in the problem this quantity is conventionally taken as a positive quantity. We thus have

$$\mu = \frac{|q|\tau}{m^*} \quad \text{and} \quad \mathbf{v} = \frac{|q|}{q}\mu\boldsymbol{\mathcal{E}}. \tag{5.78}$$

The current density $\mathbf{J}$ is given by $\mathbf{J} = nq\mathbf{v}$, therefore $\mathbf{J} = n|q|\mu\boldsymbol{\mathcal{E}}$. We recover Ohm's law

$$\mathbf{J} = \sigma\boldsymbol{\mathcal{E}}, \tag{5.79}$$

where the conductivity $\sigma = n|q|\mu = nq^2\tau/m^*$ is positive whatever the sign of the charge and whatever the convention adopted.
(**2**) When both an electric and magnetic field are applied, the force $\mathbf{F}$ becomes $\mathbf{F} = q(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \mathbf{B})$, and the velocity $\mathbf{v}$ is given by the equation

$$\mathbf{v} = \frac{q\tau}{m^*}(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \mathbf{B}). \tag{5.80}$$

Using the relations $\mathbf{J} = nq\mathbf{v}, \mu = |q|\tau/m^*$, and $\sigma = n|q|\mu$, we find

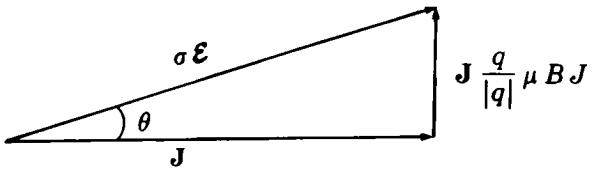
$$\boldsymbol{\mathcal{E}} = \frac{1}{\sigma}\left(\mathbf{J} + \frac{|q|}{q}\mu\mathbf{B} \times \mathbf{J}\right). \tag{5.81}$$

This formula can be written as $\mathcal{E} = \bar{\bar{\rho}}\mathbf{J}$, where the resistivity tensor referred to axes xyz is

$$\bar{\bar{\rho}} = \frac{1}{\sigma}\begin{vmatrix} 1 & -\frac{|q|}{q}\mu B & 0 \\ \frac{|q|}{q}\mu B & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}. \tag{5.82}$$

We note that in the plane xOy the vectors $\mathcal{E}$ and $\mathbf{J}$ are not collinear: as shown in the figure below, illustrating the vector relation (5.81), the angle θ between $\mathbf{J}$ and $\mathcal{E}$ is given in sign and magnitude by

$$\tan\theta = (q/|q|)\mu B$$



The conductivity tensor $\bar{\bar{\sigma}}$ is given by simply inverting the tensor $\bar{\bar{\rho}}$:

$$\bar{\bar{\sigma}} = \sigma\begin{vmatrix} \frac{1}{1+\mu^2B^2} & \frac{q}{|q|}\frac{\mu B}{1+\mu^2B^2} & 0 \\ -\frac{q}{|q|}\frac{\mu B}{1+\mu^2B^2} & \frac{1}{1+\mu^2B^2} & 0 \\ 0 & 0 & 1 \end{vmatrix}. \tag{5.83}$$

(**3**) In the slender-bar geometry we assume $J_y = J_z = 0$. Using the vector relation (5.81) or the matrix (5.82) we find

$$J_x = \sigma\mathcal{E}_x. \tag{5.84}$$

This is the same relation as in the absence of a magnetic field. Conduction is not changed by the presence of a magnetic field for a single carrier type (in the simple transport model we have adopted): there is no "magnetoresistance" effect. However we see that a transverse electric field appears, which from the definitions of σ and μ takes the form

$$\mathcal{E}_y = \frac{J_xB}{nq}. \tag{5.85}$$

Measurement of this "Hall field" $\mathcal{E}_y$ thus allows us to measure the carrier density n and sign. Knowing n, a measurement of $\sigma = n|q|\mu$ gives the mobility of the carriers.

(**4**) Application. From the sign of the measured voltages we see that the carriers are positively charged holes. The bar is therefore made of p-type

silicon. The numerical application of the above formulas gives a Hall angle $\mu B \sim 5 \times 10^{-4}$ rad and $\mu = 0.05\ \mathrm{m^2\ V^{-1}.\ s^{-1}}$, $\sigma = 12$ siemens.$\mathrm{m^{-1}}$, $n = 1.5 \times 10^{21}\ \mathrm{m^{-3}}$. This is a typical material for the manufacture of transistors.

Remark: the voltmeters used must not disturb the measurement. For the longitudinal voltage this is relatively straightforward. We have to measure a voltage of order 5 V with a total current of 10 mA: an instrument with about 20,000 Ω/V will divert about 0.05 mA and thus perturb the experiment by less than 0.5%.

By contrast the measurement of the transverse voltage is much more delicate: the internal resistance between the contacts M and N, of size approximately 0.1 cm × 0.1 cm, is

$$R_i = \frac{1}{\sigma} \times \frac{L}{S} = \frac{2 \times 10^{-3}}{12 \times 10^{-6}} \simeq 160\ \Omega.$$

For an error less than 2%, the internal impedance of the voltmeter must be at least $50R_i$, i.e., 8 kΩ for 0.2 mV. This corresponds to a sensitivity of about 40 MΩ/V. We therefore need an electronic voltmeter for this measurement.

Second Part

(1) The **currents** of the two carrier types add:

$$\mathbf{J} = \mathbf{J}_1 + \mathbf{J}_2 = (\bar{\bar{\sigma}}_1 + \bar{\bar{\sigma}}_2)\boldsymbol{\mathcal{E}}. \tag{5.86}$$

We have

$$\bar{\bar{\sigma}} = \bar{\bar{\sigma}}_1 + \bar{\bar{\sigma}}_2.$$

We must first add the conductivity tensors (matrices of type (5.83)) to obtain $\bar{\bar{\sigma}}$, then invert this matrix to find ρ_{xx} and ρ_{xy}.

Setting $q_1\mu_1 B/|q_1|\mu_1 B = \theta_1$ and $q_2\mu_2 B/|q_2|B = \theta_2$, both assumed small, and expanding to second order in θ_1 and θ_2 we get

$$\bar{\bar{\sigma}} = \begin{vmatrix} \sigma_1(1-\theta_1^2) + \sigma_2(1-\theta_2^2) & \sigma_1\theta_1 + \sigma_2\theta_2 & 0 \\ -(\sigma_1\theta_1 + \sigma_2\theta_2) & \sigma_1(1-\theta_1^2) + \sigma_2(1-\theta_2^2) & 0 \\ 0 & 0 & \sigma_1 + \sigma_2 \end{vmatrix}. \tag{5.87}$$

(2) As $J_y = J_z = 0$, the required quantities follow immediately from the tensor $\bar{\bar{\rho}}$:

$$\mathcal{E}_x = \rho_{xx} J_x \quad \text{and} \quad \mathcal{E}_y = \rho_{yx} J_x.$$

Using the fact that $\sigma_{xx} = \sigma_{yy}$ and $\sigma_{xy} = -\sigma_{yx}$, we find

$$\rho_{xx} = \frac{\sigma_{xx}}{(\sigma_{xx})^2 + (\sigma_{xy})^2} \quad \text{and} \quad \rho_{yx} = \frac{\sigma_{yx}}{(\sigma_{xx})^2 + (\sigma_{xy})^2}. \tag{5.88}$$

Expanding to second order this gives

$$\sigma_{\text{long}} = \frac{J_x}{\mathcal{E}_x} = \sigma_1 + \sigma_2 - (\sigma_1\theta_1^2 + \sigma_2\theta_2^2) + \frac{(\sigma_1\theta_1 + \sigma_2\theta_2)^2}{\sigma_1 + \sigma_2}.$$

After a little algebra this can be written

$$\sigma_{\text{long}} = \sigma_1 + \sigma_2 - \frac{\sigma_1\sigma_2}{\sigma_1 + \sigma_2}(\theta_1 - \theta_2)^2$$

or with $\sigma_1 = |q|n_1\mu_1$ and $\sigma_2 = |q|n_2\mu_2$,

$$\sigma_{\text{long}} = |q|(n_1\mu_1 + n_2\mu_2)\left[1 - \frac{n_1\mu_1 n_2\mu_2(q_1\mu_1 - q_2\mu_2)^2 B^2}{e^2(n_1\mu_1 + n_2\mu_2)^2}\right]. \tag{5.89}$$

We see that in this case the longitudinal conductivity depends to second order on the magnetic field unless the carriers have the same sign ($q_1 = q_2$) and the same mobility ($\mu_1 = \mu_2$). This result, the variation of resistance with magnetic field, or magnetoresistance, is generic as soon as we are not dealing with a single carrier type with a single relaxation time.

In the same way, the Hall constant is given (still to second order) by

$$\frac{\mathcal{E}_y}{J_x} = B \cdot R_H = \rho_{yx} = \frac{\sigma_1\theta_1 + \sigma_2\theta_2}{(\sigma_1 + \sigma_2)^2}, \tag{5.90}$$

$$R_H = \frac{q_1 n_1 \mu_1^2 + q_2 n_2 \mu_2^2}{q^2(n_1\mu_1 + n_2\mu_2)^2}. \tag{5.91}$$

In the domain of validity of these formulas we see that:

- The magnetoresistance is always positive (resistance **increases** with magnetic field).
- We can measure the resistance for $B = 0$, the magnetoresistance (which under our assumptions is a very small effect), and the Hall constant R_H. From these three independent measurements it is in general impossible to determine the four unknowns (n_1, μ_1, n_2, μ_2) even if it is possible to guess the sign of the charges.
- If we have reason to believe that $n_1 = n_2 = n$, for opposite charge carriers (semimetal or intrinsic semiconductor), we can write

$$\sigma_{\text{long}} = |q|n(\mu_1 + \mu_2)(1 - \mu_1\mu_2 B^2), \tag{5.92}$$

$$R_H = \frac{\mu_1^2 - \mu_2^2}{(\mu_1 + \mu_2)^2}\frac{1}{nq}. \tag{5.93}$$

We find that for the Hall effect it is the carriers with the greater mobility that dominate. If $\mu_1 = \mu_2$ the Hall effect disappears (but conductivity and magnetoresistance remain).

For the more realistic case of a distribution of relaxation times we must generalize the above treatment by using the Boltzmann equation. We expect the Hall effect to be slightly modified: there appears in Eq. (5.85) a correction factor of the form $< \tau^2 > / < \tau >^2$. Even for a single carrier type, a distribution of relaxation times creates magnetoresistance. This magnetoresistance must be calculated using the Boltzmann equation.
(**3**) With a single type of carriers we have $J_y = 0$; the electric field $\mathcal{E}_y$ which compensates the Lorentz force is caused by the charges which accumulate on the faces M and N once the magnetic field is switched on.

When there are two carrier types we have $J_y = J_{1y} + J_{2y} = 0$, but J_{1y} and J_{2y} are not separately zero: there are transverse currents of both carrier types which cancel. It is clear that charges cannot accumulate indefinitely on the faces, and we have a stationary regime where the two types of charges reaching the surface (e.g., electrons and holes) recombine in pairs. This recombination may give rise to measureable effects, such as light emission at the energy of the band gap.

6.

Effects of Light

In Chap. 5 we studied the transport properties of electron and hole gases regarded as independent of each other. We will now study the dynamical equilibrium between these two gases, particularly in the presence of light. This problem is important in practice, since many devices such as detecting and emitting diodes use the electro-optical properties of semiconductors.

Two reactions can occur within a semiconductor subject to intrinsic luminous irradiation (with energy greater than the band gap): the first,

$$\text{light} \rightarrow \text{electron} + \text{hole} \tag{6.1}$$

is that of **optical absorption**, which we shall discuss first. The second reaction,

$$\text{electron} + \text{hole} \rightarrow \text{energy} \tag{6.2}$$

is the process called **recombination**. If it results in the emission of light, it is called radiative recombination. The annihilation of an electron–hole pair can also occur via non-radiative processes. We shall study recombination in Sect. 6.2.

In Sect. 6.3 we will explain the working of several common optoelectronic devices that use homogeneous semiconductors: photoelectric cells, photocopiers, and television screens.

6.1 Light Absorption by Semiconductors

When a semiconductor is irradiated by light, electrons can be excited from the valence band into the conduction band by the absorption of photons, provided the photon energy is greater than E_g. Here we shall calculate this absorption in the simplest case, that of so-called "direct" absorption.

We know that a transverse electromagnetic wave propagating in a medium has a corresponding vector potential:

$$\mathbf{A}(t) = A\ \mathbf{a}_0 \cos(\omega t - \mathbf{K} \cdot \mathbf{r}), \tag{6.3}$$

where $\mathbf{a}_0$ is a unit vector orthogonal to $\mathbf{K}$, the wave vector of the light. In the following we shall be interested in the absorption of light by a small sample of area s and small thickness $\mathrm{d}x$ normal to the luminous flux. Under these conditions the wave is weakly attenuated and its constant amplitude A can be chosen as real by suitable choice of the origins of space and time. The wave's electric and magnetic fields have the forms

$$\boldsymbol{\mathcal{E}} = -\frac{\partial \mathbf{A}}{\partial t} = A\,\omega\,\mathbf{a}_0 \sin(\mathbf{K}\cdot\mathbf{r} - \omega t), \tag{6.4}$$

$$\boldsymbol{\mathcal{B}} = \nabla \times \mathbf{A} = -A\ (\mathbf{K}\times\mathbf{a}_0)\sin(\mathbf{K}\cdot\mathbf{r} - \omega t). \tag{6.5}$$

The energy flux is given by the Poynting vector

$$\Pi = \boldsymbol{\mathcal{E}} \times \frac{\boldsymbol{\mathcal{B}}}{\mu_0}. \tag{6.6}$$

Averaging over a period and using the relation $\epsilon_0\mu_0 c^2 = 1$ we get the mean energy flux

$$\overline{\Pi} = \frac{1}{2}\epsilon_0\,\omega^2\,A^2\,c\,n, \tag{6.7}$$

where n is the refractive index of the medium.

We also know that in the presence of the vector potential (6.3) the Hamiltonian of an electron is

$$\mathcal{H} = \frac{1}{2m}(\mathbf{p} + e\,\mathbf{A})^2 + V(\mathbf{r}), \tag{6.8}$$

$$\mathcal{H} = \mathcal{H}_0 + \frac{e}{2m}(\mathbf{p}\cdot\mathbf{A} + \mathbf{A}\cdot\mathbf{p}) + \frac{e^2}{2m}\mathbf{A}^2. \tag{6.9}$$

The operator $\mathcal{H}_0$ is the crystal Hamiltonian, whose eigenfunctions are the Bloch states. For light of sufficiently low intensity, the last term in Eq. (6.9) can be neglected and the second term regarded as a small time-dependent perturbation of $\mathcal{H}_0$. We easily verify that

$$[\mathbf{p}, \mathbf{A}(t)] = -i\hbar\nabla\cdot\mathbf{A}(t). \tag{6.10}$$

Now with the choice (6.3), the vector potential is such that $\nabla\cdot\mathbf{A} = 0$, and $\mathbf{p}$ and $\mathbf{A}$ commute, so Eq. (6.9) can be rewritten as

$$\mathcal{H} = \mathcal{H}_0 + \frac{e}{m}\mathbf{A}\cdot\mathbf{p}. \tag{6.11}$$

We thus have a quantum mechanical system subject to a sinusoidal perturbation, because of expression (6.3) for $\mathbf{A}$. Indeed, the Hamiltonian describing the interaction of an electron with a travelling electromagnetic wave has the form

$$\mathcal{H}_p(t) = \mathcal{H}_p\cos(\omega t - \mathbf{K}\cdot\mathbf{r}) \tag{6.12}$$

or

$$\mathcal{H}_p(t) = C\exp(i\omega t) + C^+\exp(-i\omega t) \tag{6.13}$$

with

$$C^+ = \frac{eA}{2m}\,(\mathbf{a}_0\cdot\mathbf{p})\,(\exp i\mathbf{K}\cdot\mathbf{r}). \tag{6.14}$$

As we show in Appendix 6.1, the transition probability from an initial state $|i>$ to a final state $|f>$ of higher energy can be written

$$P_{i\to f}(t) = \frac{|<f|\mathcal{H}_p|i>|^2}{\hbar^2}\;\frac{\sin^2(1/2)(\omega_{if}-\omega)t}{(\omega_{if}-\omega)^2} \tag{6.15}$$

with

$$\hbar\omega_{if} = (E_f - E_i). \tag{6.16}$$

We note from this formula that **the transition probability between two states depends sinusoidally on the time**. This probability is proportional to the square of the matrix element of the perturbation. Replacing $\mathcal{H}_p$ by its expression we obtain the following expression for the absorption transition probability, which we shall use below:

$$P_{i\to f}(t) = \frac{4\,|<f|C^+|i>|^2}{\hbar^2}\;\frac{\sin^2(1/2)(\omega_{if}-\omega)t}{(\omega_{if}-\omega)^2}. \tag{6.17}$$

For the case of optical excitation of a semiconductor, the initial state $|\mathbf{k}_v>$ is in the valence band, and the final state $|\mathbf{k}_c>$ in the conduction band.

6.1a The Fermi Golden Rule

We consider the probability, for a **given** initial state $|i>$, of finding an electron in **one or another** of a set of final states of energy close to E_f. If $n(E_f)$ is the density of these states the **total** probability of finding the electron in any one of these states is

$$P(t) = \frac{4}{\hbar}\int |<f|C^+|i>|^2\, n(E_f)\frac{\sin^2(1/2)(\omega_{if}-\omega)t}{(\omega_{if}-\omega)^2}\,d\omega_{if}. \tag{6.18}$$

In this integral the only non-negligible contribution comes from the region where ω_{if} is very close to ω, because of the denominator $(\omega_{if}-\omega)^2$. The variation of the function $(\Delta\omega)^{-2}\sin^2\Delta\omega t/2$ versus $\Delta\omega = \omega_{if}-\omega$ is shown in Fig. 6.1. We see that only a narrow frequency bandwidth of order $1/t$ contributes to the integral. We can therefore extend the limits of the integral to $\pm\infty$ and use the fact that both $n(E_f) = n(E_i+\hbar\omega_{if})$ and the matrix element are approximately constant over the interval $\hbar/t$ to write

$$P(t) = \frac{4}{\hbar}| < f|C^+|i > |^2 n(E_f) \int_{-\infty}^{+\infty} \frac{\sin^2(1/2)(\omega_{if} - \omega)t}{(\omega_{if} - \omega)^2} d\omega_{if} \quad (6.19)$$

or, setting

$$x = \frac{1}{2}(\omega_{if} - \omega)t,$$

$$P(t) = \frac{1}{\hbar}| < f|C^+|i > |^2 n(E_f) 2\pi t, \quad (6.20)$$

where we have used the fact that

$$\int_{-\infty}^{+\infty} \frac{\sin^2 x}{x^2} = \pi.$$

The result (6.20) is remarkable in that while the probability of transition from a given state to another given state is a sinusoidal function of time (6.17), **the total probability of transition from a given initial state to one or another of a set of final states which are very close to each other is proportional to the time. We may therefore define a transition probability per unit time W:**

$$W = \frac{P(t)}{t} = \frac{2\pi}{\hbar}| < f|C^+|i > |^2 n(E_f). \quad (6.21)$$

This equation is called the **Fermi Golden Rule**. In the proof leading to this formula we have not used the nature of the system subject to the radiation. This formula applies in several areas of physics, whenever a system possessing nearby states is illuminated by monochromatic radiation.

The fact that the only significant contributions to the integral (6.18) come from energies such that $\omega_{if} \simeq \omega$ is an expression of conservation of energy for states which were eigenstates of the unperturbed system. This conservation is exact in the limit of weak perturbations and long time scales. The central peak of Fig. 6.1 has height $t^2/4$ and becomes sharper and sharper as its width decreases as t^{-1}. The area under the peak increases linearly in time. In this limit we can write Eq. (6.21) as

$$W = \frac{2\pi}{\hbar}| < f|C^+|i > |^2 \delta(E_f - E_i - \hbar\omega) \quad (6.22)$$

and the transition probability is obtained by integration over the final states.

6.1b Selection Rules

Using expression (6.14) for C^+ and replacing $|i>$ and $|f>$ by the corresponding Bloch states, the perturbation matrix element becomes

$$\frac{eA}{2m} \int \exp(-i\mathbf{k}_f \cdot \mathbf{r}) u^*_{c,\mathbf{k}_f}(\mathbf{r}) \exp(i\mathbf{K} \cdot \mathbf{r}) \times$$

$$(\mathbf{a}_0 \cdot \mathbf{p}) \exp(i\mathbf{k}_i \cdot \mathbf{r}) u_{v,\mathbf{k}_i}(\mathbf{r}) d^3\mathbf{r}, \quad (6.23)$$

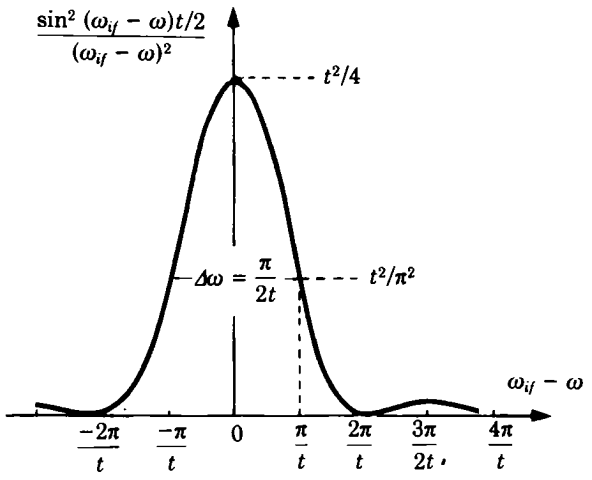


Fig. 6.1. Variation of the function $(\omega_{if}-\omega)^{-2}\sin^2(\omega_{if}-\omega)t/2$.

where $u(\mathbf{r})$ are the periodic parts of the Bloch functions, indexed v for the valence band and c for the conduction band. This integral can be regarded as the matrix element of the periodic operator $\mathbf{a}_0 \cdot \mathbf{p}$ between two functions of Bloch form, one characterized by $\mathbf{k}_f - \mathbf{K}$, the other by $\mathbf{k}_i$. This matrix element vanishes unless, from Appendix 2.1,

$$\mathbf{k}_f = \mathbf{k}_i + \mathbf{K}. \tag{6.24}$$

Now, the light wave vector is very small on the scale of the Brillouin zone, as the wavelength of light with $\hbar\omega \sim E_g$ is around 10^4 times larger than the lattice constant. We may therefore neglect $\mathbf{K}$ in Eq. (6.24) and write

$$\mathbf{k}_i = \mathbf{k}_f \tag{6.25}$$

for allowed transitions.

Transitions of this kind can therefore only occur when the electron wave vector after excitation is approximately equal to the wave vector before excitation. For this reason the transitions are called vertical or direct. They are represented in Fig. 6.2 by the arrow (a). An "oblique" transition such as (b) is forbidden for the above process.

In comparing matrix elements or estimating their order of magnitude, it is useful to remember (see Chap. 2) that the periodic parts of Bloch functions resemble atomic wave functions, and write

$$\psi_{v,\mathbf{k}_i}(\mathbf{r}) = N^{-1/2} u_{v,\mathbf{k}_i}(\mathbf{r}) \exp(\mathrm{i}\mathbf{k}_i \cdot \mathbf{r}), \tag{6.26}$$

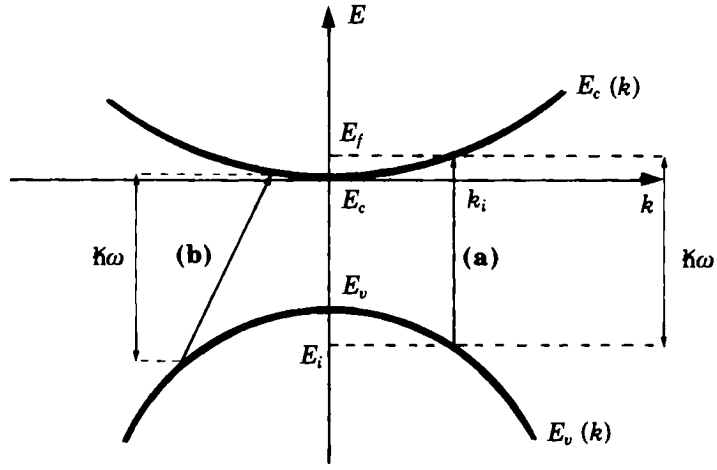


Fig. 6.2. Direct (a) and indirect (b) absorption in a semiconductor.

where N is the number of unit cells of the crystal. With this convention the periodic parts are normalized over the unit cell and have an amplitude comparable to the atomic functions. Only matrix elements with $\mathbf{k}_i = \mathbf{k}_f$ appear, i.e., matrix elements of $\mathbf{p}$ between valence and conduction functions with the form

$$\begin{aligned}&\frac{1}{N}\int_{crystal\ volume} u^*_{v,\mathbf{k}_i}(\mathbf{r})\ \mathbf{p}\ u_{c,\mathbf{k}_i}(\mathbf{r}) d^3\mathbf{r}\\ &= \int_{unit\ cell} u^*_{v,\mathbf{k}_i}(\mathbf{r})\ \mathbf{p}\ u_{c,\mathbf{k}_i}(\mathbf{r})\, d^3\mathbf{r}\\ &= <\mathbf{p}>_{\mathbf{k}_i}. \end{aligned} \tag{6.27}$$

The components of $<\mathbf{p}>_{\mathbf{k}_i}$ will have the same order of magnitude as the matrix elements of $\mathbf{p}$ between two atomic states. Finally the matrix element appearing in the transition probability will be

$$<\mathbf{k}_{f,c}|C^+|\mathbf{k}_{i,v}> = \frac{eA}{2m}\mathbf{a}_0 \cdot <\mathbf{p}>_{\mathbf{k}_i}\ \delta(\mathbf{k}_i - \mathbf{k}_f). \tag{6.28}$$

To simplify the notation we shall write

$$\mathbf{a}_0 \cdot <\mathbf{p}>_{\mathbf{k}_i} = p_{\mathbf{k}_i}. \tag{6.29}$$

6.1c Calculation of the Absorption Coefficient

The total probability W_T of photon absorption is calculated by summing Eq. (6.22) over all the initial states of the solid (valence band states) and all possible final states (conduction band states):

$$W_T = \sum_{\substack{\mathbf{k}_{i,v} \\ \mathbf{k}_{f,c}}} \frac{2\pi}{\hbar} \frac{e^2 A^2}{4m^2} p^2_{\mathbf{k}_i} \, \delta(\mathbf{k}_i - \mathbf{k}_f) \, \delta(E_f - E_i - \hbar\omega) \tag{6.30}$$

with

$$E_f - E_c = \frac{\hbar^2 k^2_{f,c}}{2m_e}, \tag{6.31}$$

$$E_i - E_v = -\frac{\hbar^2 k^2_{i,v}}{2m_h} \tag{6.32}$$

for a "standard" band structure (Fig. 6.2).

Summing over $\mathbf{k}_i$ we get

$$W_T = \frac{2\pi}{\hbar} \frac{e^2 A^2}{4m^2} \sum_{\mathbf{k}_f} p^2_{\mathbf{k}_f} \, \delta \left[\frac{\hbar^2 k_f^2}{2} \left(\frac{1}{m_e} + \frac{1}{m_h} \right) + E_g - \hbar\omega \right]. \tag{6.33}$$

We can transform the discrete summation into an integral by using the density of states in $\mathbf{k}$ space. For a sample of volume $s\,dx$ this density is $s\,dx/4\pi^3$ so that

$$W_T = \frac{2\pi}{\hbar} \frac{e^2 A^2}{4m^2} \frac{s\,dx}{4\pi^3} \int d^3\mathbf{k}_f \, p^2_{\mathbf{k}_f} \delta \left[\frac{\hbar^2 k_f^2}{2} \left(\frac{1}{m_e} + \frac{1}{m_h} \right) + E_g - \hbar\omega \right]. \tag{6.34}$$

The periodic functions u_v and u_c vary slowly with $\mathbf{k}$ (see for example the $\mathbf{k} \cdot \mathbf{p}$ method, Appendix 2.4), so we can regard $p^2_{\mathbf{k}_f}$ as independent of k and write

$$p^2_{\mathbf{k}_f} = p^2. \tag{6.35}$$

We define the reduced mass m_r by

$$m_r^{-1} = m_e^{-1} + m_h^{-1}. \tag{6.36}$$

We set $x = (\hbar^2/2m_r)\mathbf{k}_f^2 + E_g - \hbar\omega$. It remains to integrate

$$\frac{e^2 A^2 p^2}{8\pi^2 m^2 \hbar} s \, dx \cdot 2\pi \left(\frac{2m_r}{\hbar^2} \right)^{3/2} \int (\hbar\omega - E_g + x)^{1/2} \, \delta(x) \, dx. \tag{6.37}$$

We obtain

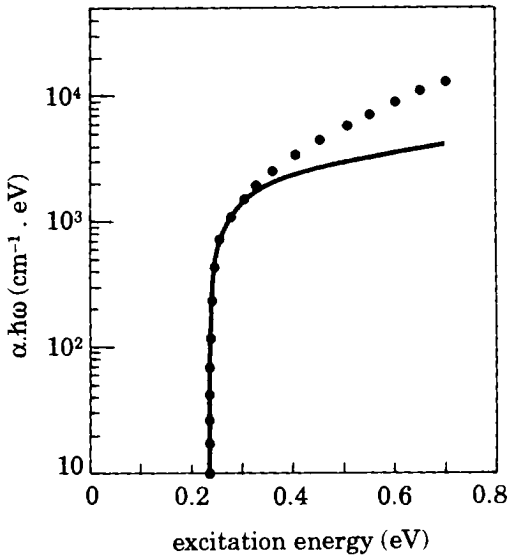


Fig. 6.3. Absorption threshold of InSb at $T = 5$ K. The measurements (points) are compared with the predictions of Eq. (6.40). The deviation at high energy is suppressed by taking account of the correct density of states, which is no longer parabolic at large $\mathbf{k}$, and of the variation of the matrix element $p_{\mathbf{k}_f}$ with $\mathbf{k}_f$.

$$W_T = \frac{s\,dx}{4\pi}\frac{e^2A^2}{\hbar^4 m^2}(2m_r)^{3/2}p^2(\hbar\omega - E_g)^{1/2}. \tag{6.38}$$

The average incident photon flux is $\overline{\Pi}/\hbar\omega$ per unit area, and by definition the fraction of incident photons absorbed over a depth dx is

$$W_T = \alpha\frac{\overline{\Pi}}{\hbar\omega}s\,dx. \tag{6.39}$$

Substituting Eqs. (6.7) and (6.38) in Eq. (6.39) we get

$$\alpha = \frac{(2m_r)^{3/2}}{2\pi\epsilon_0 cn}\frac{e^2p^2}{m^2}\frac{1}{\hbar^3\omega}(\hbar\omega - E_g)^{1/2}. \tag{6.40}$$

This establishes Eq. (2.106) for the case of direct transitions. This result was obtained by J. Bardeen, F.J. Blatt, and L.H. Hall (Atlantic City Photoconductivity Conference, 1954).

Absorption varies from one semiconductor to another through the values m_r, p and the refractive index n. An average order of magnitude is

$$\alpha(\mathrm{cm}^{-1}) = 4.10^4[(h\nu - E_g)(\mathrm{eV})]^{1/2}. \tag{6.41}$$

For $h\nu = 1.1$ eV and $E_g = 1$ eV we find $\alpha \sim 10^4$ cm^{-1}, showing that a semiconductor is an efficient absorber. It is sometimes used as a filter transparent to long wavelengths.

The experimental dependence of α $(h\nu - E_g)^{1/2}$ can be seen as proof of the existence of direct permitted transitions between band extrema. Figure 6.3 shows the results found for indium antimonide, InSb. We see that near the band gap, which is 0.22 eV for InSb, Eq. (6.40) is well satisfied. But when the initial or final states have energies in the valence or conduction bands comparable with the band gap, the above approximations are no longer sufficient.

6.1d Excitons

Figure 6.4 shows the absorption coefficient of GaAs for photon energies close to E_g ($E_g = 1.52$ eV at low temperature). We note the presence of a peak near E_g, which is not predicted by the calculation of Sect. 6.1c. This results from the creation through photon absorption of **excitons**, whose definition we now give.

The ground state of the semiconductor corresponds to a full valence band with an empty conduction band. According to what we have said, the first excited state should correspond to an electron at the bottom of the conduction band with a hole at the top of the valence band. In this state these two charges are assumed to be far away in real space. We have created a non-interacting electron-hole pair by absorbing a photon of energy $h\nu = E_g$.

In fact the electron, because of its negative charge, feels the Coulomb attraction of the positively charged hole. It is feasible that a bound electron-hole state can exist whose energy will be less than that of the dissociated electron-hole pair. To find the binding energy of this **exciton** pair we solve a "hydrogen-like" problem of two charged particles of effective masses m_e and m_h, which also involves the dielectric constant of the medium (the problem is analogous to that of the binding energy of an electron and a donor, treated in Sect. 3.2). To obtain the motion about the center of mass, we take the reduced mass (6.36), so that the binding energy is

$$E_{\text{exc}} = -\frac{m_r}{m_0}\frac{1}{\epsilon_r^2}\frac{E_1}{n^2}, \tag{6.42}$$

where $E_1 = -13.6$ eV. As $m_h > m_e$ in general we have $m_r \lesssim m_e$, and E_{exc} is of the order of the donor binding energy (a few meV for the III–V compounds of Sect. 3.5). We thus obtain a hydrogenic series of levels. To energy E_{exc} we must add the kinetic energy of the center of mass, with a mass $m_e + m_h$. The lowest optical excitation energy of the solid will thus not be $h\nu = E_g$, but $h\nu = E_g + E_{\text{exc}}$, and in some cases we get a series of peaks corresponding to different values of n.

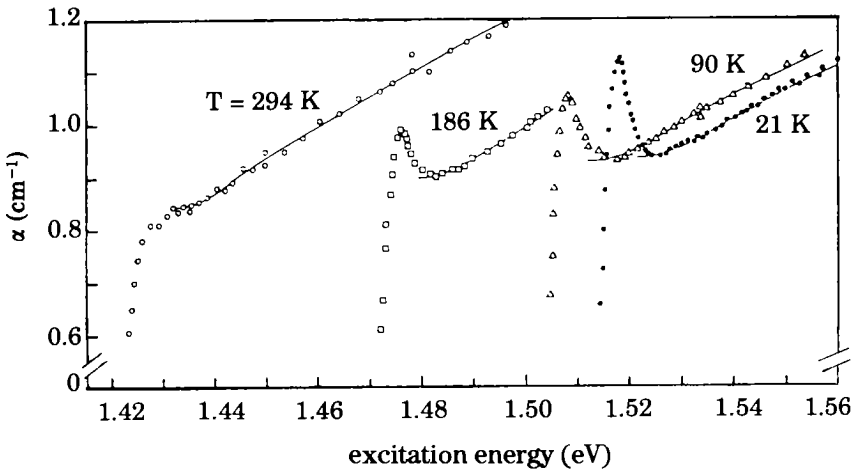


Fig. 6.4. Absorption in GaAs near threshold, at various temperatures. As the temperature increases the band gap decreases, basically because of the thermal expansion of the lattice (Appendix 2.3). The peak is caused by absorption leading to the creation of excitons and is more prominent at low temperature: if kT is of the order of the exciton binding energy, broadening of the peak will become significant.

Figure 6.5 indeed shows the exciton levels $n = 1, 2, 3$ as well as an even lower excitation energy, corresponding to the creation of an exciton bound to a neutral donor ($D^0 - X$), a system consisting of the donor nucleus, two electrons, and a hole.

It can be shown that the formation of an exciton is possible only for vertical transitions (cf. Fig. 6.2), which occur in most III–V semiconductors. From Eq. (6.42) the binding energy of the exciton is greater, and the exciton thus more readily observable, the larger m_r (and thus m_e) is, and the smaller ϵ_r is. As m_r and E_g are small in the same materials (cf. Sect. 2.4), excitons are more easily observed in semiconductors with large band gaps.

6.2 Recombination

When a semiconductor is illuminated, its properties are modified by the creation of electron-hole pairs, and we study these changes here. We expect qualitatively that as the number of electrons and holes increases the conductivity should rise: this is called photoconductivity. It is easily observable, and is currently used to measure light intensity in photoelectric cells. If a crystal is submitted to continuous irradiation the number of electron-hole pairs should steadily increase in time, and the conductivity would tend to infinity, contrary to experiment. The limiting factor is **the destruction of electron-hole pairs by recombination processes**.

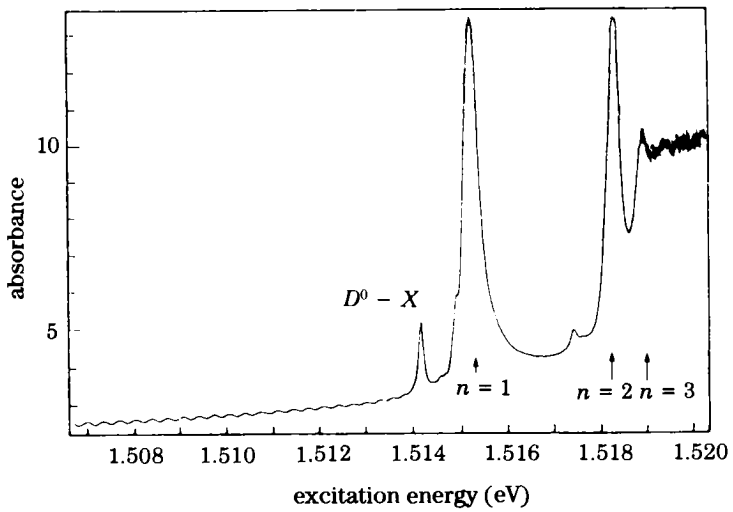


Fig. 6.5. Absorption measured at $T = 6$ K of a sample of pure GaAs containing residual donors. The absorbance is proportional to the absorption coefficient α. Note the peaks corresponding to the levels $n = 1, 2$, and 3 of the exciton, as well as a peak associated with an exciton bound to a neutral donor ($D^0 - X$).

There are other means of establishing electron and hole concentrations departing from thermal equilibrium: injecting excess carriers through a metallic contact, or bombarding the semiconductor with charged particles, which can create excess electron-hole pairs as they decelerate. The latter effect is used to detect particles. In all these cases the return to thermodynamic equilibrium once the external excitation stops must involve the disappearance of the excess electrons and holes. These recombination processes play an important role in semiconductor devices: diodes, transistors, and light-sensitive devices. When light is emitted by the semiconductor we speak of photoluminescence, electroluminescence, thermoluminescence, etc., depending on the type of excitation.

Consider a semiconductor in which we have created deviations Δn and Δp from the equilibrium concentrations. We call $G_0(T)$ the rate of creation of electron-hole pairs through thermal agitation. If we leave the system to evolve freely, the concentrations will return to the equilibrium values n_0, p_0. We may thus express the variation of n and p by the equations

$$n = n_0 + \Delta n, \tag{6.43}$$

$$\frac{dn}{dt} = \frac{d\,\Delta n}{dt} = G_0\,(T) - \frac{n}{\tau_n}, \tag{6.44}$$

where τ_n is a quantity which can vary with p. In equilibrium $\mathrm{d}(\Delta n)/\mathrm{d}t = 0$, $n = n_0$ and

$$G_0 = \frac{n_0}{\tau_{n_0}}, \tag{6.45}$$

where τ_{n_0} is the value of τ_n for $n = n_0$. Equation (6.45) can thus be written as

$$\frac{d\Delta n}{dt} = \frac{n_0}{\tau_{n_0}} - \frac{n}{\tau_n}. \tag{6.46}$$

Similarly we can write

$$\frac{d\Delta p}{dt} = \frac{p_0}{\tau_{p_0}} - \frac{p}{\tau_p}. \tag{6.47}$$

Equations (6.46) and (6.47) do no more than introduce the *ad hoc* quantities τ_n, τ_p. These quantities may well depend on n and p and in the general case we can say no more without explicit analysis of the microscopic recombination process. The kinematics of recombination can indeed be complex.

Things are simpler if we consider a **doped** semiconductor subject to moderate light excitation. We consider for example an n-type semiconductor, i.e., one where $n_0 >> p_0$. If the deviations Δn and Δp are much smaller than n_0, n will change little during the return to equilibrium. We may thus regard τ_p as independent of Δn and write $\tau_p = \tau_{p_0}$. Then

$$\frac{d\Delta p}{dt} = -\frac{\Delta p}{\tau_p} \quad \text{with} \quad \Delta p = p - p_0. \tag{6.48}$$

For this reason τ_p is called the lifetime of the minority carriers, here the holes.

The solution of Eq. (6.48) has the form

$$\Delta p = \Delta p_0 \exp(-t/\tau_p) + \text{const} \tag{6.49}$$

and if electrical neutrality holds at all times, $\Delta n = \Delta p$.

We can make this description more precise by considering a particular recombination process: the direct recombination of electron-hole pairs where the recombination involves the simultaneous disappearance of an electron and a hole. The evolution of the electron number is given by the equation

$$\frac{dn}{dt} = \frac{dp}{dt} = -A\,np + G, \tag{6.50}$$

where G is the total generation rate. Indeed the number of electrons recombining per unit time is proportional to the number of electrons present, as each electron has the same probability of recombining. Further this probability is, for an electron, proportional to the number of free holes it may encounter, hence the form of the first term. The factor A depends on the semiconductor but not on n or p. We have added the quantity G which represents the generation rate for pairs.

In thermodynamic equilibrium, i.e., in the absence of excitation, $dn/dt = 0$; $n = n_0$; $p = p_0$. Equation (6.50) gives the thermal pair generation rate as a function of the annihilation rate:

$$G_0(T) = A n_0 p_0 = A n_i^2. \tag{6.51}$$

If there is an additional generation rate g, e.g., by light, Eq. (6.50) allows us to write Eq. (6.51) as

$$\begin{aligned} \frac{dn}{dt} &= \frac{d\Delta n}{dt} = -A n p + G_0 + g, \\ \frac{d\Delta n}{dt} &= \frac{d\Delta p}{dt} = -A(n_0 + \Delta n)(p_0 + \Delta p) + G_0 + p, \\ \frac{d\Delta n}{dt} &= \frac{d\Delta p}{dt} = -A p_0 \Delta n - A(n_0 + \Delta n)\Delta p + g. \end{aligned} \tag{6.52}$$

In an n-type sample, where $n_0 >> p_0$, and if the injection of carriers is small (Δn and $\Delta p << n_0$) we have

$$\frac{d\Delta n}{dt} = \frac{d\Delta p}{dt} = -A n_0 \Delta p + g. \tag{6.53}$$

One defines

$$\frac{1}{\tau_p} = A n_0. \tag{6.54}$$

In the absence of excitation

$$\frac{d\Delta n}{dt} = -\frac{\Delta n}{\tau_p}, \tag{6.55}$$

$$\frac{d\Delta p}{dt} = -\frac{\Delta p}{\tau_p}. \tag{6.56}$$

The time constant for the disappearance of **electrons** and **holes** is the **same in an n-type semiconductor**, but it is the **lifetime of the minority carriers**. This lifetime decreases as the concentration of majority carriers increases.

Remarks

If g is constant in time, then in a steady state,

$$p = p_0 + g\tau_p. \tag{6.57}$$

If at some arbitrary time t_1 the light is switched off then, setting $t_1 = 0$,

$$p(t) = p_0 + g\tau_p \exp\left(-\frac{t}{\tau_p}\right). \tag{6.58}$$

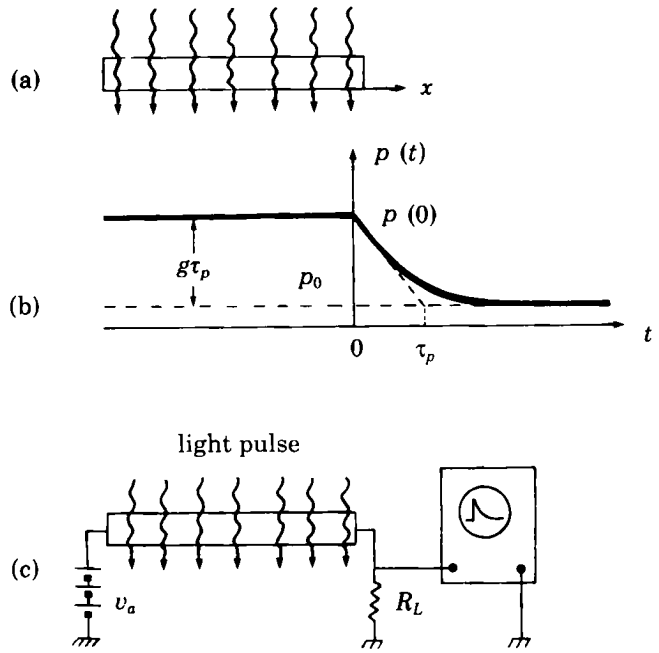


Fig. 6.6. (a) n-type sample subject to continuous irradiation; (b) behavior of the concentration of minority carriers when the irradiation is switched off; (c) measurement of the lifetime. One injects a light pulse of duration short compared with the lifetime. The variation of the resistance of the sample follows that of the number of carriers. (After Sze: "Semiconductor Devices," J. Wiley, 1985.)

This is the principle of the measurement of the lifetime of the minority carriers illustrated in Fig. 6.6.

The result we have obtained may seem paradoxical. **The lifetime of the minority carriers governs the dynamics of recombination.** We can understand this by referring to the recombination probabilities of Eq. (6.53). The recombination probability of an electron, no matter what its origin, is $An_0\Delta p/n_0 = A\Delta p$, and thus depends on Δp. It is therefore not uniquely determined. The recombination probability of an additional electron is $An_0\Delta p/\Delta n = An_0 = \tau_p^{-1}$. This is the inverse lifetime of the minority carriers. The recombination probability of a hole, which must have been created by photoexcitation since we neglected $p_0 << \Delta p$, is $An_0\Delta p/\Delta p$, i.e., the same. We should remark that the recombination probability of a hole is large because it sees many electrons.

Of course in a p-type semiconductor the reverse occurs; recombination is governed by the lifetime of the minority carriers which are then the electrons.

Recombination Processes

Recombination processes can be classified into three main categories.

direct recombination

Direct recombination is the reverse of the creation of an electron-hole pair by a photon. We consider the ensemble (photon + crystal). If we neglect the interaction between radiation and matter resulting from the force exerted by the electric field of the light on the electrons, the state (photon + valence-band electron) is degenerate with the state where the photon is absorbed and the electron is in the conduction band. If we leave the system in one of these states it will stay there indefinitely.

This is no longer true if we take account of the interaction between radiation and matter. As we saw in Sect. 6.1 and Appendix 6.1, in the presence of this interaction the state (photon + electron in the valence band) is no longer an eigenstate of the system. If we leave the system in this state at time $t = 0$, then after time t the quantum state will have a non-zero projection on the state (absorbed photon + conduction-band electron). There is thus a finite probability of absorption of the photon and the creation of a pair. Conversely an electron in the conduction band can relax to the valence band by the same effect, with light emission. We call this **radiative recombination**. Appendix 6.2 gives the calculation of the radiative recombination probability. We can also have direct non-radiative recombination via processes involving several electrons. An electron falls back into the valence band, the corresponding energy being given to another conduction electron whose energy in the band increases. This is called Auger recombination.

recombination through traps or deep impurities

In this process recombination occurs in two stages well separated in time. For example a conduction electron may first be captured by an impurity whose level lies deep in the band gap. At a later time this occupied center may capture a hole from the valence band (or, equivalently, emit an electron into the valence band), finally ensuring the recombination of an electron-hole pair. As shown in Appendix 6.2 the probability of a radiative transition decreases very rapidly as the photon energy increases. Performing

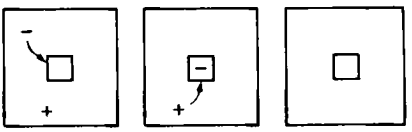


Fig. 6.7. Recombination by trapping: (a) electron capture; (b) hole capture; (c) recombination.

the transition in two stages, thanks to the trapping, can make this process much more probable than the emission of a single photon.

The process is shown schematically in Fig. 6.7: a recombination center alternately captures an electron and a hole and thus "catalyzes" their recombination.

In fact, in the intermediate stage it is possible for an electron to be re-excited into the conduction band before capturing a hole. In this case we speak of **a slow trap**, reserving the term **recombination center** for the case where the capture probability of a hole is larger than the probability of re-emission of the electron towards the conduction band. It appears that the kinematics of recombination by deep centers can be complex.

We give here some orders of magnitude to emphasize the importance of recombination via deep centers: we shall see later that the direct recombination time in germanium is of the order of 1 s. In the purest known crystals the measured time is several milliseconds. It has been shown that the presence of a relative concentration of 10^{-7} copper atoms in germanium reduces this time to 10^{-6} s. The study of recombination processes via deep centers is a very active field of the physics of semiconductors.

Given that most crystalline imperfections (impurities, gaps, dislocations, grain boundaries) can produce states in the band gap, we can see here once again why the manufacture of **pure single crystals** (cf. Sect. 4.9) is so important (both terms are necessary: chemically pure single crystals without crystallographic defects) if one wishes to obtain long lifetimes.

surface recombination

Even a perfectly pure and perfectly regular crystal has an external surface which breaks the periodicity. This break means that not all the chemical bonds of the surface atoms can be satisfied. There are therefore quantum states, localized near the surface, with an energy within the band gap. The presence of surface impurities also contributes to the presence of "surface states" which can be very effective recombination states.

Figure 6.8 illustrates the various recombination processes.

6.3 Photoconductivity and its Applications

6.3a Detection of Electromagnetic Radiation

A direct consequence of the creation of electron-hole pairs is photoconductivity. If there is an excess Δn and Δp of carriers, the photoconductivity of the semiconductor varies as $\Delta\sigma$:

$$\Delta\sigma = \Delta n\, \mu_e\, e + \Delta p\, \mu_h\, e. \tag{6.59}$$

A commonly used detector of visible light is cadmium sulphide CdS. In CdS the band gap E_g is direct (extrema of the valence and conduction

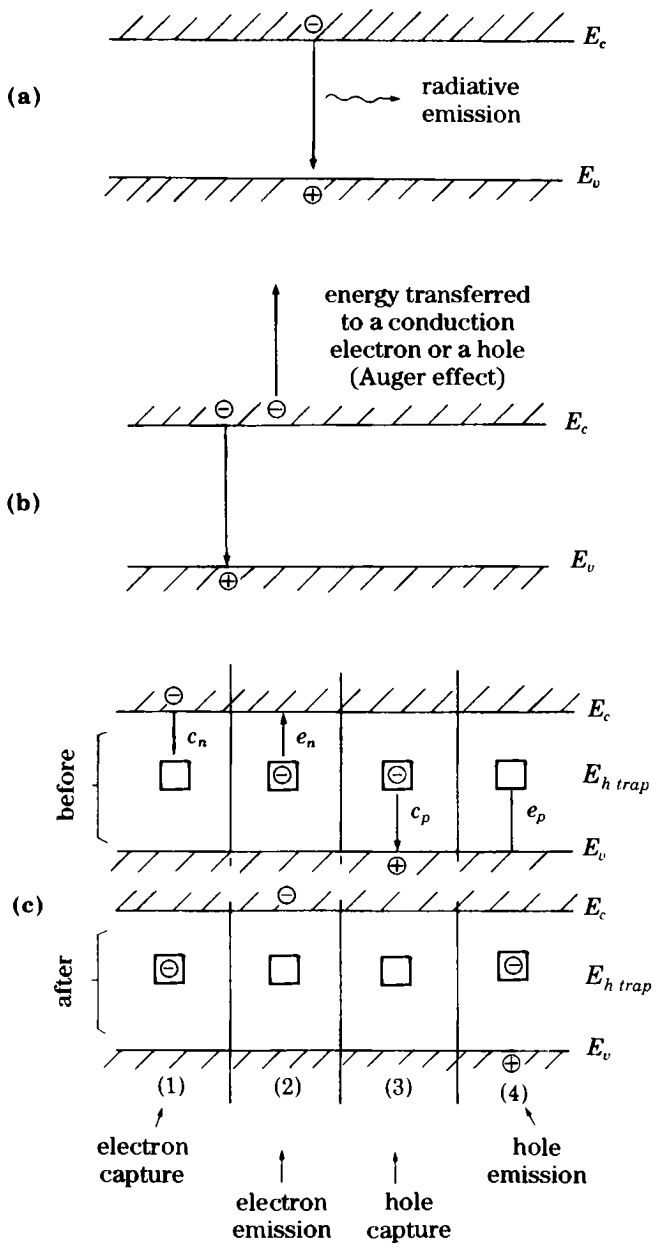


Fig. 6.8. The various recombination processes: (a) direct radiative recombination; (b) direct Auger recombination; (c) recombination via trapping on a deep center. In the latter case the figure shows the state of the system before and after each of the stages (1), (2), (3), (4). (After S.M. Sze, Physics of Semiconductor Devices.)

bands are at $k = 0$) and the width of the gap is 2.4 eV at 300 K. The absorption threshold is therefore 2.4 eV, corresponding to a wavelength of 520 nm, in the green region of the visible spectrum: green and blue are absorbed while longer wavelengths (yellow, red) are transmitted. A CdS crystal thus has a yellow–orange color when seen in transmission.

There are many semiconductors with band gaps less than 1 eV, whose absorption thresholds are in the infrared. For example the three lead salts PbS, PbSe, and PbTe have band gaps of order 0.2 eV. Some alloys $Hg_{1-x}Cd_xTe$ have an E_g of 0.04 eV, which corresponds to $\lambda = 30$ μm, in the far infrared. These materials are used to build infrared detectors.

In a material containing impurities with quantum states in the band gap, light can excite electrons from these localized states into the conduction band. In this case there is light absorption at energies less than E_g and creation of conduction electrons without the creation of free holes. Recombination occurs when the electron falls back to the localized level. In this case one speaks of **extrinsic** absorption and photoconductivity. Germanium doped with gold is used as a sensitive infrared detector between 10 and 20 μm.

6.3b Electrophotography

The xerographic photocopying process is an important application of photoconductivity. The photoconducting material is amorphous selenium, a semiconductor whose band gap is around 2.1 eV at 300 K. This material is deposited by vacuum evaporation in the form of a thin film on a metallic substrate. Selenium has very high resistivity of around 10^{14} Ω.m. The

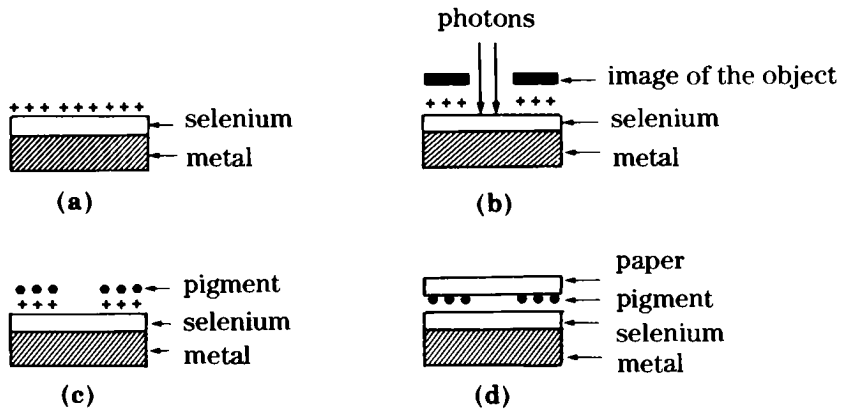


Fig. 6.9. Principle of electrophotography: (a) deposit of positive charges over the surface; (b) recording the image; (c) deposit of pigment; (d) transfer of the pigment to paper. (After Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

first stage of electrophotography consists of depositing a positive charge on the semiconductor surface using an electric discharge (corona effect). The charge stays on the surface because of selenium's high resistance. This is shown in Fig. 6.9(a). The second stage consists of sending the luminous image one wishes to record on to the selenium (Fig. 6.9(b)). The illuminated regions become conducting and the charge moves away while the dark parts of the image remain charged. The resistivity of the semiconductor is so high that the charge does not leak laterally. The result of this operation is a distribution of positive charges that replicates the dark regions of the image being copied. We have obtained an **electric image**. The third stage (Fig. 6.9(c)) consists of coating the electric image with black pigment which preferentially fixes itself electrostatically to the charged regions that have not been illuminated in the second stage. The last stage is the transfer of the black pigment to ordinary paper (Fig. 6.9(d)). The pigment is finally fixed to the paper by heating and provides a permanent positive image.

6.3c The Vidicon Tube

The Vidicon tube is the light-sensitive part of the early TV camera. The tube, shown schematically in Fig. 6.10, contains a thin layer of a photoconductor sensitive in the visible — for example As_2S_3 or PbO, for which E_g is, respectively, 2.5 and 2.3 eV. The material must have a very high resistivity $\sim 10^{13}$ Ω.m in the dark. Photons reach the photoconducting material after crossing a transparent conducting electrode. The photoconductor is continuously scanned from below by an electron beam, which scans its surface sequentially. If the electron beam arrives at a point that is illuminated on the front, the material conducts, and a current proportional to the light intensity at this point will be produced.

We thus obtain an electrical signal containing information about the light received at a given point of the screen. This signal is then processed,

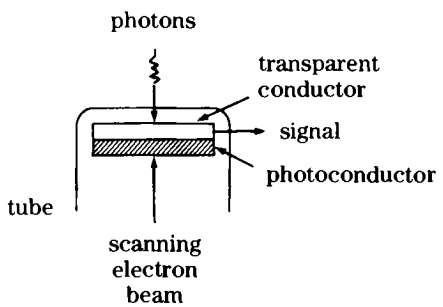


Fig. 6.10. Schematic view of a Vidicon tube.

and then may for example be reproduced on the cathode-ray tube of a TV set.

We note that the production of the image on the cathode-ray tube also uses semiconductors. The screen is covered with powder, usually alloys such as ZnS–ZnSe, which emit visible light when excited by a high-energy beam of electrons. This phenomenon is called cathodoluminescence. ZnS and ZnSe are semiconductors whose band gap is in the ultraviolet or the blue, and which are doped by various elements, generally copper or silver associated with a halogen. These impurities create deep levels which cannot be described using effective mass theory. Luminescence occurs between a delocalized band state and a deep level, or between two deep levels. By judicious choice of the doping agents one can cover the visible spectrum. The three color bands of color television are obtained by using three different powders with adapted doping chemicals.

Image converters, which allow night vision by transforming an infrared image into a visible one, consist of a Vidicon tube followed by a cathode-ray tube.

Appendix 6.1

Quantum System Submitted to a Sinusoidally Varying Perturbation

We are interested in a system (here a semiconductor) submitted to a travelling electromagnetic wave. The system Hamiltonian is

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1. \tag{6.60}$$

Here $\mathcal{H}_0$ describes the semiconductor in the absence of the electromagnetic field and $\mathcal{H}_1$ its interaction with the field. We assume $\mathcal{H}_1$ small compared with $\mathcal{H}_0$; it has time dependence:

$$\mathcal{H}_1(t) = \mathcal{H}_1 \cos(\omega t - \mathbf{K} \cdot \mathbf{r}), \tag{6.61}$$

where ω is the wave frequency and $\mathbf{K}$ its wave vector.

We assume that at the initial time t_0 the system is in one of the known eigenstates $|n>$ of $\mathcal{H}_0$:

$$\mathcal{H}_0 |n> = E_n |n> . \tag{6.62}$$

As this state is not an eigenstate of $\mathcal{H}$ the system must evolve according to the Schrödinger equation:

$$i\hbar \frac{d}{dt} |\psi(t)> = \mathcal{H} |\psi(t)> . \tag{6.63}$$

We seek the probability of finding the system in the eigenstate $|f>$ at time t(with $t > t_0$). Hence we project the state $|\psi(t)>$ along the basis $|n>$ of eigenstates of $\mathcal{H}_0$:

$$|\psi(t)> = \sum_n \gamma_n(t) \exp(-iE_n t/\hbar) |n> . \tag{6.64}$$

If the Hamiltonian $\mathcal{H}_1$ were time-independent, the γ_n would be constant. Applying the Schrödinger equation to $|\psi(t)>$ we find the relation between the coefficients $\gamma_n(t)$ and their time derivatives $\mathrm{d}\gamma_n(t)/\mathrm{d}t$:

$$
\begin{aligned}
& i\hbar \sum_n \frac{d}{dt}\gamma_n(t)\exp(-iE_nt/\hbar)|n> \\
& + i\hbar \sum_n \gamma_n(t)\left(-\frac{iE_n}{\hbar}\right)\exp(-iE_nt/\hbar)|n> \\
& = \sum_n (\mathcal{H}_0 + \mathcal{H}_1)\gamma_n(t)\exp(-iE_nt/\hbar)|n>,
\end{aligned}
$$

which, on using Eq. (6.62), give

$$
\begin{aligned}
& \sum_n i\hbar \frac{d}{dt}\gamma_n(t)\exp(-iE_nt/\hbar)|n> \\
& = \sum_n \mathcal{H}_1\gamma_n(t)\exp(-iE_nt/\hbar)|n> . \qquad (6.65)
\end{aligned}
$$

Multiplying by $< k|$ we get

$$
i\hbar \frac{d}{dt}\gamma_n(t) = \sum_n \gamma_n \exp i(E_k - E_n)t/\hbar < k|\mathcal{H}_1|n> . \qquad (6.66)
$$

We thus have a set of differential equations to solve, with the initial condition that at $t = t_0 = 0$ the system is in the state $|i>$, i.e., $\gamma_i(t_0) = 1, \gamma_k(t_0) = 0$ for $k \neq i$. The probability $P_{i\to f}$ of finding the system in the final state f at t when it was in the initial state i at $t = 0$ is

$$
P_{i\to f}(t) = |< f|\psi(t)>|^2 = |\gamma_f(t)|^2. \qquad (6.67)
$$

We have assumed $\mathcal{H}_1$ small compared with $\mathcal{H}_0$, so we set

$$
\mathcal{H}_1 = \lambda\mathcal{H}_1', \qquad (6.68)
$$

where λ is a small parameter. We can expand in powers of λ, and solve Eq. (6.66) by equating coefficients of powers of λ:

$$
\begin{aligned}
& i\hbar\left(\frac{d}{dt}\gamma_k^0(t) + \lambda\frac{d}{dt}\gamma_k^1(t) + \lambda^2\frac{d}{dt}\gamma_k^2(t) + \ldots\right) \\
& = \sum_n \left[\gamma_n^0(t) + \lambda\gamma_n^1(t) + \lambda^2\gamma_n^2(t) + \ldots\right] \times \\
& \exp[i(E_k - E_n)t/\hbar] < k|\lambda\mathcal{H}_1|n> . \qquad (6.69)
\end{aligned}
$$

To order zero:

$$
i\hbar\frac{d}{dt}\gamma_k^0(t) = 0. \qquad (6.70)
$$

To first order:

$$i\hbar\lambda\frac{d}{dt}\gamma_k^1(t) = \sum_n \gamma_n^0(t)\exp[i(E_k - E_n)t/\hbar] < k|\lambda\mathcal{H}_1'|n>, \tag{6.71}$$

and so on.

Retaining just the first order in λ and using the initial conditions we get

$$\gamma_k^0(t) = \delta_{ik},$$
$$i\hbar\frac{d}{dt}\gamma_k^1(t) = \gamma_i\exp[i(E_k - E_i)t/\hbar] < k|\mathcal{H}_1|i> . \tag{6.72}$$

For the final state $|f>$, such that $\gamma_f(t_0) = 0$,

$$\gamma_f^1(t) = \frac{1}{i\hbar}\int_0^t \exp[i(E_f - E_i)t/\hbar] < f|\mathcal{H}_1|i> dt. \tag{6.73}$$

In this approximation, called the Born approximation, the transition probability from $|i>$ to $|f>$ is

$$P_{i\to f}(t) = |\gamma_f^1(t)|^2. \tag{6.74}$$

We assumed $\gamma_f^1(t)$ small, so $P_{i\to f} << 1$.

We now give the results more explicitly for the interaction of a solid with the electromagnetic field. The interaction Hamiltonian is (cf. Eqs. (6.13) and (6.14)):

$$\mathcal{H}_1 = \frac{eA}{2m}\mathbf{a}_0[\exp i(\omega t - \mathbf{K}\cdot\mathbf{r}) + \exp[-i(\omega t - \mathbf{K}\cdot\mathbf{r})]]\cdot\mathbf{p}, \tag{6.75}$$

so that

$$\mathcal{H}_1 = C\exp i\omega t + C^+\exp(-i\omega t) \tag{6.76}$$

with

$$C^+ = \frac{eA\exp i\mathbf{K}\cdot\mathbf{r}}{2m}\,\mathbf{a}_0\cdot\mathbf{p}. \tag{6.77}$$

We recall that with the gauge Eq. (6.3) $\mathbf{A}$ and $\mathbf{p}$ commute. We set

$$E_f - E_i = \hbar\omega_{if} > 0 \tag{6.78}$$

as we are interested in **photon absorption** from the initial state. Expression (6.73) for $\gamma_f^1(t)$ is then

$$\begin{aligned}\gamma_f^1(t) =& \frac{1}{i\hbar}\int_0^t \exp i\omega_{if}t[< f|C\exp i\omega t|i> + < f|C^+\exp(-i\omega t)|i>]\,dt \\ =& \frac{1}{i\hbar}\left[< f|C|i> \frac{\exp i(\omega_{if}+\omega)t - 1}{i(\omega_{if}+\omega)} + \right. \\ & \left. < f|C^+|i> \frac{\exp i(\omega_{if}-\omega)t - 1}{i(\omega_{if}-\omega)}\right].\end{aligned} \tag{6.79}$$

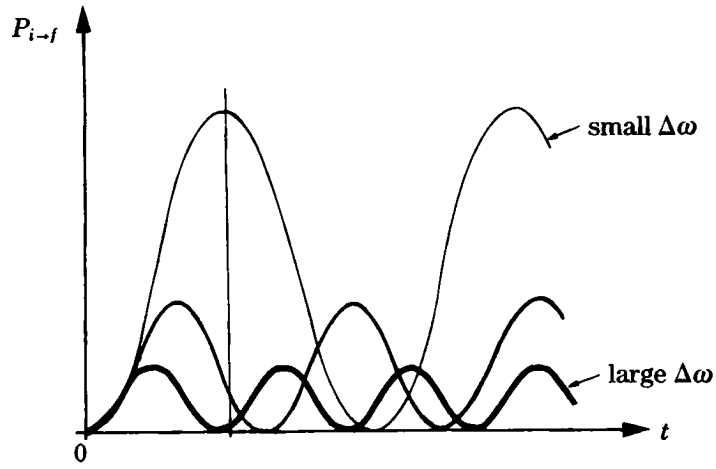


Fig. 6.11. Variation of the absorption transition probability $P_{i\to f}$ versus time for different values of $\Delta\omega = \omega_{if} - \omega$.

We consider the neighborhood of $\omega = \omega_{if}$: The first term is of the order of ω^{-1}, the period of the electromagnetic wave, while the second is of the order of t, much longer than the period of the wave. The first term is therefore negligible compared with the second, so that

$$P_{i\to f}(t) = \frac{4}{\hbar^2} | < f|C^+|i > |^2 \frac{\sin^2[(\omega_{if} - \omega)t/2]}{(\omega_{if} - \omega)^2} \tag{6.80}$$

and

$$P_{i\to f}(t) = \left(\frac{eA}{2\hbar m}\right)^2 | < f| \exp(i\mathbf{K}\cdot\mathbf{r})\mathbf{a}_0 \cdot \mathbf{p}|i > |^2 \frac{\sin^2[(\omega_{if} - \omega)t/2]}{[(\omega_{if} - \omega)/2]^2}. \tag{6.81}$$

The variation of $P_{i\to f}(t)$ as a function of time for several values of $\Delta\omega = \omega_{if} - \omega$ is shown in Fig. 6.11. $P_{i\to f}$ is a sinusoidal function of time whose peak value is proportional to $1/(\Delta\omega)^2$. For very small t ($t\Delta\omega << 1$ for all $\Delta\omega$) all the curves coincide, and grow as t^2, getting out of phase for larger t. Conversely $P_{i\to f}$ can be regarded as a function of ω with a resonance of width $\Delta\omega = \pi/2t$ around $\omega = \omega_{if}$ (cf. Fig. 6.1).

To calculate the **emission** probability from an excited state we have to distinguish between induced and spontaneous emission. The calculation of induced emission is symmetrical with the one we have just performed for absorption: one simply reverses the roles of $|i>$ and $|f>$. The calculation of spontaneous emission is complex but involves the same matrix element as for absorption, so that the same selection rules do apply.

Appendix 6.2

Calculation of the Radiative Recombination Probability

As we have seen in Sects. 6.1 and 6.2, there is a close connection between light absorption by a semiconductor and radiative recombination, as both involve the same quantum states and the same electromagnetic interaction.

We show here that if the absorption coefficient is known one can deduce the radiative lifetime. The basis of the proof consists of noting that in thermal equilibrium the rate of creation of electron-hole pairs is equal to the recombination rate. Using the **detailed balance** of these processes for each frequency interval $\nu, \nu+\mathrm{d}\nu$ we can write that the emission rate $R(\nu)\mathrm{d}\nu$ is equal to the pair generation rate

$$R(\nu)\,d\nu = P(\nu)\rho(\nu)d\nu, \tag{6.82}$$

where $P(\nu)$ is the probability per unit time of absorbing a photon of energy $h\nu$ and creating a pair, and $\rho(\nu)\mathrm{d}\nu$ is the photon density between frequencies ν and $\nu+\mathrm{d}\nu$ for a unit volume. In thermal equilibrium this is given by the Planck blackbody law modified to account for the refractive index n of the medium:

$$\rho(\nu)d\nu = \frac{8\pi\,\nu^2\,n^2}{c^3}\,\frac{1}{\exp\left(h\,\nu/kT\right)-1}\,\frac{d(n\,\nu)}{d\,\nu}d\,\nu. \tag{6.83}$$

This formula is easily proved by noting that the density of photon states in the volume Ω, $(\Omega/8\pi^3)4\pi k^2\mathrm{d}k$, multiplied by 2 (for the 2 polarization states of light), should be multiplied by the Bose factor $[\exp(h\nu/kT)-1]^{-1}$. As the wave vector k is $2\pi n\nu/c$, the number of occupied states between ν and $\nu+\mathrm{d}\nu$ is

$$2\times\frac{\Omega}{8\pi^3}4\pi\left(2\pi n\frac{\nu}{c}\right)^2 d\left(2\pi n\frac{\nu}{c}\right)\times\frac{1}{\exp\left(h\,\nu/kT\right)-1} \tag{6.84}$$

giving Eq. (6.83).

We also have to relate the probability $P(\nu)$ per unit time of absorbing a photon of energy $h\nu$ to the light-absorption coefficient $\alpha(\nu)$. The electric field of a wave propagating a distance x into a medium of refractive index n and absorption coefficient α is

$$\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}_0 \exp i\,(2\pi\,\nu t - kx)\ \exp(-\alpha\, x/2). \tag{6.85}$$

Here α and n depend on ν. The travelling energy is proportional to

$$I(x) = \overline{(\mathrm{Re}\ \boldsymbol{\mathcal{E}})^2} \simeq \frac{\boldsymbol{\mathcal{E}}_0^2}{2} \exp(-\alpha x). \tag{6.86}$$

The quantity α^{-1}, the inverse of the absorption coefficient, can be regarded as the mean free path of a photon of energy $h\nu$ in the medium before it produces a band-to-band transition at this energy. In fact $\alpha\ \mathrm{d}x$ is the probability that a photon is absorbed between x and $x + \mathrm{d}x$. The change of intensity over $\mathrm{d}x$ is

$$dI = I(x + dx) - I(x) = -I(x)\alpha\, dx \tag{6.87}$$

and integration of this equation gives just the form (6.86). The probability that a photon **will not be absorbed** over the distance x is thus $\exp(-\alpha x)$ and the probability that it is absorbed **exactly** between x and $x+\mathrm{d}x$ is then

$$\alpha\, dx \exp(-\alpha\, x). \tag{6.88}$$

The photon mean free path is then

$$\int_0^\infty dx\ \alpha\, x\ \exp(-\alpha\, x) = \alpha^{-1}. \tag{6.89}$$

The mean lifetime $\tau(\nu)$ of such a photon is thus $\alpha^{-1} v_g^{-1}$ where v_g is the wave group velocity in the medium

$$v_g = \frac{d\omega}{dk} = c\frac{d\nu}{d\, n\nu}, \tag{6.90}$$

finally giving the inverse photon lifetime

$$\alpha\, v_g = \frac{1}{\tau(\nu)} = \alpha\, c\frac{d\nu}{d\, n\nu}. \tag{6.91}$$

The probability $P(\nu)$ per unit time of absorbing a photon is the inverse of the lifetime:

$$P(\nu) = \frac{1}{\tau(\nu)}, \tag{6.92}$$

so, substituting Eqs. (6.83), (6.91), and (6.92) into Eq. (6.82),

$$R(\nu)\, d\nu = \frac{8\pi}{c^2} \frac{\nu^2 n^2 \alpha}{\exp(h\nu/kT) - 1} d\nu. \tag{6.93}$$

This is called the Van Roosbroeck–Shockley relation [Physical Review **94** 1558, (1954)]. It is the fundamental relation between the expected emission spectrum and the absorption spectrum. The total number of recombinations per unit volume and per second is obtained by integrating over frequency. Setting $u = h\nu/kT$:

$$R = \int_0^\infty R_\nu \, d\nu = \frac{8\pi n^2}{c^2} \left(\frac{kT}{h}\right)^3 \int_0^\infty \frac{n^2(\nu)\alpha(\nu)u^2}{e^u - 1} du. \tag{6.94}$$

We note in Eq. (6.93) that if $\alpha(\nu)$ vanishes, corresponding to $h\nu < E_g$, the emission will also vanish at frequency ν. Formula (6.93) allows us to transform the absorption spectrum of a semiconductor of known refractive index into its emission spectrum. The integral in Eq. (6.94) is negligible except over a very narrow frequency band of width kT/h near the fundamental absorption because of the factor e^u in the denominator and because $\alpha = 0$ for $h\nu < E_g$. We see that the light emission by a semiconductor is relatively monochromatic. This is the reason why the electroluminescent diodes used in control panel lights or in fiber optic telecommunications emit well defined colors.

Although the calculation above was for band-to-band transitions, it also holds for transitions between any pairs of states whatsoever. We see from Eq. (6.93) that if we consider a transition from the conduction band to a deep level in the band gap, the first stage of recombination through impurity centers, the factor $\exp(h\nu/kT)$ is much smaller, as $h\nu \sim E_g$ for the band–band transition is replaced by $h\nu \sim E_g/2$ for recombination by centers; hence the probability of the latter process is greatly enhanced.

In thermal equilibrium (Eq. (6.51))

$$R = G_0(T) = A\, n_i^2. \tag{6.95}$$

The radiative lifetime defined by Eq. (6.54) is then given for n-type material by

$$\frac{1}{\tau_p} = \frac{R}{n_i^2} n_0 = \frac{R}{p_0}. \tag{6.96}$$

In an intrinsic material with $\Delta p = \Delta n << n_i$ we have from Eq. (6.53),

$$\frac{d\,\Delta n}{dt} = -2A\, n_i\, \Delta n + g \tag{6.97}$$

and the intrinsic lifetime is defined by

$$\frac{1}{\tau_i} = \frac{2R}{n_i}. \tag{6.98}$$

The following table gives the values of the lifetimes calculated from Eqs. (6.94), (6.96), and (6.98).

Radiative Recombination at 300 K

Material	E_g (eV)	n_i ($\times 10^{20}\,\mathrm{m}^{-3}$)	τ (intrinsic)	τ for $10^{23}\,\mathrm{m}^{-3}$ majority carriers (μs)
Si	1.08	0.00015	4.6 h	2500
Ge	0.66	0.24	0.61 s	150
GaSb	0.71	0.043	0.009 s	0.37
InAs	0.31	16	15 μs	0.24
InSb	0.18	200	0.62 μs	0.12
PbS	0.41	7.1	15 μs	0.21
PbTe	0.32	40	2.4 μs	0.19
PbSe	0.29	62	2.0 μs	0.25
GaP	2.25			3000

R. N. Hall, Proc. Institution of Electrical Engineering 106B Suppl. No. *17, 923* (1959).

We observe the wide range of recombination times for different semiconductors.

Appendix 6.3

Semiconducting Clusters for Non-Linear Optics

In recent years much interest has been focussed on "optronics," a new technology in which electronics would be replaced by optics (or electron current by photon flux). This needs non-linear optical devices using non-linear optical materials. In such a material the absorbed radiation power is not a linear function of the incident power. This can be realized in particular with small clusters of semiconducting materials.

For usual semiconducting samples or devices of macroscopic size, the size effects are negligible. Indeed the quantum confinement effect (as measured by the energy of the lowest state in a finite cubic box) of a microelectronic transistor of typical size of $L = 3$ μm corresponds to an energy shift $\delta E = (\hbar^2/2m)(k_x^2 + k_y^2 + k_z^2) = 3\hbar^2/2mL^2 = 10^{-8}$ eV. This energy is very small compared to the thermal energy and plays no role. Even if m is replaced by the effective mass m^* of order $m/10$ this confinement energy is small. However if one considers very small clusters of size ranging from 1 to 10 nm the confinement effect becomes large and can be observed. Using the effective mass approximation one can write the energy of the lowest state corresponding to the conduction band as

$$E_{c,\text{cluster}} = E_{c,\text{bulk}} + 3\hbar^2/2m_e L^2. \tag{6.99}$$

In the same manner one can write the highest energy corresponding to the valence band as

$$E_{v,\text{cluster}} = E_{v,\text{bulk}} - 3\hbar^2/2m_h L^2. \tag{6.100}$$

In consequence the new energy for the gap is

$$E_{g,\text{cluster}} = E_{g,\text{bulk}} + (3\hbar^2/2L^2)(m_e^{-1} + m_h^{-1}). \tag{6.101}$$

This increase in energy may be partially compensated by excitonic effects that we neglect here for the sake of simplicity (see Chap. 6.1d for excitonic effects).

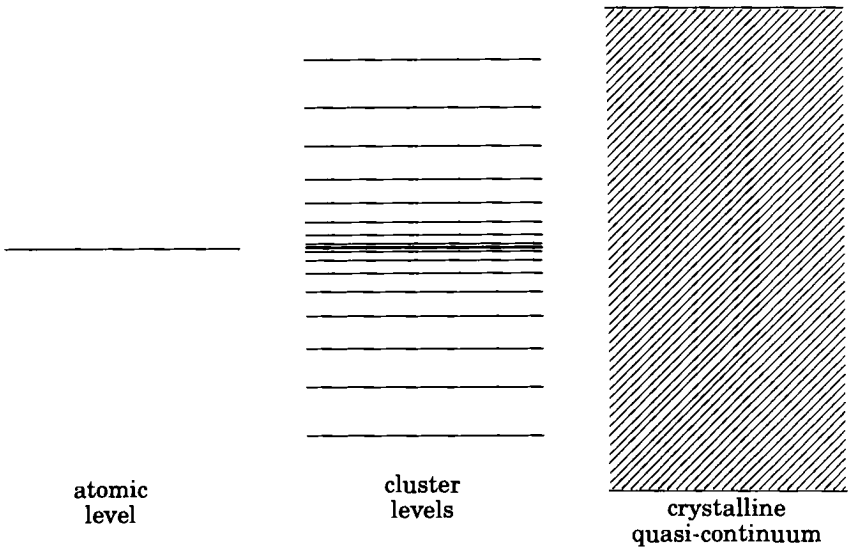


Fig. 6.12. Comparative scheme of the energy levels for atoms, clusters, and large crystals. In large crystals, the distance between energy levels is so small that one speaks of a quasi-continuum.

The next effect is that the splitting between conduction (or valence) levels is increased drastically for the same reason. There no longer exists a quasi-continuum of states like in a large crystal but a finite number of discrete levels. This is schematized in Fig. 6.12. Finally the energy level scheme near the new gap is shown in Fig. 6.13.

The non-linear absorption of these small clusters is related to the fact that their optical absorption is very large because the "oscillator strength" is spread over a limited number of states. This means that the absorption probability for photons is larger than in crystals of macroscopic size. This can be understood from Eq. (6.21). The absorption probability for photons corresponding to the excitation from a single level to a single final level of width ΔE is proportional to the density of final states $n(E_f) = 1/\Delta E$. In a bulk semiconductor the term which corresponds to ΔE is of the order of the bandwidth (a few eV) whereas in the cluster the width of the discrete levels is typically 10 to 100 hundred times smaller. Correspondingly the absorption probability is very large.

The absorption spectra of glasses containing clusters of $CdS_{1-x}Se_x$ is shown in Fig. 6.14. The smaller the size of the cluster, the higher the absorption energy. Because of the enhanced absorption, it is possible to "saturate" the optical transition, that is to equalize the populations of the fundamental and excited states, using moderate radiation power. Here the fundamental

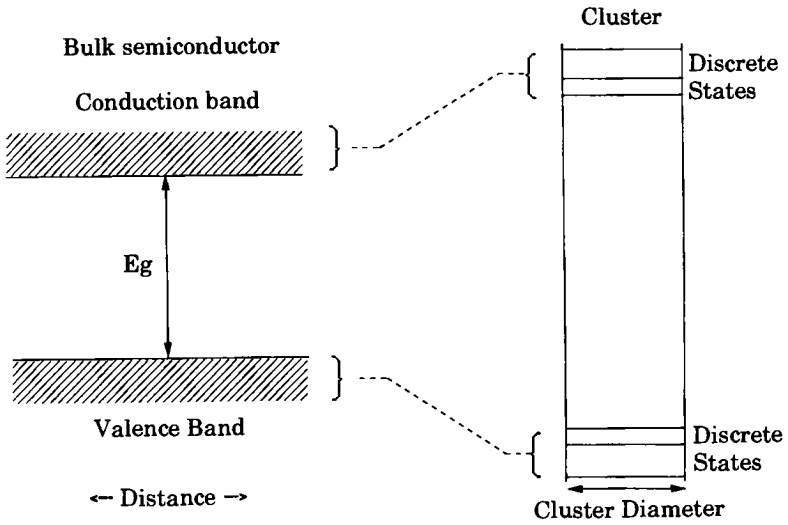


Fig. 6.13. Energy levels near the forbidden band gap for bulk and small clusters of semiconductors. The increase in the distance between valence states and excited levels is a consequence of quantum confinement.

state corresponds to a full valence band and the excited state corresponds to the occupation of the lower state in the conduction band. Of course saturation can be achieved if the recombination is slow enough or if the intensity of the laser irradiation (pump laser) is large enough. Once saturation exists it is no longer possible to absorb photons of the same wavelength because the number of electrons excited by light from the valence to the conduction band is equal to the number of electrons which transit from a state of the conduction band to an empty state of the valence band under the light illumination. This phenomenon, which is called "stimulated emission," is due to the fact that the transition probability from i to f, $P_{i,f}$ that one can compute from Eqs. (6.79)–(6.81) is equal to the transition probability from f to i: $P_{f,i}$. In this saturation situation, the system has become non-linear. There exists a "self-induced transparency" because no more light can be absorbed at that wavelength.

Another illustration of quantum confinement is the recent observation of light emission in the visible range by porous silicon. The band gap of ordinary crystalline silicon is situated in the infrared range and for this reason silicon cannot be used as a light-emitting material. Porous silicon is usually obtained by electropolishing of silicon in aqueous hydrofluoric acid and has been shown recently (1990) to emit light in the visible range. This fact (emission of light at an energy larger than the gap energy of bulk silicon) is explained by the fact that this material which is highly porous (50% porosity) is made of an irregular columnar structure with a typical

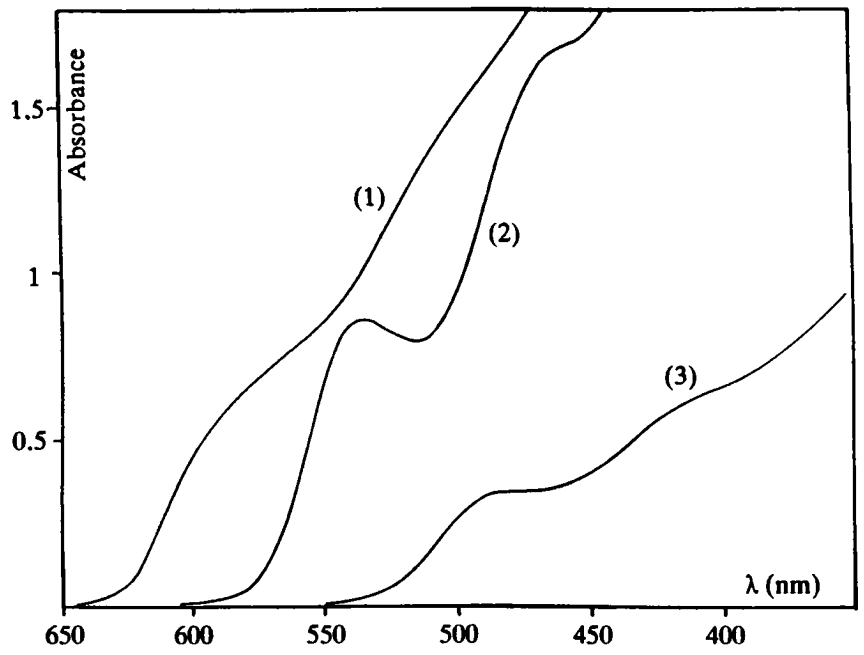


Fig. 6.14. Absorption spectra of three glasses containing small clusters of $CdS_{1-x}Se_x$. The cluster diameter is 12 nm for (1), 5 nm for (2), and 2.5 nm for (3). The cluster diameter can be measured by electron microscopy. The quantum confinement effect due to the size reduction from (1) to (3) creates a "blue shift" of the absorption threshold. The apparition of structure is a consequence of the fact that the distance between energy levels is also increased.

size of a few nanometers. Also, in this material the quantum confinement is the essential mechanism for the increase of the effective band gap, as schematized in Fig. 6.13.

7.

Carrier Injection by Light

When a semiconductor is subject to a flux of radiation of energy exceeding that of the band gap the rate of creation of electron-hole pairs varies spatially as light is progressively absorbed. Under these conditions there appear diffusion currents due to the non-equilibrium injection of carriers by light.

We shall see that the system rearranges itself in order to maintain electrical neutrality. The excess carriers created near the surface redistribute themselves over a finite distance, the diffusion length, determined by the diffusion coefficient and the recombination time of the minority carriers. The concepts introduced in this chapter will be useful for the study of more complex systems, such as inhomogeneous semiconductors, which are the building blocks of solid state electronic devices: junctions and transistors. These devices are studied in Chaps. 8–10.

7.1 Basic Equations for Semiconductor Devices

We recall here the current Eqs. (5.31) and (5.32) involving the drift and diffusion currents:

$$\mathbf{J}_e = ne\mu_e \, \boldsymbol{\mathcal{E}} + eD_e \nabla n, \tag{7.1}$$

$$\mathbf{J}_h = pe\mu_h \boldsymbol{\mathcal{E}} - eD_h \nabla p, \tag{7.2}$$

$$\mathbf{J} = \mathbf{J}_e + \mathbf{J}_h. \tag{7.3}$$

The continuity equations, which express the particle number conservation, have to include recombination terms. These have to be considered for each type of particle. For the minority carriers, n_p electrons in p-type material, p_n holes in n-type material:

$$\frac{\partial n_p}{\partial t} = g_n - \frac{n_p - n_p^0}{\tau_n} + \frac{\nabla \cdot \mathbf{J}_e}{e}, \tag{7.4}$$

$$\frac{\partial p_n}{\partial t} = g_p - \frac{p_n - p_n^0}{\tau_p} - \frac{\nabla \cdot \mathbf{J}_h}{e}, \tag{7.5}$$

where g_n and g_p are the creation rates arising from external excitations, and n_p^0, p_n^0 the equilibrium concentrations of minority carriers. If the system is electrically neutral (a question discussed in Sect. 7.2 and Appendix 7.1), the concentration of majority carriers is obtained from the neutrality relation

$$n_n - n_n^0 = p_n - p_n^0 \tag{7.6}$$

and an analogous relation in p-type material:

$$n_p - n_p^0 = p_p - p_p^0. \tag{7.7}$$

7.2 Charge Neutrality

Consider a homogeneously doped semiconductor in which one has created carrier concentrations $n_0 + \Delta n$ and $p_0 + \Delta p$. If Δn differs from Δp there will be a space charge of density

$$\rho = e(\Delta p - \Delta n). \tag{7.8}$$

We wish to show that in fact $\Delta n = \Delta p$ except in the presence of a very strong field. The charge and electric field are related by Poisson's equation

$$\nabla \cdot \boldsymbol{\mathcal{E}} = \frac{\rho}{\epsilon} = \frac{e\,(\Delta p - \Delta n)}{\epsilon_0\,\epsilon_r}. \tag{7.9}$$

We take $p = 10^{18}\mathrm{m}^{-3}$, $\Delta p - \Delta n = 10^{-2}p$, and $\epsilon_r \sim 10$, $\epsilon_0 \simeq 10^{-11}$ F.m^{-1}. Then $\nabla \cdot \boldsymbol{\mathcal{E}}$ is of the order of 1.6×10^7 V $\cdot$ m^{-2} and the electric field produced by the charges in a slab of thickness d is $1.6 \times 10^7 d$ (V $\cdot$ m^{-1}). Over a 1 cm slab, $\mathcal{E} = 1.6 \times 10^5$ V $\cdot$ m^{-1}. For ordinary fields we will thus have charge neutrality, or more precisely **charge quasi-neutrality**, even in the presence of fields and currents.

It is also interesting to consider the time variation of the charge density ρ in the simple case where we neglect recombination and diffusion. The usual charge conservation law then applies:

$$\nabla \cdot \mathbf{J} = \nabla \cdot \sigma \boldsymbol{\mathcal{E}} = -\frac{d\rho}{dt} \tag{7.10}$$

or using Eq. (7.8),

$$\frac{d\,(\Delta n - \Delta p)}{dt} = -\frac{\sigma}{\epsilon_0\,\epsilon_r}(\Delta n - \Delta p). \tag{7.11}$$

If at time $t = 0$ there is a charge imbalance $(\Delta n - \Delta p)_0$, this imbalance decreases according to

$$(\Delta n - \Delta p) = (\Delta n - \Delta p)_0 \exp(-t/\tau_0), \tag{7.12}$$

where

$$\tau_0 = \frac{\epsilon_0 \epsilon_r}{\sigma} \tag{7.13}$$

is a characteristic time for the medium called the **dielectric relaxation time**. This is the time required for the charges to screen the perturbation. This time is very short, even in a semiconductor where the conductivity is lower than in a metal. In a silicon crystal with $\sigma = 100\ \Omega \cdot \mathrm{m}^{-1}$ and $\epsilon_0 \epsilon_r \sim 10^{-10}$ $\mathrm{F} \cdot \mathrm{m}^{-1}$ we get $\tau_0 \sim 10^{-12}$ s. If a space charge is created it will therefore disappear very quickly.

We may therefore assume that in homogeneously doped materials under normal conditions **charge neutrality holds even in the presence of external excitation. This is nevertheless only an approximation, and we should really refer to it as charge quasi-neutrality.** We give a more complete treatment including diffusion in Appendix 7.1, which shows that the approximation is excellent.

We shall see in Chap. 8 that in materials with strongly inhomogeneous doping there can exist large electric fields and large net charge densities.

7.3 Injection or Extraction of Minority Carriers

We can now discuss the electron and hole currents in a semiconductor where n or p depart from their equilibrium values. We shall consider one-dimensional motion in the x direction resulting from concentration gradients along x. We consider only extrinsic materials, e.g., n type, and as an example we study the effect of irradiation by light at the end of a semiconducting bar (Fig. 7.1). If the photon energy is high enough (3 or 4 eV) the absorption coefficient is about 10^6 cm^{-1}. This means that the light intensity is multiplied by $1/e$ over 10 nm, and that electron-hole pairs are created at the surface only. Appendix 7.2 gives a more detailed treatment of this phenomenon. In steady state there is a concentration gradient near the surface.

The equations governing the phenomenon are the current equation (7.2) and the charge conservation equation (7.5). In the bulk of the material the generation rate g_p is zero and if there is no electric field in the system, substitution of Eq. (7.2) in Eq. (7.5) gives

$$\frac{\partial p_n}{\partial t} = -\frac{p_n - p_n^0}{\tau_p} + D_h \frac{\partial^2 p_n}{\partial x^2}. \tag{7.14}$$

In steady state we have

$$0 = -\frac{p_n - p_n^0}{\tau_p} + D_h \frac{\partial^2 p_n}{\partial x^2}. \tag{7.15}$$

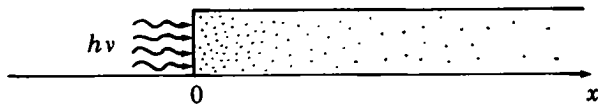


Fig. 7.1. n-type semiconductor crystal illuminated at one end.

The boundary conditions are $p_n(x = 0) = p_n^0 + \Delta p_x(0)$ and $p_n(x \to \infty) = p_n^0$. The solution of Eq. (7.15) is

$$p_n(x) = p_n^0 + \Delta p_n\,(0) \exp(-x/L_h) \tag{7.16}$$

with

$$L_h = \sqrt{D_h\,\tau_p}. \tag{7.17}$$

The length L_h is called the **minority carrier diffusion length.**

For the irradiation of Fig. 7.1 the diffusion profile is shown in Fig. 7.2. The length L_h characterizing the injection effects is relatively large. With $D_h = 25 \times 10^{-4}$ m^2/s and $\tau_p = 10^{-6}$ s we get $L_h = 5 \times 10^{-5}$ m. As there exists a hole concentration gradient there will be a diffusion current depending on x:

$$J_h(x) = -D_h\, e \frac{\partial\, \Delta p}{\partial x}. \tag{7.18}$$

At the surface

$$J_h(x = 0) = e \frac{D_h}{L_h} \Delta p_n(0). \tag{7.19}$$

Expression (7.19) is needed for an understanding of the p–n junction and the junction transistor. In the present example the "diffusion velocity" D_h/L_h is 50 m $\cdot$ s^{-1}. The electric field $\mathcal{E}$ which would give the same value of drift velocity is given by $\mathcal{E} \sim D_h/(L_h \mu_h) \sim 500$ V $\cdot$ m^{-1} for silicon. This large value shows that the transport of minority carriers may in some cases be essentially a diffusion current. This discussion is taken up more thoroughly in Appendix 7.1. The electron concentration is found from charge quasi-neutrality (7.6). In short: light creates a packet of excess holes, minority carriers, at $x = 0$. The majority carriers (electrons) surround them and screen them, so as to maintain charge quasi-neutrality in the sample.

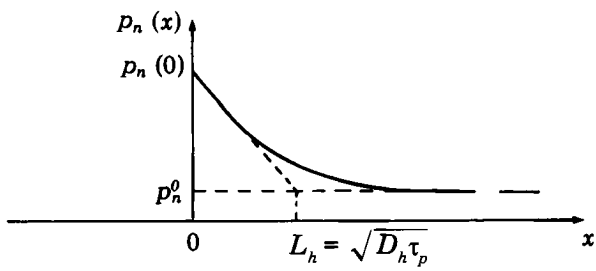


Fig. 7.2. Hole concentration profile.

Of course we get an equivalent situation for the injection of electrons in p-type material. Conversely we shall show in the next chapter that in an inhomogeneously doped semiconductor, where this screening cannot occur, charge neutrality cannot be enforced everywhere.

Appendix 7.1

Charge Quasi-Neutrality

We give here a simple example showing that the surface injection of carriers leads not only to a diffusion current of the minority carriers, but also to the existence of an electric field which appears spontaneously in the system. However this field is very small, and we can show that the deviation from charge neutrality remains very small.

We take the geometry of Fig. 7.1. We assume that irradiation creates a departure from equilibrium at the surface of n-type material. For simplicity in the following we suppress the index n on the concentrations.

We start from the charge conservation equation and the expression for the current, which are (cf. Sect. 7.1)

$$\frac{\partial p}{\partial t} = -\frac{p - p_0}{\tau_p} - \frac{\nabla \cdot \mathbf{J}}{e}, \tag{7.20}$$

$$\mathbf{J}_h = pe\mu_h \, \boldsymbol{\mathcal{E}} - eD_h \nabla p, \tag{7.21}$$

which in one dimension gives

$$\frac{\partial p}{\partial t} = -\frac{p - p_0}{\tau_p} - \mu_h \, \mathcal{E} \frac{\partial p}{\partial x} + \mu_h \, p \frac{\partial \mathcal{E}}{\partial x} + D_h \frac{\partial^2 p}{\partial x^2}. \tag{7.22}$$

Up to now we have omitted terms containing the electric field under the assumption of strict charge neutrality. We thus found the stationary solution (7.16):

$$p(x) = p + \Delta p(0) \exp(-x/L_h) \tag{7.23}$$

and

$$J_h(x) = -e \, D_h \frac{\partial p}{\partial x} = e \frac{D_h}{L_h} \Delta p(0) \exp(-x/L_h). \tag{7.24}$$

We consider the total current $\mathbf{J} = \mathbf{J}_e + \mathbf{J}_h$. From Eqs. (7.1) and (7.2):

$$\mathbf{J} = \mathbf{J}_e + \mathbf{J}_h$$
$$= n\, e\, \mu_e\, \boldsymbol{\mathcal{E}} + p\, e\, \mu_h\, \boldsymbol{\mathcal{E}} + e\, D_e\, \nabla n - e\, D_h\, \nabla p. \tag{7.25}$$

In steady state

$$\nabla \cdot \mathbf{J} = -\frac{d\rho}{dt} = 0 \tag{7.26}$$

and $\mathbf{J}$ is constant along the bar. But the current vanishes far from the surface, and therefore must vanish everywhere, so that from Eq. (7.25):

$$(n\, e\, \mu_e + p\, e\, \mu_h)\, \boldsymbol{\mathcal{E}} = e\, D_h\, \nabla p - e\, D_e\, \nabla n. \tag{7.27}$$

Equation (7.27), which is rigorous, shows that we can have strict charge neutrality—i.e., $\nabla n = \nabla p$ at each point, and thus $\boldsymbol{\mathcal{E}} = 0$ since in this case there is no charge in the system—only if the electron and hole diffusion coefficients are equal. The physical reason is that electrons and holes are created in equal numbers at the surface, but as the diffusion coefficients are different the electrons and holes diffuse differently, leading to the appearance of a space charge. This charge creates the field $\boldsymbol{\mathcal{E}}$ of Eq. (7.27) and thus allows the drift current on the left-hand side to compensate the differences between the diffusion currents.

To evaluate the order of magnitude of the electric field $\boldsymbol{\mathcal{E}}$ we note that we can neglect the hole drift current in Eq. (7.27), as there are very few holes in n-type material. We then have

$$\boldsymbol{\mathcal{E}} = \frac{1}{n\, \mu_e}(D_h\, \nabla p - D_e\, \nabla n). \tag{7.28}$$

We assume that charge neutrality holds to a first approximation, so that $\nabla n = \nabla p$ and

$$\boldsymbol{\mathcal{E}} \sim \frac{D_h}{n\, \mu_e}\left(1 - \frac{D_e}{D_h}\right) \nabla p. \tag{7.29}$$

We set

$$\frac{D_e}{D_h} = \frac{\mu_e}{\mu_h} = b \tag{7.30}$$

with b a parameter of order unity. The charge required to create the field $\boldsymbol{\mathcal{E}}$ is obtained from Poisson's equation:

$$\rho(x) = \epsilon_0\, \epsilon_r \nabla \cdot \boldsymbol{\mathcal{E}} = \frac{\epsilon_0\, \epsilon_r}{n\, \mu_e} D_h (1 - b) \frac{\Delta p\,(0)}{L_h^2} \exp(-x/L_h). \tag{7.31}$$

To obtain Eq. (7.31) we used Eq. (7.29) and the approximate solution (7.23) of the systems (7.20) and (7.21). At $x = 0$, using $L_h^2 = D_h \tau_p$:

$$\rho(0) = \frac{\epsilon_0\, \epsilon_r}{n\, \mu_e}(1 - b) \frac{\Delta p\,(0)}{\tau_p}. \tag{7.32}$$

The charge density ρ is the deviation from strict electrical neutrality:

$$\rho = (\Delta p - \Delta n)e \tag{7.33}$$

or

$$(\Delta p - \Delta n)_{x=0} = \frac{\epsilon_0\ \epsilon_r}{e\ n\ \mu_e}(1-b)\frac{\Delta p(0)}{\tau_p}, \tag{7.34}$$

which using Eq. (7.13) becomes

$$\frac{(\Delta p - \Delta n)_{x=0}}{\Delta p(0)} = \frac{\tau_0}{\tau_p}(1-b). \tag{7.35}$$

This value is very small, as the dielectric relaxation time is always very short compared with the lifetime τ_p: for example $\tau_0 \sim 10^{-12}$ s while $\tau_p \sim 10^{-6}$ s. The deviation from strict equality of the electron and hole concentrations is in this case only 10^{-6} times the deviation of the minority carriers from the equilibrium concentration. It is of course exactly zero if $b = 0$, i.e., if $D_h = D_e$ because then the electrons and holes diffuse at precisely the same speed.

Appendix 7.2

Problems on Photoexcitation, Recombination, and Photoconductivity

We consider a semi-infinite sample of n-type silicon with electron concentration n_0, steadily illuminated by light of wavelength λ or energy $h\nu$ and intensity I_0 (Fig. 7.3). The wavelength considered here is 500 nm (blue–green), with a corresponding absorption coefficient $\alpha = 10^4$ cm^{-1}.

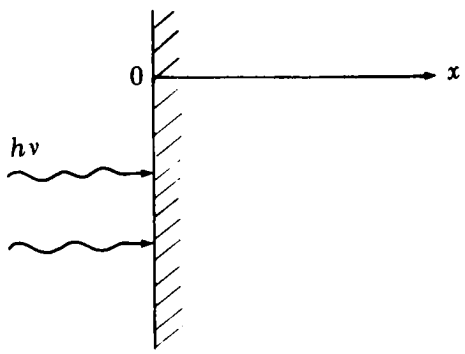


Fig. 7.3. A semiconducting sample is illuminated by photons of energy $h\nu$.

Problems

1. Why is the electron–hole pair creation rate g non-uniform in the sample? Find its variation as a function of x. Assume that each photon absorption produces an electron-hole pair.

2. Write down the differential equation for the excess hole concentration $\Delta p(x)$ (with respect to the equilibrium concentration). Show that the solution of this equation is the sum of two decreasing exponentials in x, and

ıpare their characteristic lengths. Take the hole recombination time as 1 μs and their diffusion coefficient as $D_h = 25 \times 10^{-4}$ m^2 s^{-1}.

As boundary condition assume that there is no recombination at the face of the sample, i.e., the hole current vanishes at $x = 0$. Give an ression for $\Delta p(x)$ and sketch its variation.

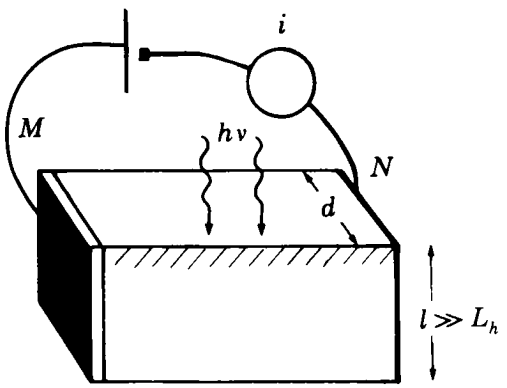
3. Calculate the surface density of excess holes, and show from the re-s of **2** that Eq. (7.16) describes the variation of $\Delta p(x)$ correctly. For = 1 W/m^2 what is the surface hole concentration? What is the surface tron concentration?

4. An external bias is used to set up an electric field $\mathcal{E}$ between the trodes M and N, in the geometry of Fig. 7.4. The thickness l of the ıple is assumed large compared with L_h, and its width is d.

Calculate the current i through the cross section of the illuminated ıiconductor, and show under the above assumptions that it exceeds the k current i_0 by

$$\Delta i = i_0 \left(1 + \frac{\mu_h}{\mu_e}\right) \frac{I_0 \, \tau}{h \, \nu \, n_0 \, l}, \tag{7.36}$$

ere μ_h and μ_e are the hole and electron mobilities, respectively. This ression does not involve the absorption coefficient α. Also give the form Δi when $h\nu$ is close to the band gap energy.



7.4. Principle of the measurement of photoconductivity.

5. Keeping the illumination and the sample as in **4** we now extract the ess carriers from the back face via an electrode P, (Fig. 7.5). Assuming ıin that $l >> L_h$, and following **3**, find the excess hole concentration (x), using the facts that the hole current vanishes at $x = 0$ and the ess holes are extracted at $x = l$. What is the current $J_p(l)$?

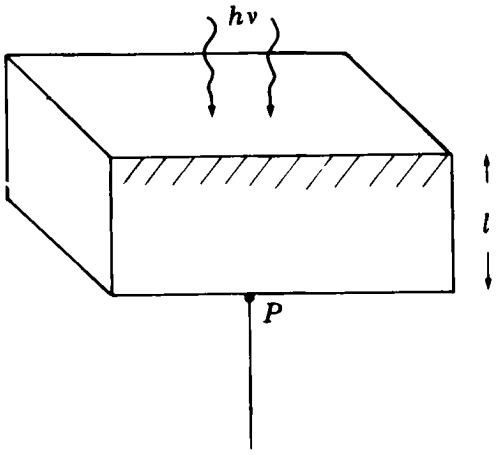


Fig. 7.5. Extraction of the excess carriers at the back of the sample. The front surface is covered with a transparent conductor such as I.T.O. (indium–tin–oxide).

Solutions

1. The intensity reaching depth x is

$$I = I_0 \exp(-\alpha x). \tag{7.37}$$

Between x and $x + \mathrm{d}x$ an intensity

$$dI = \alpha\ I_0\ dx\ \exp(-\alpha x) \tag{7.38}$$

is absorbed. We verify that $\int_0^\infty dI = I_0$. The creation energy for an electron-hole pair is $h\nu = hc/\lambda$, so that the number of pairs $g(x)\mathrm{d}x$ created per unit area between x and $x + \mathrm{d}x$ is

$$g(x)\ dx = \alpha\ I_0\ dx \exp(-\alpha x)/h\nu \tag{7.39}$$

with $g(x)$ expressed in particles/m^3 s.

2. The conservation equation for the minority carriers (holes) involves the creation rate $g(x)$:

$$\frac{\partial \Delta p\ (x)}{\partial t} = D_h \frac{\partial^2\ \Delta p(x)}{\partial x^2} - \frac{\Delta p}{\tau} + \frac{\alpha\ I_0}{h\nu} \exp(-\alpha x). \tag{7.40}$$

We seek the steady-state solution. This is the solution of a second order differential equation with constant coefficients and a function of x on the right-hand side. The solution of the homogeneous equation is

$$\Delta p\ (x) = A \exp(-x/L_h) \tag{7.41}$$

with A = constant and $L_h = (D_h \tau)^{1/2}$. In fact, when x is large compared with L_h and α^{-1} there are no more excess holes and $\Delta p(x)$ vanishes. A particular integral of Eq. (7.40) is

$$\Delta p\ (x) = C \exp(-\alpha x), \tag{7.42}$$

where C is fixed by the condition

$$C\ \left(\alpha^2\ D_h - \frac{1}{\tau}\right) + \frac{\alpha\ I_0}{h\nu} = 0, \tag{7.43}$$

$$C = \frac{\alpha\ \tau\ I_0}{h\nu\ (1 - \alpha^2\ L_h^2)}. \tag{7.44}$$

Numerically, with $D_h = 25 \times 10^{-4}$ m^2/s, $\tau = 1$ μs, we get $L_h = 50$ μm, the characteristic length for diffusion. Optical absorption occurs over a distance of order $\alpha^{-1} = 10^{-4}$ cm $= 1$ μm, much smaller than L_h. The product αL_h is much larger than 1, and C is negative:

$$C \simeq \frac{-\tau\ I_0}{\alpha\ L_h^2\ h\nu}. \tag{7.45}$$

The general solution of Eq. (7.40) is thus the sum of two terms, the first varying rapidly and the second slowly:

$$\Delta p\ (x) = C \exp(-\alpha x) + A \exp(-x/L_h), \tag{7.46}$$

where C is given by Eq. (7.44). The constant A is determined by the condition that the current should vanish at $x = 0$, with

$$J_h\ (x) = -D_h \frac{\partial \Delta p}{\partial x}, \tag{7.47}$$

$$= D_h[\alpha\ C \exp(-\alpha x) + (A/L_h) \exp(-x/L_h)]. \tag{7.48}$$

The first component of the current pulls holes out of the solid, the second pulls them into it. From $J_h(0) = 0$ we deduce

$$A = -\alpha\ C\ L_h, \tag{7.49}$$

$$\Delta p(x) = \frac{\alpha\tau\ I_0}{h\nu\ (\alpha^2\ L_h^2 - 1)}[\alpha L_h \exp(-x/L_h) - \exp(-\alpha x)]. \tag{7.50}$$

3. The excess hole density at $x = 0$ is

$$\Delta p(0) = \frac{\alpha\tau\ I_0}{h\nu\ (\alpha\ L_h + 1)}. \tag{7.51}$$

Since $\alpha L_h >> 1$, we have

$$\Delta p(0) \simeq \frac{\tau\ I_0}{h\nu\ L_h}, \tag{7.52}$$

$$\Delta p(x) \simeq \Delta p(0) \exp(-x/L_h) \tag{7.53}$$

which is Eq. (7.16). This expression correctly describes the variation of $\Delta p(x)$ for $x > \alpha^{-1}$ (cf. Fig. 7.6).

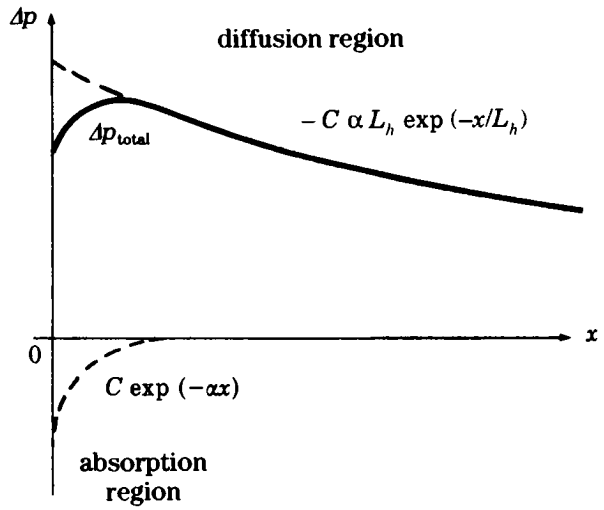


Fig. 7.6. The heavy curve shows $\Delta p(x)$, and the two dashed curves the two terms of Eq. (7.50) which build $\Delta p(x)$.

The energy of a blue–green photon of 500 nm is given by $hc/\lambda = 2.48$ eV. For $I_0 = 1$ W/m^2 we have 2.5×10^{18} photons/m^2.s. The excess surface hole concentration given by Eq. (7.52) is about 5×10^{16}/m^3. It equals the excess electron concentration.

4. The current through the cross section of the semiconductor and the electric field $\mathcal{E}$ are related by

$$i = \mathcal{E} d \int_0^l \sigma\,(x)\, dx. \tag{7.54}$$

For the dark current

$$\sigma_0 = e(n_0\, \mu_n + p_0\, \mu_p), \tag{7.55}$$

$$\sigma_0 \simeq e\, n_0\, \mu_n \tag{7.56}$$

since the semiconductor is n type. When illuminated,

$$\Delta\sigma\,(x) = e\, \mu_n\, \Delta n\,(x) + e\, \mu_p\, \Delta p\,(x), \tag{7.57}$$

$$\Delta\sigma\,(x) = e\, \Delta p\,(x)\,(\mu_n + \mu_p), \tag{7.58}$$

$$\frac{\Delta i}{i_0} = \frac{i - i_0}{i_0} = \frac{\int_0^l \Delta\sigma\,(x)\, dx}{l\, \sigma_0}, \tag{7.59}$$

$$\frac{\Delta i_0}{i_0} = \frac{\mu_e + \mu_h}{\mu_e} \frac{\int_0^l \Delta p\,(x)\, dx}{l\, n_0}. \tag{7.60}$$

The integral in the numerator can be found from Eq. (7.50). Because $\alpha^{-1} << L_h << l$,

$$\int_0^l \Delta p\,(x)\,dx \simeq \int_0^\infty \Delta p\,(x)\,dx = -C\left(\alpha\,L_h^2 - \frac{1}{\alpha}\right). \tag{7.61}$$

Since $\alpha L_h >> 1$, we get

$$\frac{\Delta i_0}{i_0} \simeq \left(1 + \frac{\mu_h}{\mu_e}\right)\frac{\tau I_0}{h\nu}\frac{1}{n_0 l}. \tag{7.62}$$

This expression involves only the photon number $I_0/h\nu$, and not the variation of α with $h\nu$. This comes from the approximation $\alpha L_h >> 1$. We have detector of light that is fairly insensitive to photon energies.

In contradistinction, when we use photons with energy close to the band ap, α is small, and we can have $\alpha^{-1} > l$, where L_h is not modified. Then $_h << l << \alpha^{-1}$, $\alpha L_h << 1$,

$$\int_0^l \Delta p\,(x)\,dx = -C\,\alpha\,L_h^2 + C\frac{1-\exp(-\alpha l)}{\alpha}, \tag{7.63}$$

$$\simeq -C\,\alpha\,L_h^2 + Cl \simeq Cl. \tag{7.64}$$

Moreover C is now positive:

$$C \simeq \frac{\alpha\,\tau\,I_0}{h\nu} \tag{7.65}$$

o that the photoconductivity signal

$$\Delta i_0 \simeq \frac{i_0}{n_0}\left(1 + \frac{\mu_h}{\mu_e}\right)\frac{\alpha\,\tau\,I_0}{h\nu} \tag{7.66}$$

s proportional to αI_0 in the energy range where α is small. The photoonductivity spectrum, i.e., the curve $\Delta i_0(h\nu)$, for a source with photon umber varying weakly with $h\nu$, will exhibit the variation of α close to the bsorption threshold, and in particular the exciton peaks (see Sect. 6.1d).

5. The general steady solution of the differential Eq. (7.40) now has the orm

$$\Delta p\,(x) = A\exp(-x/L_h) + B\exp(x/L_h) + C\exp(-\alpha x), \tag{7.67}$$

here C is given by Eq. (7.44), and A and B are determined by the boundry conditions

$$J_h = 0, \quad \text{i.e.,} \quad D_h[(A-B)/L_h + C\alpha] = 0 \quad \text{at} \quad x = 0; \tag{7.68}$$

$$\Delta p = 0, \quad \text{i.e.,} \quad A\exp(-l/L_h) + B\exp(l/L_h) + C\exp(-\alpha l) = 0$$
$$\text{at} \quad x = l. \tag{7.69}$$

We still have $\alpha^{-1} << L_h << l$, hence $\exp(-\alpha l) << \exp(-l/L_h) << 1$, o that

$$A \simeq -C\alpha\,L_h\exp(l/L_h)\;/\;[\exp\,(l/L_h) + \exp(-l/L_h)], \tag{7.70}$$

$$B \simeq C\,\alpha\,L_h\exp(-l/L_h)\;/\;[\exp\,(l/L_h) + \exp\,(-l/L_h)], \tag{7.71}$$

hence

$$J_h\ (l) = \frac{D_h}{L_h}[A\exp(-l/L_h) - B\exp(l/L_h) + \alpha\ L_h\ C\exp(-\alpha l)], \quad (7.72)$$

$$J_h\ (l) \simeq -2\ D_h\ C\ \alpha/[\exp(l/L_h) + \exp(-l/L_h)]. \quad (7.73)$$

Replacing C by its expression in the same approximation we get

$$J_h\ (l) \simeq 2\frac{D_h\ \alpha^2\ \tau\ I_0}{\alpha^2\ L_h^2\ h\nu[\exp\ (l/L_h) + \exp(-l/L_h)]},$$

$$J_h\ (l) \simeq\ 2\frac{I_0}{h\nu}\exp(-l/L_h). \quad (7.74)$$

The current collected at the back face of the illuminated sample is a diffusion current. The situation is comparable with that of a p–n–p transistor (see Chap. 10) in which holes are injected from the emitter to the base and extracted at the collector.

The p–n Junction

1 Introduction: Inhomogeneous Semiconductors

the remaining chapters we shall discuss the physics of various "electron achines" such as diodes, transistors, etc. These devices are now vital for eryday life and for the industrial development of the second half of the th century. They mostly consist of semiconducting structures which are homogeneous either through their chemistry or because of doping. While assical mechanical machines are assembled from variously shaped compo- nts, the "electron machines" we will study are manufactured essentially spatially varying the concentration of impurities.

The p–n junction is a vital component of all these devices, both be- use of its direct applications and because an understanding of its physics plains the junction transistor and many other devices. A p–n junction nsists of a semiconducting crystal whose concentration of shallow impu- ties varies with x so that a p-type region, where $N_a - N_d > 0$, is next to n-type region, i.e., to one where $N_a - N_d < 0$. This is therefore an inho- ogeneous semiconductor, where the electron and hole concentrations vary ith position, even in thermodynamic equilibrium, i.e., in the absence of a rrent. Now, if there is a concentration gradient there must be a diffusion rrent. For the electrons, for example, this is

$$\mathbf{J}_{e,d} = e\, D_e\, \nabla n.$$

For the electron current to vanish, the drift current $ne\,\mu_e\,\mathcal{E}$ must exactly ncel it:

$$n\,\mu_e\,\mathcal{E} = -D_e\,\nabla n. \tag{8.1}$$

There must therefore be an electric field, which in the absence of external harges can only result from a charge density ρ within the system. This ows that:

In thermodynamic equilibrium, in an inhomogeneously doped system, charge neutrality cannot be preserved everywhere.

Once charge neutrality is violated an electrostatic potential $V(x)$ appears, given by Poisson's equation

$$\Delta V(x) = -\frac{\rho\,(x)}{\epsilon_r\,\epsilon_0} \tag{8.2}$$

with

$$\rho(x) = e[\,p\,(x) - n\,(x) + n_d^+\,(x) - n_a^-\,(x)\,], \tag{8.3}$$

where the concentrations of free carriers n, p and ionized impurities n_d^+, n_a^- now depend on position.

This macroscopic potential is added to the usual crystalline potential and shifts the energy levels by an amount $-eV(x)$. The general shape of the bands is shown in Fig. 8.1.

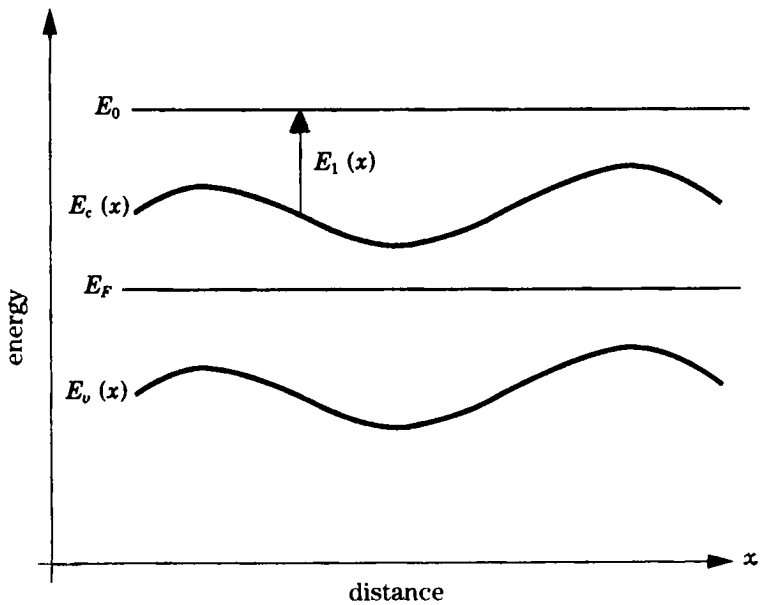


Fig. 8.1. Band shape in an inhomogeneous semiconductor at thermodynamic equilibrium.

For a system in thermodynamic equilibrium the chemical potential, or Fermi level, which is the partial derivative of the thermodynamic potential with respect to the particle number, is spatially constant.

We have to consider an energy origin E_0 which is independent of the charge distribution, for example, the energy of a static electron at infinity (this is the zero of energy in the hydrogen atom, compared with which the ground state has energy -13.6 eV). Then if the energy of a given

conduction-band electron is E and $E_1(x)$ is its energy with respect to the bottom of the conduction band, we have

$$\begin{aligned} E &= E_c(x) + E_1(x) \\ &= E_{c_0} - eV(x) + E_1(x) \end{aligned} \tag{8.4}$$

and the Fermi function is

$$f = \frac{1}{1 + \exp[(E - E_F)/kT]} = \frac{1}{1 + \exp[(E_c + E_1 - E_F)/kT)]}. \tag{8.5}$$

The number of conduction electrons at abscissa x is then given by

$$\begin{aligned} n\ (x) &= \int_{E_c(x)}^{\infty} n[E - E_c\ (x)]\ f\ dE \\ &= \int_0^{\infty} \frac{n\ (E_1)\ dE_1}{1 + \exp[(E_1 + E_c(x) - E_F)/kT]}. \end{aligned} \tag{8.6}$$

There will be analogous Eqs. (8.6′), (8.6″)... giving $p(x), n_d^0(x)$, $n_a^-\ (x)$ as we have seen in Chap. 4. The total charge density is given by Eq. (8.3). Solving the equilibrium junction problem then reduces to solving Eqs. (8.2), (8.3), (8.6), (8.6′),... so as to find the potential distribution and the carrier densities as functions of the coordinates.

When the Fermi level is far from the band levels (see Sect. 4.4), Eq. (8.6) gives for each n:

$$n\ (x) = N_c \exp[-(E_c(x) - E_F)/kT], \tag{8.7}$$

$$n\ (x) = \text{const.} \exp[e\ V(x)/kT], \tag{8.8}$$

and

$$\frac{1}{n}\nabla n = \frac{e}{kT}\nabla V = -\frac{e}{kT}\boldsymbol{\mathcal{E}}. \tag{8.9}$$

Substituting in Eq. (8.1) we recover the Einstein relation (5.30). There is of course an equivalent result for the holes. Formula (8.7) shows that the electrons are more numerous in regions where the Fermi level is closer to the bottom of the conduction band.

8.2 The Equilibrium p–n Junction

In the following we shall limit ourselves to what is called an abrupt junction. It consists of a semiconducting crystal in which the impurity concentration changes discontinuously from a majority concentration of donors to a majority concentration of acceptors at $x = 0$. Of course in reality it is impossible to set up such a system and the transition from the n to the p region is fairly gradual, as shown schematically in Fig. 8.2(a), but we shall

see that we can often approximate this type of profile by a discontinuous one (Fig. 8.2(b)).

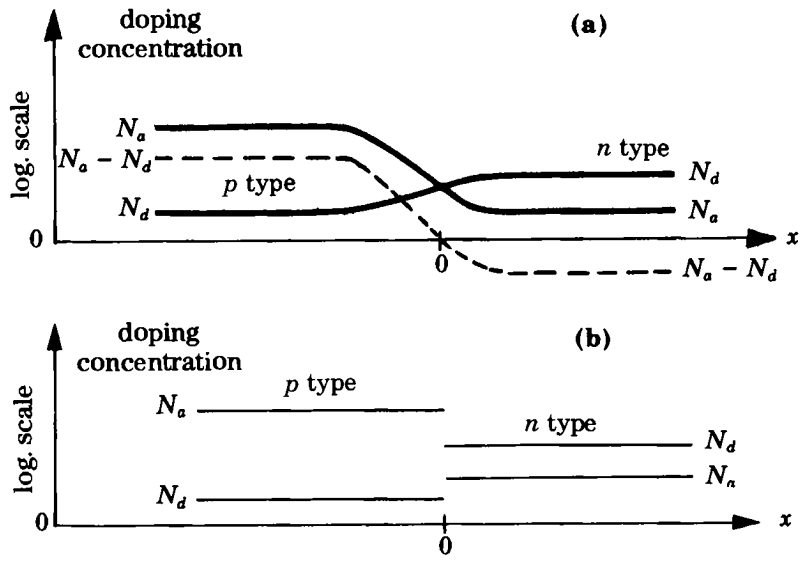


Fig. 8.2. Doping profile of a p–n junction: (a) real profile; (b) approximation as a discontinuous profile.

8.2a Qualitative Discussion

Let us assume that we have made an abrupt junction by bringing together two crystals of different types. The band diagram of the two separate systems is shown in Fig. 8.3. The Fermi levels in the two crystals are different. We assume we are in the saturation regime where $n = N_d$ and $p = N_a$, with the Fermi levels in the band gap; the materials are assumed to be non-compensated (this simplifies the notation without loss of generality in the following).

We denote by n_p^0 and p_p^0 the electron and hole concentrations in the p-type material and n_n^0, p_n^0 the corresponding concentrations in the n-type material, before they are brought into contact. The carrier concentrations are given by Eqs. (4.14), (4.16), (4.31), and (4.37):

$$p_p^0 = N_a = N_v \exp[-(E_{F,p} - E_{v,p})/kT], \tag{8.10}$$

$$n_p^0 = n_i^2/p_p^0 = N_c \exp[-(E_{c,p} - E_{F,p})/kT], \tag{8.11}$$

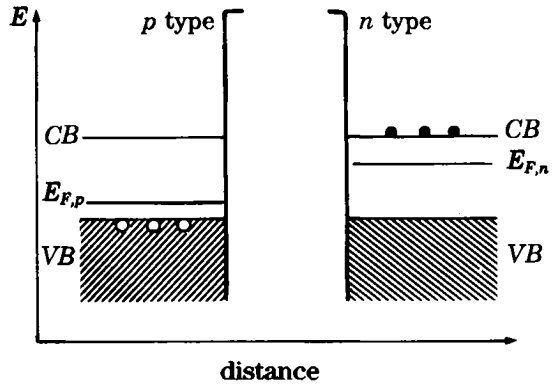


Fig. 8.3. Energy levels in two separated n- and p-type crystals. The hatched zones represent filled states, the heavy dots conduction electrons, and the open dots holes. (After Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

$$p_n^0 = n_i^2/n_n^0 = N_v \exp[-(E_{F,n} - E_{v,n})/kT], \tag{8.12}$$

$$n_n^0 = N_d = N_c \exp[-(E_{c,n} - E_{F,n})/kT], \tag{8.13}$$

$$n_n^0\, p_n^0 = n_p^0\, p_p^0 = N_c\, N_v \exp(-E_g/kT) = n_i^2. \tag{8.14}$$

As the chemical potentials differ from left to right, a reaction will occur if the two systems are brought into contact to make a junction; electrons will pass from the system where the chemical potential is higher (on the right) towards the p region on the left, and the holes will move from left to right. This movement of free carriers may also be described in an exactly equivalent fashion by saying that the concentration gradient creates a diffusion current of electrons from right to left and of holes from left to right. The additional electrons in the p material are the excess minority carriers: they recombine with the majority holes, with a similar effect for the additional holes in the n region. The electrons leave ionized donors of concentration N_d in the n material, and the holes leave ionized acceptors of concentration N_a in the p material. A region of non-zero charge is then created near the junction. This is called the space-charge region and is shown in Fig. 8.4.

However, this process cannot continue indefinitely as the space charge creates an electric field that opposes the diffusion of majority carriers. This is how equilibrium is achieved. The space-charge zone is a region where the number of free carriers is very low; for this reason it is also called the depletion region. The space-charge zone bears a charge, namely that of the fixed donor or acceptor sites that are no longer neutralized.

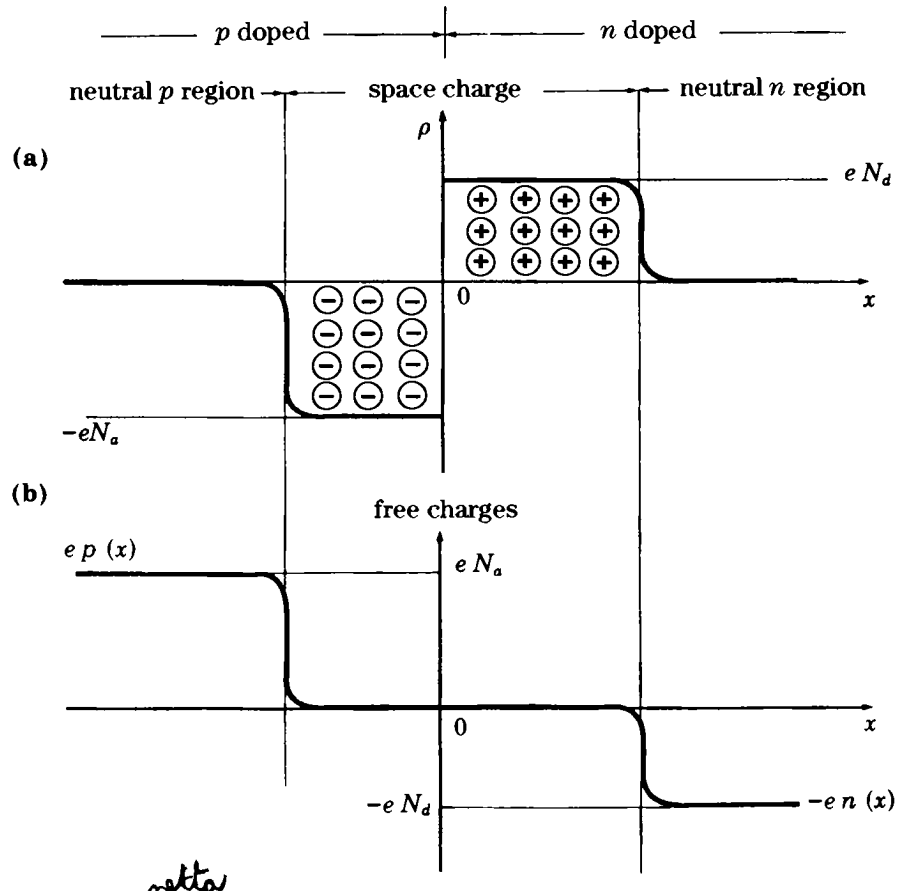


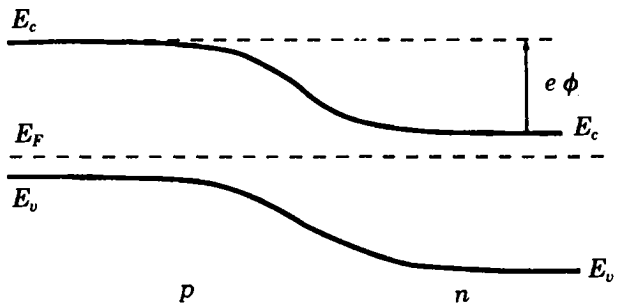
Fig. 8.4. (a) Net charge density $\rho(x)$, and (b) free charge density, in a p–n junction.

8.2b Potential Difference between the p and n Regions

Equilibrium thermodynamics gives the internal electrostatic potential difference ϕ between the p and n regions. The overall shift of the bands in the junction region must have the shape shown in Fig. 8.5 as in equilibrium the chemical potential must be constant throughout the system. Far from the junction the n and p materials must retain their original properties, i.e., constant free carrier concentrations, equal to the initial donor and acceptor concentrations ($p_p^0 = N_a, n_n^0 = N_d$), and constant electrostatic potential as the current must vanish.

If the electrostatic potentials in the n and p regions far from the junction are V_n and V_p, then setting

$$\phi = V_n - V_p, \tag{8.15}$$



8.5. Band profile across an equilibrium p–n junction.

ıave

$$E_{c,p} - E_{c,n} = -e\,(V_p - V_n) = e\,\phi \tag{8.16}$$

using Eqs. (8.10), (8.11), (8.13), and (8.14),

$$\phi = \frac{E_g}{e} + \frac{kT}{e} \log \frac{N_a\,N_d}{N_c\,N_v}. \tag{8.17}$$

The internal potential ϕ is thus slightly smaller than the band gap. In on at room temperature $E_g = 1.12$ eV; $N_c \sim 3 \times 10^{25}$ m^{-3}; $N_v \sim 10^{25}$ '; $kT/e = 0.025$ V; if $N_a = 3 \times 10^{23}$ m^{-3}, $N_d = 10^{21}$ m^{-3} we have $\phi =$ 5 V. We note that we also have from Eqs. (8.10)–(8.13),

$$\frac{n_p^0}{n_n^0} = \frac{p_n^0}{p_p^0} = \exp(-e\,\phi/kT). \tag{8.18}$$

: Calculation of the Space Charge

from the junction electrical neutrality holds between mobile and fixed ges. In contrast, close to the junction, the Fermi level is far from the ds. This is why there are few mobile charges here (see Eqs. (8.10)–3)). Further, the density of carriers given by Eq. (8.8) varies exponen-y, i.e., very rapidly with the potential. This means that if the potential the shape shown in Fig. 8.6(a) the space-charge density given by Eq.) will have the shape shown by the thick curves in Fig. 8.6(b).

In Sect. 8.1 we set ourselves the aim of integrating the Poisson equation the potential with a local charge depending on this potential through laws of statistical mechanics. We see that to a good approximation we regard ρ as constant and equal to $-eN_a$ in the p region over the range veen $-d_p$ and 0, and equal to $+eN_d$ between 0 and d_n; it remains to ermine the distances d_p, d_n. We replace the real charge density curve of

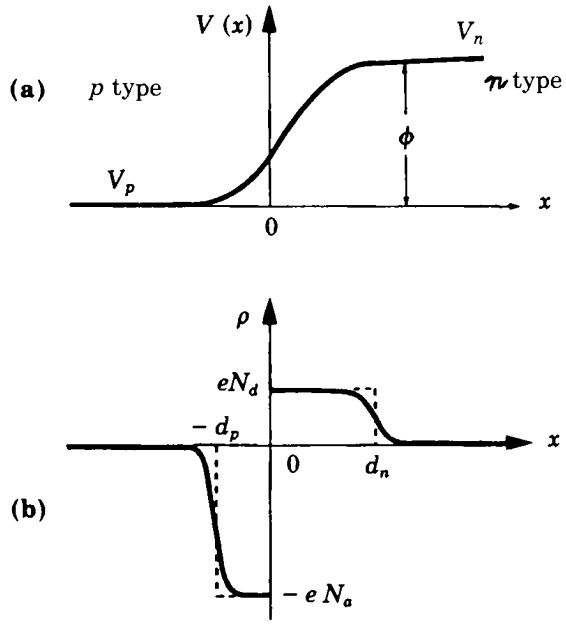


Fig. 8.6. (a) Electrostatic potential across an equilibrium p–n junction; (b) corresponding charge density. The dashed curve is the approximate form of ρ used in the calculation in the text.

Fig. 8.6(b) by the dashed curve. We then have to integrate the very simple equations

$$-\frac{\partial^2 V}{\partial x^2} = \frac{\partial \mathcal{E}}{\partial x} = -\frac{e\,N_a}{\epsilon_0\,\epsilon_r} \qquad \text{for} \quad -d_p \le x \le 0, \tag{8.19}$$

$$-\frac{\partial^2 V}{\partial x^2} = \frac{\partial \mathcal{E}}{\partial x} = +\frac{e\,N_d}{\epsilon_0\,\epsilon_r} \qquad \text{for} \quad 0 \le x \le d_n. \tag{8.20}$$

Noting that the electric field vanishes in equilibrium for $x = -d_p$ and d_n, since the current is zero at these points, we get

$$\mathcal{E} = -\frac{e\,N_a}{\epsilon_0\,\epsilon_r}(x + d_p) \quad \text{for} \quad -d_p \le x \le 0, \tag{8.21}$$

$$\mathcal{E} = \frac{e\,N_d}{\epsilon_0\,\epsilon_r}(x - d_n) \quad \text{for} \quad 0 \le x \le d_n. \tag{8.22}$$

We note that at $x = 0$:

$$N_a\,d_p = N_d\,d_n. \tag{8.23}$$

s equation states that the total space charge of the p region exactly pensates that of the n region. It reflects the global neutrality of the tion, since the "manufacture" began with two electrically neutral re- s. A second integration gives the potentials, with $V = V_p$ for $x = -d_p$:

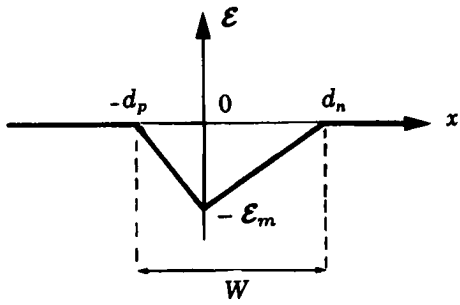
$$V(x) = \frac{e}{2\,\epsilon_0\,\epsilon_r} N_a (x + d_p)^2 + V_p \quad \text{for} \quad -d_p \le x < 0, \tag{8.24}$$

$$V(x) = \frac{e\,N_a}{2\,\epsilon_0\,\epsilon_r} d_p^2 - \frac{e\,N_d}{\epsilon_0\,\epsilon_r}\left(\frac{x^2}{2} - d_n\,x\right) + V_p \quad \text{for} \quad 0 \le x \le d_n. \tag{8.25}$$

$x = d_n$ we have $V = V_n$ and

$$\phi = (V_n - V_p) = \frac{e}{2\,\epsilon_0\,\epsilon_r}[N_a\,d_a^2 + N_d\,d_p^2]. \tag{8.26}$$

The quantities d_n, d_p are found from Eqs. (8.17) and (8.23). The form of electric field is shown in Fig. 8.7. It is negative, repelling the electrons ing from the n-type region on the right and the holes coming from p-type region on the left, and thus separates the free carriers from the e-charge region.



8.7. Electric field across an equilibrium p–n junction.

The width of the space-charge region can be found from Eqs. (8.23) and 3):

$$W = d_p + d_n = \left(\frac{2\,\epsilon_0\,\epsilon_r\,\phi}{e}\right)^{1/2}\left(\frac{N_a + N_d}{N_a\,N_d}\right)^{1/2}. \tag{8.27}$$

With $\phi \sim 1$ V; $N_d, N_a \sim 10^{21}$ m^{-3}; $\epsilon_0\epsilon_r \sim 10^{-10}$ F·m^{-1} we get $W \sim 10^{-6}$ m. If $|\mathcal{E}|$ is the modulus of the maximum field at the junction,

$$|\mathcal{E}_m| = \frac{e\,N_d\,d_n}{\epsilon_0\,\epsilon_r} = \frac{e\,N_a\,d_p}{\epsilon_0\,\epsilon_r}. \tag{8.28}$$

We note that the triangle in Fig. 8.7 has area ϕ, or

$$\phi = \frac{1}{2}|\mathcal{E}_m|(d_n + d_p) = \frac{1}{2}|\mathcal{E}_m|W, \tag{8.29}$$

where $W = (d_n + d_p)$ is the width of the space-charge region.

Let us consider the example of a junction where $N_a >> N_d$, called a $p^+ - n$ junction; then $d_p << d_n$, and the space-charge region mainly extends into the less doped region, with $W \sim (2\epsilon_0\epsilon_r\phi/eN_d)^{1/2}$. The maximum electric field in the less doped region is of the order of ϕ/W or about 10^6 V/m for a junction where the space charge extends over 1 μm.

For the "abrupt-junction" approximation used in this section to hold we require that in a real junction the region over which the doping varies from p type to n type should be much smaller than the size W we have just calculated. Suppose that the junction was made by diffusion of phosphorus into silicon originally of p type. If the diffusion operates for time t, the width of the profile is of the order of $L = (D_{\text{phosph}}t)^{1/2}$, where D_{phosph} is the diffusion coefficient of phosphorus. For phosphorus in silicon at 1000° C we have $D_{\text{phosph}} \sim 10^{-17}\text{m}^2 \cdot \text{s}^{-1}$. If diffusion lasts about 10^3 s, we get $L = 10^{-7}$ m: one can therefore manufacture junctions where the width of the space-charge region is much larger than that of the region where the doping changes. Such junctions can reasonably be regarded as abrupt.

A real junction is never made by assembling two different crystals (n- and p-type) because their surface properties (surface states to be discussed in Sect. 9.4) would totally perturb the above situation.

8.2d Currents in the Equilibrium Junction

As we have seen above, the total current through the junction vanishes because the chemical potential is constant. This is true for the electron and hole currents separately. It is interesting to understand the mechanism by which each of these currents cancels out. In the uniform regions where the charge density is zero, left and right of the deserted region, the current of each type of carrier vanishes because the electric field is zero. By contrast there is a very strong electric field in the deserted zone. Although the number of carriers is small, the drift current for each type is large, and the total current vanishes because there is an equally large opposing diffusion current. The diffusion current is large because the concentration gradient is large in this region (Eq. (8.9)). It is interesting to find the order of magnitude of these currents. For simplicity let us consider a symmetrical junction. Then at $x = 0$ the Fermi level is exactly in the middle of the band gap and the concentrations are the intrinsic ones. The drift current $J = n_i e\mu\mathcal{E}$ is then about 0.1 A/cm^2 for silicon where $n_i \sim 10^{16}$ m^{-3}; $\mu \sim 0.1$ m$^2 \cdot$ V^{-1} $\cdot$s^{-1} and for an electric field of about 10^6 V$\cdot$m^{-1} as we have just seen. This is the typical value for the drift current at the center of the junction. This reasoning holds for electrons and holes.

In summary, in equilibrium, far from the junction, the current vanishes because the carriers have a constant concentration and there is no electric

field. In the center of the space-charge region two very large currents, the drift and diffusion currents, cancel exactly.

8.3 The Non-Equilibrium Junction

Let us now assume that we apply an external voltage by means of a dc generator connected across the junction. We say that the junction is biased. The external voltage V_e is counted positive if it tries to make the p-doped side positive with respect to the n-doped side, and thus tends to decrease the height of the potential barrier between the two regions. The chemical potential cannot be constant in the whole system and a current will circulate. However the space-charge region has very high resistivity since it has few free carriers. We can thus say that the potential drop will occur in the space-charge region, the width of this zone possibly being modified by the presence of the external potential. Outside this region of very high resistivity the p and n regions can be regarded as approximately equipotential as their resistance is low.

The effect of applying an external potential V_e is shown schematically in Fig. 8.8. The distances d_n and d_p are modified by the existence of V_e and become d'_n and d'_p, which are both functions of V_e.

A calculation analogous to that of Sect. 8.2c gives

$$\phi - V_e = \frac{e}{2\,\epsilon_0\,\epsilon_r}[N_a\,d'^2_p + N_d\,d'^2_n] \tag{8.30}$$

and

$$N_a\,d'_p = N_d\,d'_n. \tag{8.31}$$

We note that **the electric fields and size of the space-charge region are modified without changing their order of magnitude.**

8.3a Energy Level Diagram

To understand the operation of a biased junction, it is useful to consider the energy diagram in the three cases of zero, forward, and reverse bias, as shown in Figs. 8.9(a), 8.9(b), and 8.9(c). Applying a potential difference across the junction means imposing a difference in chemical potential between left and right. We can do this in practice by making two contacts of the same metal at the extremities of the junction and holding them at different potentials.

The band profiles are arcs of parabolas as shown by Eqs. (8.24) and (8.25).

Outside the space-charge region, the p and n conductors remain equipotential (apart from the small variation corresponding to the ohmic drop in the two materials when a current flows). We then have

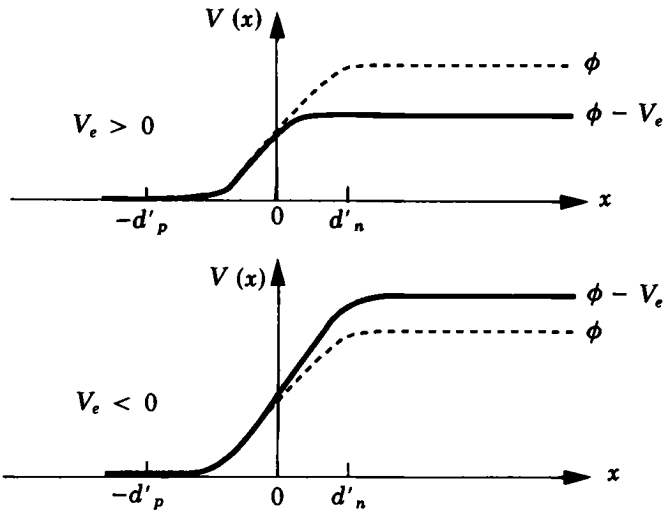


Fig. 8.8. Profile of the electrostatic potential of the junction for an applied positive or negative voltage V_e. The dashed curves show the potential for $V_e = 0$.

$$E_{F,p} - E_{F,n} = -e\, V_e. \tag{8.32}$$

8.3b Calculation of the Current

To calculate the current we make use of our description of the equilibrium currents in the unbiased diode. We saw that in the space-charge region **two very large currents** (drift and diffusion) exactly cancel. To calculate the current in the presence of an external voltage we assume first that the resultant current is small compared with the opposing components of the equilibrium current. Then in the current equation for (for example) the electrons

$$\mathbf{J}_e = n\, e\, \mu_e\, \mathcal{E} + e\, D_e \nabla\, n, \tag{8.33}$$

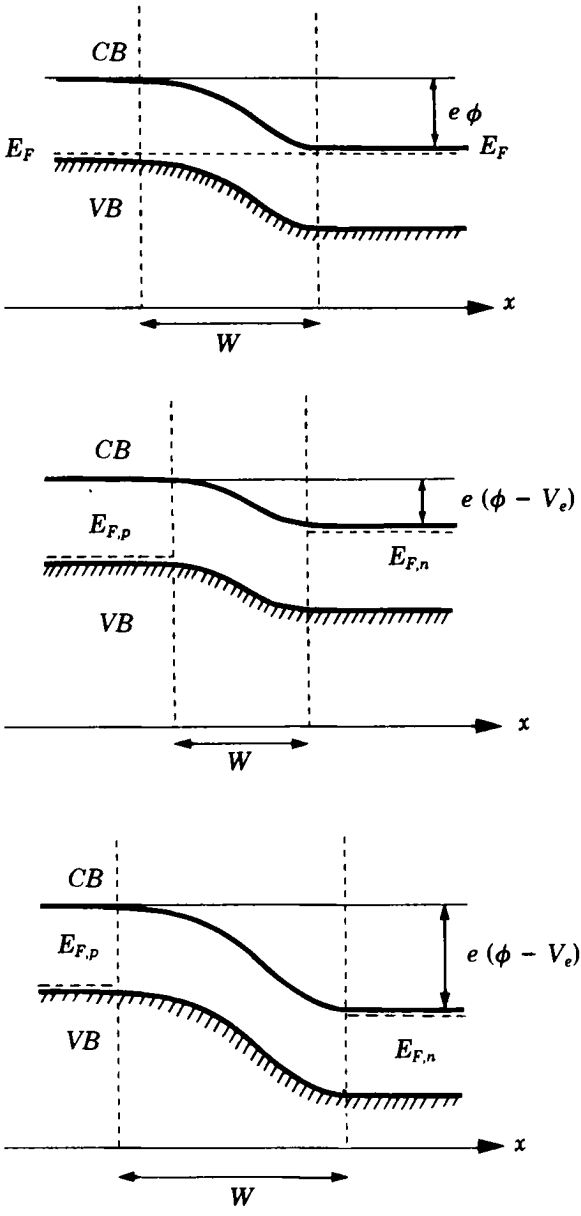
we can neglect $\mathbf{J}_e$ compared with the two terms on the right-hand side and write in the presence of currents

$$n\, e\, \mu_e\, \mathcal{E}\, \# - e\, D_e \nabla\, n \tag{8.34}$$

or, using the Einstein relations,

$$n\,(x) \# \text{ const. } \exp \frac{e\, V\,(x)}{kT}. \tag{8.35}$$

The latter equation shows that for currents that are not too large (we will specify this later) the electrons are in thermal equilibrium in the space-charge region. Then using Eq. (8.35) at the points $-d'_p$ and d'_n, we have



of band profiles for a p–n junction which is (a) unbiased; (b)
) reverse biased.

$$\frac{n_{(x=-d'_p)}}{n_{(x=+d'_n)}} = \exp\left[-e(\phi - V_e)/kT\right]. \tag{8.36}$$

But the electron concentration in the n-type material, for $x \geq d'_n$, has the fixed value N_d, so that

$$n_{(x=-d'_p)} = N_d \exp\left[-e(\phi - V_e)/kT\right]. \tag{8.37}$$

Using Eq. (8.18),

$$n_{(x=-d'_p)} = n_p^0 \exp \frac{e\,V_e}{kT}. \tag{8.38}$$

This shows that in the presence of the external voltage the concentration of minority carriers is modified at the edge of the space-charge region. For $x = -d'_p$ we have a deviation from the equilibrium population of amount

$$\Delta\, n_{(x=-d'_p)} = n_{(x=-d'_p)} - n_p^0 = n_p^0 \left[\exp\left(\frac{e\,V_e}{kT}\right) - 1\right]. \tag{8.39}$$

Depending on whether V_e is positive or negative, the concentration of minority carriers is increased or decreased. We thus say that there is **injection** or **extraction of minority carriers**. This injection or extraction takes place from the majority carrier region. We can say that the concentration deviation is created by the change in the height of the potential barrier, which changes the number of electrons able to diffuse from the n region towards the semiconducting p region. We expect this number to depend on V_e through a Boltzmann factor. The p region behaves as a semiconducting strip whose surface population of minority carriers is maintained out of equilibrium. There is therefore diffusion and an associated diffusion current. This situation was considered in the preceding chapter (Sect. 7.3 and Appendix 7.2). The current is given by Eq. (7.19), or here, for $x = -d'_p$:

$$J_e = e\,\Delta\, n_{(x=-d'_p)} \cdot \frac{D_e}{L_e}, \tag{8.40}$$

$$J_e = e\, n_p^0 \frac{D_e}{L_e} \left[\exp\left(\frac{e\,V_e}{kT}\right) - 1\right]. \tag{8.41}$$

Similarly for the holes, analogous reasoning gives the diffusion current at the edge of the n region, so that for $x = +d'_n$:

$$J_h = e\, p_n^0 \frac{D_h}{L_h} \left[\exp\left(\frac{e\,V_e}{kT}\right) - 1\right]. \tag{8.42}$$

If there is no recombination in the space-charge region these two currents are constant in this zone and add. The total current is therefore

$$J = J_e + J_h$$

$$= e\left(n_p^0 \frac{D_e}{L_e} + p_n^0 \frac{D_h}{L_h}\right)\left[\exp\left(\frac{e\,V_e}{kT}\right) - 1\right], \tag{8.43}$$

$$= J_s\left[\exp\left(\frac{e\,V_e}{kT}\right) - 1\right]. \tag{8.44}$$

The quantity J_s is called the **saturation current**:

$$J_s = e\left(n_p^0 \frac{D_e}{L_e} + p_n^0 \frac{D_h}{L_h}\right) = e\,n_i^2\left(\frac{D_e}{L_e\,N_a} + \frac{D_h}{L_h\,N_d}\right). \tag{8.45}$$

Because of the factor n_i^2, J_s varies very rapidly with temperature. The turation current is small compared with the current components at the nter of the barrier. Taking $N_a = N_d = 10^{21}$ m^{-3} and $D/L \sim 5$ m$\cdot$ s^{-1} get $J_s \sim 10^{-11}$ A/cm^2 for a silicon junction, or six orders of magnitude low the drift current at the center of the barrier. This order of magnide justifies the hypotheses made in writing Eq. (8.37). Even for forward as ($V_e > 0$) the law (8.44) will be correct as long as $J_s \exp(eV_e/kT)$ reains small enough compared with the drift current at the center of the rrier, and thus remains applicable over a very wide range of current (from $^{-11}$ A/cm^2 up to about 10^{-3} A/cm^2 in the example considered).

The law (8.44) is called Shockley's law, and dates from 1949. Its form shown in Fig. 8.10. When V_e is very large and negative we say that the ltage is reversed and the current is $-J_s$. For $V_e > 0$ we say that the ltage is direct; the current is then positive and increases exponentially th the applied voltage. This behavior is shown in Figs. 8.10(a) and 8.10(b) large and small scales. Figure 8.10(b) is drawn for $J_s = 2.19 \times 10^{-11}$ /cm^2. Notice the difference in the scales of the ordinates for direct and verse voltages.

It is interesting to note that for a strong reverse bias the electrons, the inority carriers in the p region, that reach the barrier are accelerated the electric field. They cross the space-charge region from right to left, ile the electrons in the n region, where they are the majority carriers, nnot diffuse into the p region because they cannot cross the very high tential barrier. Then for $x = -d_p'$ the electron concentration falls to zero d there is a concentration gradient for $x < -d_p'$. We thus have a semiinducting strip in which we have created a deviation $\Delta n_p = -n_p^0$ from e equilibrium populations at $x = -d_p'$. In this situation electron-hole irs are generated to compensate this deviation and diffusion results. The ctron contribution to the saturation current can then be interpreted as e maximum electron current one can draw from p-type material by exacting all the minority electrons at its surface. This current can only flow ntinuously if there is continuous reappearance of the minority electrons d thus generation of electron-hole pairs by the material. For this reason is current is often called the "electron generation current." Similarly the her component of J_s is often called the "hole generation current."

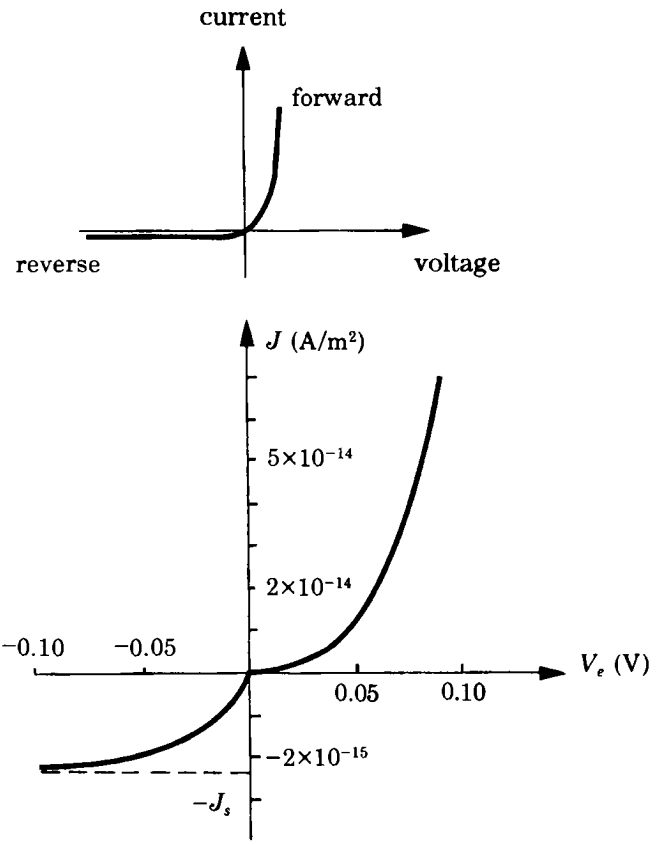


Fig. 8.10. Current–voltage characteristic of a *p–n* junction. (a) Linear scale; (b) enlarged view of the region near the origin. The current density scale is expanded by a factor 10 for the reverse region, causing the change of slope at the origin.

We have assumed that recombination is negligible in the space-charge region. To justify this we note that if the mean electric field in this zone is of the order of 10^6 V/m an electron of mobility 0.1 $m^2 \cdot V^{-1} \cdot s^{-1}$ has a drift velocity of 10^5 m/s and the region of 10^{-6} m thickness is crossed by the minority carriers in 10^{-11} s, much shorter than the recombination time. By contrast the diffusion of the majority carriers takes a time of order W^2/D or about 10^{-9} s. The recombination time may not always be extremely long compared with this time and the above theory must be modified.

Shockley's equation shows that the *p–n* junction is an efficient rectifier, as the current depends very strongly on the sign of V_e; for this reason it is often called a *p–n* diode. In real diodes there are several limitations which we have neglected:

– we assumed generation and recombination of pairs to be negligible in the space-charge region;

– we limited ourselves to moderate currents;

– we neglected the sides of the junction, where surface recombinations occur.

These effects, and the resistance of the p and n regions, may modify the current–voltage characteristics of the diode.

8.3c Current and Concentration Distributions

We note from Eq. (8.39) that if V_e is positive Δn is > 0: We say that minority carriers are injected. If the potential is strictly constant (by charge neutrality) for $x < -d'_p$ and $x > d'_n$ the current at the edge of the space charge is purely a diffusion current. This is shown in Fig. 8.11.

In a steady state we have

$$\nabla \cdot \mathbf{J} = -\frac{d\rho}{dt} = 0, \tag{8.46}$$

and the total current (electrons + holes) must be constant along x.

If there is injection of excess minority carriers, the carriers injected at the edge of the space-charge region diffuse and give rise to the diffusion current shown. As the total current must be constant, the majority carrier current must also vary with x, as shown in Fig. 8.11. An excess of majority carriers, extremely small in relative value, must therefore appear to compensate the charge of the injected minority carriers. Far from the junction the diffusion currents are small and the current is thus entirely a drift current of majority carriers, with uniform concentrations again. Hence there are five successive regions in a non-equilibrium junction: homogenous p type, p diffusion, space charge, n diffusion, and homogenous n type; by contrast there are no diffusion regions in an equilibrium junction, and thus only three regions.

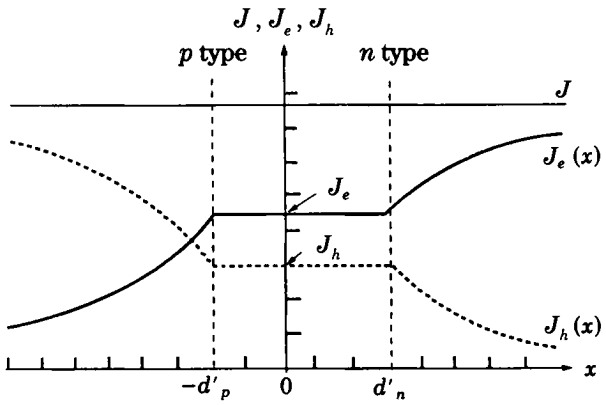


Fig. 8.11. Total current density (J), electron current density (J_e), and hole current density (J_h) across a p–n junction. (After Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

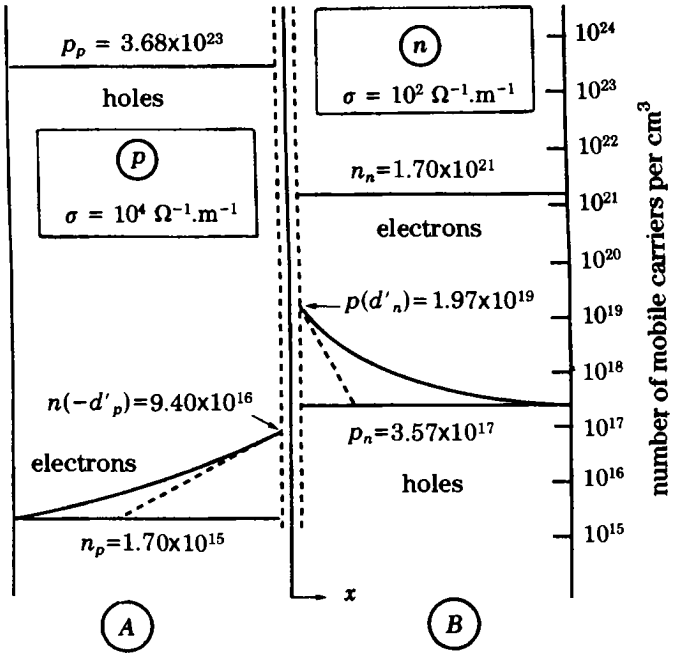


Fig. 8.12. Concentration profile in a forward biased germanium p–n junction (V_e = 0.1 V). (After Middlebrook, "An Introduction to Junction Transistor Theory," J. Wiley, 1957, 1965.)

Figure 8.12 gives the carrier densities in a germanium p–n junction as a function of x for a forward bias of 0.1 V. Note the logarithmic scale of the ordinate. The excess of majority carriers is negligible. The excess of minority carriers resulting from injection is very small compared with the majority carrier concentrations but very large compared with the equilibrium concentration of minority carriers. We see the five spatial regions: at the left and right of the figure the unperturbed n and p zones (not shown); then closer to the space charge, regions A and B where the current is mainly a diffusion current; and finally at the center the very narrow space-charge region.

Remark: For the non-equilibrium p–n junction it is useful to introduce the notion of "quasi-Fermi level": as this is not an equilibrium, the Fermi level is not constant in space. However from the concentration $n(x)$, by analogy with expression (8.7) we can define the electron quasi-Fermi level E_{Fe} by

$$n(x) = N_c \exp[-(E_c(x) - E_{Fe})/kT] \tag{8.47}$$

or

$$E_{Fe} = E_{c_0} - eV(x) + kT \log [n(x)/N_c]. \tag{8.48}$$

ilarly we define the quasi-Fermi level E_{Fh} for the holes:

$$E_{Fh} = E_{v_0} - eV(x) - kT \log [p(x)/N_v]. \tag{8.49}$$

an equilibrium junction E_{Fe} and E_{Fh} are equal, so that

$$E_{c_0} - E_{v_0} + kT \log [n(x)\ p(x)/N_c\ N_v] = 0 \tag{8.50}$$

ich is the law of mass action (Eq. (4.22)). In this case the quantities $_e, E_{Fh}$ do not depend on position, so that we recover the relation (8.7) its equivalent for holes. In a non-equilibrium junction the quasi-Fermi els vary with position. We now show that the spatial variation of E_{Fe} is ıted to the corresponding current J_e:

$$J_e = e\ \mu_e\ n\ \mathcal{E} + e\ D_e \frac{\partial n}{\partial x}, \tag{8.51}$$

$$J_e = -e\ \mu_e\ n\ \frac{\partial V}{\partial x} + e\ \mu_e \frac{kT}{e} \frac{\partial n}{\partial x}, \tag{8.52}$$

$$= n(x)\ \mu_e \frac{\partial}{\partial x}[-e\ V(x) + kT\ \mathrm{Ln}\ n(x)]. \tag{8.53}$$

ng the definition (8.48) of E_{Fe}, we obtain

$$J_e = n(x)\ \mu_e \frac{\partial E_{F_e}}{\partial x}. \tag{8.54}$$

ıilarly we have

$$J_p = p(x)\ \mu_h \frac{\partial E_{Fh}}{\partial x}. \tag{8.55}$$

e quasi-Fermi levels are actually the electrochemical potentials of trans- rt theory.

The notion of quasi-Fermi levels provides a description of the carrier ısity even within the space-charge region. We show that J_e varies little this region: from the conservation equation for the minority carriers in ady state:

$$\frac{\partial n}{\partial t} = \frac{1}{e} \frac{\partial J_e}{\partial x} - \frac{n - n_0}{\tau_n} = 0 \tag{8.56}$$

$$\frac{\partial J_e}{\partial x} = e \left(\frac{n - n_0}{\tau_n} \right),$$

deduce

$$\frac{\partial J_e}{\partial x} < \frac{e}{\tau_n} \delta n_{\max}. \tag{8.57}$$

The variation of J_e occurs over a distance of order of the diffusion length L_e, or about 0.1 mm, while the width of the space-charge region $d_p + d_n$ is about one micron. Then using the Einstein relation we have

$$J_e \leq \frac{eL_e}{\tau_n}\delta n_{\max}, \tag{8.58}$$

$$\frac{\partial E_{Fe}}{\partial x} \leq \frac{eL_e}{\tau_n \mu_e}\frac{\delta n_{\max}}{n} = \frac{\delta n_{\max}}{n}\frac{kT}{L_e} < \frac{kT}{L_e}. \tag{8.59}$$

The variation of E_{Fe} over the distance $(d_p + d_n)$ is certainly less than $kT(d_p + d_n)/L_e$, and thus much smaller than kT. We can say nothing about the variation of E_{Fe} in diffusion regions except that on the n side $\delta n << N_d$ (and $\delta p << N_a$ on the p side), so that E_{Fe} coincides with the Fermi level on the n side, and the quasi-Fermi levels have the variation shown in Fig. 8.13 across a non-equilibrium junction.

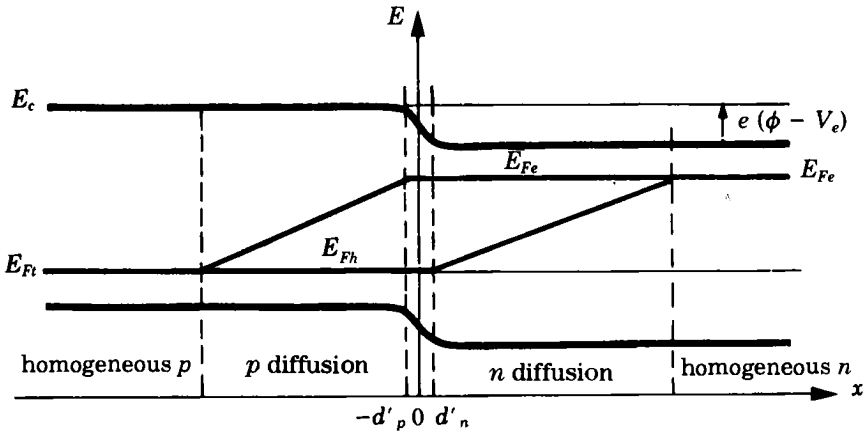


Fig. 8.13. Variation of the quasi-Fermi levels across a forward biased p–n junction. The slopes are not to scale. The width of the p (or n) diffusion zone is several times L_h (or L_e).

At the limit of the space charge we recover Eq. (8.35). The quasi-Fermi level of each species is approximately constant across the space-charge region because, if we neglect recombination, there is no other process allowing the electrons and holes to interact. The quasi-Fermi levels, which are the electrochemical potentials for the electrons and holes, are then decoupled.

d Capacitance of a Junction

he presence of external bias, the size of the space-charge region is altered. he case $N_a >> N_d$ we get from Eq. (8.33),

$$d_n'^2 = \frac{2\ \epsilon_0\ \epsilon_r\ (\phi - V_e)}{e\ N_d}. \tag{8.60}$$

ere is an accumulation of charge between the two conducting materials, hence a capacitance. The charge per unit area is $eN_d d_n'$ or

$$Q_c = [2\epsilon_0\ \epsilon_r(\phi - V_e)e\ N_d]^{1/2}. \tag{8.61}$$

ve change the bias V_e by a small amount dV_e we have a differential acitance per unit area:

$$C_d = \frac{d\ Q_c}{dV_e} = 2^{-1/2}[\epsilon_r\ \epsilon_0\ e\ N_d]^{1/2}[\phi - V_e]^{-1/2}. \tag{8.62}$$

For $\epsilon_r\epsilon_0 \sim 10^{-10}$ F·m^{-1}; $N_d \sim 10^{21}$ m^{-3}; $(\phi - V_e) \sim 1$ V, the capacitance $C_d \sim 0.01\ \mu$F/cm^2. This shows that in the equivalent electrical circuit a p–n junction there is a capacitor in parallel across the junction. The ue of the capacitance depends on the bias voltage V_e.

A device of this type is called a "varactor," a contraction of "variable ctor." The variation of capacitance with applied voltage is shown in Fig. 1. The ac behavior of p–n junctions is studied in Appendix 8.1. The trol of a capacitance by an applied voltage has many applications; for mple, varactors are used in the automatic frequency control of radio eivers.

e Breakdown of a p–n Junction

ve apply a reverse voltage exceeding a certain value to a diode, the erse current increases very rapidly, as shown in Fig. 8.15(a). This sudden rease is called "breakdown," and can occur as a result of two different chanisms.

The first of these is the Zener effect, the direct tunneling of charge riers between bands under large reverse bias. Figure 8.15(c) shows the id diagram in this case. The inverse polarization lowers the energies on n side with respect to those on the p side. For a sufficient voltage there l be occupied states in the valence band on the p side at the same energy empty states in the conduction band on the n side. There is therefore possibility of tunnelling between the p and n sides. The width d of the rgy barrier decreases as the reverse bias increases, and the probability tunnelling depends exponentially on the width of the barrier. Thus the erse current increases rapidly above the threshold V_Z.

Another mechanism which can cause breakdown is the multiplication of number of carriers through an "avalanche." Within the junction there

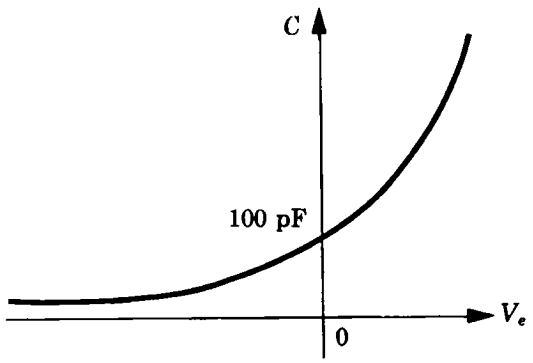


Fig. 8.14. Capacitance of a p–n junction as a function of applied voltage.

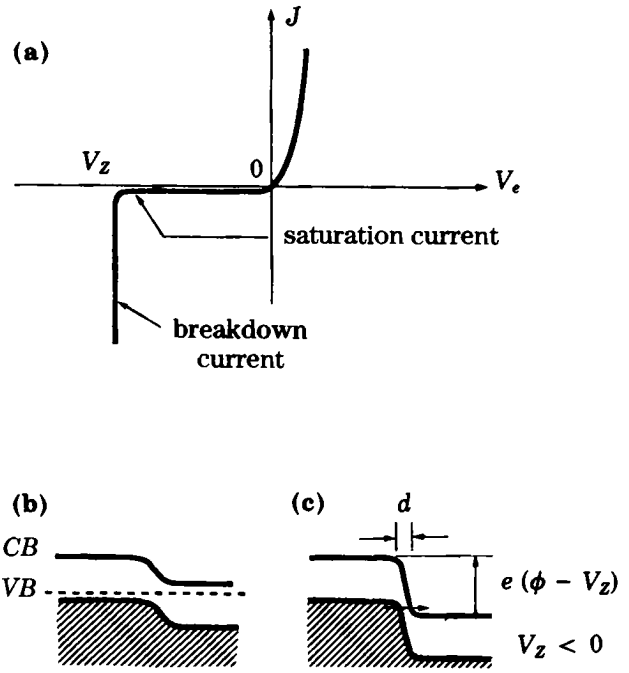


Fig. 8.15. (a) Current–voltage characteristics of a Zener diode; (b) band profile for $V_e = 0$; (c) band profile for $V_e = V_Z << 0$. (After Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

is a very strong electric field. The electrons can be accelerated by this field until they acquire very large kinetic energies, at which point the electron gas is no longer in thermodynamic equilibrium with the lattice. If an electron

has a kinetic energy greater than the band gap it can create an electron-hole pair by giving up its energy. These new carriers are themselves accelerated, and can create new pairs, hence the name "avalanche" for this process, which clearly leads to a rapid increase of the reverse current.

8.3f Transient Response of a p–n Junction

An important characteristic of the behavior of a p–n diode in the large signal regime is its recovery time. Consider a junction forward biased at voltage $V_1(>0)$, in steady state, and assume that at time $t=0$ the applied voltage is suddenly changed to a new reverse value $V_2(<0)$. The non-equilibrium carrier distributions will reach a new steady state only after a characteristic time τ_r.

Due to the rapid transit time, we can assume that the free carrier concentrations at the edges of the space region instantaneously take their new equilibrium values. We can describe the evolution of the non-equilibrium carrier densities from the initial steady state $(n_1(x), p_1(x))$ to the final steady state $(n_2(x), p_2(x))$ by introducing the distributions $\Delta n = n - n_1$ and $\Delta p = p - p_1$. These distributions obey the equations

$$\frac{\partial \Delta n}{\partial t} = D_e \frac{\partial^2 \Delta n}{\partial x^2} - \frac{\Delta n}{\tau_n} \quad (p \text{ region}), \tag{8.63}$$

$$\frac{\partial \Delta p}{\partial t} = D_h \frac{\partial^2 \Delta p}{\partial x^2} - \frac{\Delta p}{\tau_p} \quad (n \text{ region}), \tag{8.64}$$

with the boundary conditions

$$t = 0 \rightarrow \Delta n = \Delta p = 0, \quad \text{everywhere except at } x = 0, \tag{8.65}$$

$$x = 0 \rightarrow \begin{cases} \Delta n = n_p[\exp(eV_2/kT) - \exp(eV_1/kT)], & (8.66) \\ \Delta p = p_n[\exp(eV_2/kT) - \exp(eV_1/kT)]. & (8.67) \end{cases}$$

The solution of these equations is shown schematically in Fig. 8.16. The distribution Δn (or Δp), initially confined near $x = 0$, diffuses into the p (or n) region, while $\Delta n(x=0)$ (or $\Delta p(x=0)$) remains constant. At times $t << \tau_n, \tau_p$, Δn and Δp have a spatial extent of the order of $(D_e t)^{1/2}$ and $(D_h t)^{1/2}$, respectively.

This gives a current density

$$\Delta J \sim \frac{eD_e\ \Delta n\ (0)}{\sqrt{D_e t}} + \frac{eD_h\ \Delta p\ (0)}{\sqrt{D_h\ t}} \propto \frac{1}{\sqrt{t}}. \tag{8.68}$$

For $t \gtrsim \tau_n, \tau_p$ recombination is important, bringing Δn and Δp to the new steady state

$$\Delta n\ (x) = \Delta n\ (x=0) \times \exp(x/L_e),$$
$$\Delta p\ (x) = \Delta p\ (x=0) \times \exp(-x/L_h). \tag{8.69}$$

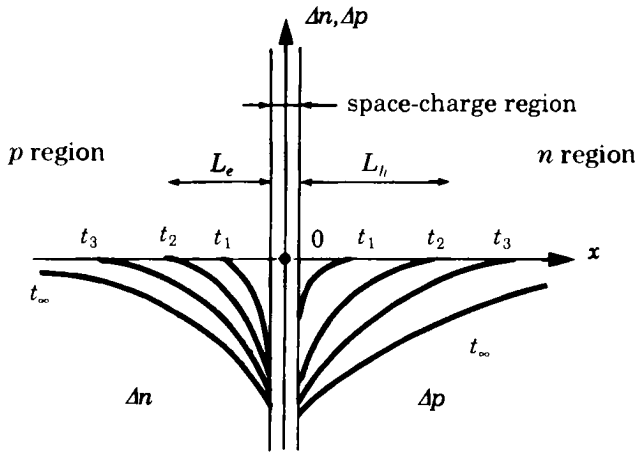


Fig. 8.16. Non-equilibrium densities of electrons (Δn) and holes (Δp) as functions of position for various times, when the applied voltage changes from $V_1 > 0$ to $V_2 < 0$. For clarity the curves are not drawn to $x = 0$ where they take the same value.

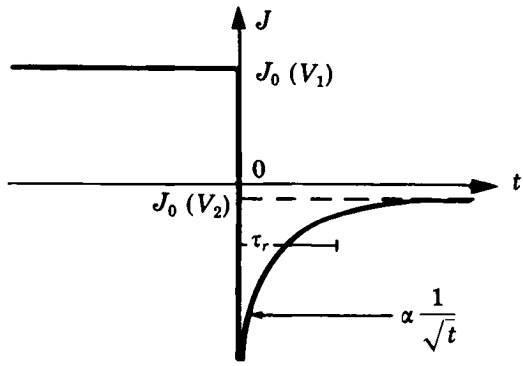


Fig. 8.17. Response of the current J when the applied voltage changes from $V_1 > 0$ to $V_2 < 0$ at time $t = 0$.

For $t >> \tau_n, \tau_p$ we have $J \simeq J(V_2)$. The recovery time is the larger of the two times τ_n, τ_p. For short times the carriers injected earlier maintain the junction as apparently conducting, even though V_2 is negative, and a significant reverse current flows. This is shown in Fig. 8.17.

This effect is obviously a hindrance to using a p–n junction as a rectifier, since it will not work above a frequency $f \sim 1/\tau_r$. High-frequency p–n rectifiers have to be made with short recombination times, for example, by introducing impurities creating recombination centers in the semiconductor.

ıppendix 8.1

roblem: Non-Stationary p–n Junctions and their igh-Frequency Applications

oblems **3** and **4** are independent of **1** and **2**. Solution of problem **1** is ough for attempting the second part of this study.

Aim of the problem: p–n junctions are currently used as ac rectifiers. wever, when the ac frequency is raised, new phenomena appear that ake p–n junctions difficult to use as rectifiers, but make them valuable for ıer applications. Here we present an introduction to these phenomena.

In all of these problems we assume that the junction is subject to a ne-dependent voltage of the form $V = V_p - V_n = V_0 + \mathrm{Re}(\delta V e^{i\omega t})$, where ' is infinitesimally small (the "small signal" domain) and we consider the ear response of the system, i.e., terms up to the first order in δV. (The e of the diode as a rectifier clearly departs from these assumptions.) As the steady case, we must distinguish on the one hand phenomena associ- ed with charges stored in the space-charge region, and on the other hand enomena involving transfer of carriers across the junction. We assume at the voltage applied to the device appears fully and instantaneously at e edges of the space-charge region.

We shall adopt the following notation: ϵ_r is the relative dielectric con- ınt of the semiconductor, N_a and N_d the doping of the p and n regions, spectively, N_c and N_v the equivalent densities of states of the conduction d valence bands, $\mu_e, D_e, \tau_n, L_e, n_p$ (and $\mu_h, D_h, \tau_p, L_h, p_n$) the mobility, fusion coefficient, recombination time, diffusion length, and equilibrium ncentration of electrons in the p region (and of holes in the n region), d_p d d_n the respective widths of the space-charge regions on the p and n

First Part: Abrupt p–n Junction

The energy level scheme of an abrupt planar p–n junction in equilibrium is given in Fig. 8.5. The plane of the junction is taken as $x = 0$ and we call $\phi(> 0)$ the internal potential of the equilibrium junction.

1. Using the standard approximation of the space charge leading to bending of the bands in parabolic arcs, give an expression for the charge Q per unit area stored as fixed charges in the band-bending regions on either side of the junction plane as a function of V and the dopings N_a, N_d.

2. Deduce from the foregoing an expression for the current density $\delta J_1(t)$ associated with the variation of Q over time, and show how for small signals this effect gives the junction the behavior of a capacitor whose capacitance depends on V_0 (a "varactor"). Give the values of the "differential capacitance" C (defined in Sect. 8.3) per unit area and of the associated complex impedance Z_1. We recall that the complex impedance per unit area is the ratio $Z = \Delta V/\Delta J$, where ΔV and ΔJ are δV and δJ expressed in complex notation.

3. First neglecting δJ_1, determine the current density $J = J_0 + \delta J_2(t)$ corresponding to the passage of mobile carriers across the junction. One makes the usual (Shockley) approximations, which imply that the electrons and holes are each separately in thermal equilibrium in the space-charge region (Eq. (8.35)) and neglects recombination in this region. Write down the equations governing the evolution of the minority carrier densities (electrons in the p region and holes in the n region) in neutral regions, and the boundary conditions. Using the fact that δV is infinitesimal, write these equations and boundary conditions to zeroth and first orders in δV.

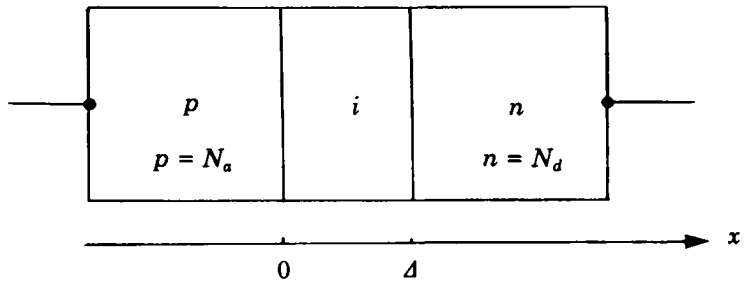
4. Solve these equations and find expressions for J_0 and δJ_2. Deduce as a function of V_0 the contribution Z_2 to the complex impedance per unit junction area associated with the transfer of mobile carriers. Explain physically the ω dependence of this impedance.

5. Combining the results of **2** and **4**, give an expression for the impedance per unit area of the junction. In a real diode the impedance does not tend to zero for infinite ω. Why?

6. We have so far taken δV as "infinitesimal." Up to what order of magnitude of δV_L (in volts) is this an acceptable approximation?

Second Part: Extension to Large Signal Regime: The p–i–n Diode

For high frequencies the allowable range of δV for a linear response can be extended beyond δV_L because of an effect related to the width of the space charge. This effect is particularly important in the *pin* diode studied here. The *pin* (p–i–n) diode, generally made of silicon, is derived from the p–n junction by adding an undoped silicon zone (i = "intrinsic") of width $\Delta \sim 100\ \mu$m between the p and n regions (Fig. 8.18).



. 8.18. Schematic view of a *pin* diode.

The three following questions deal with the energy level scheme of the ı diode in equilibrium. Lacking an analytic solution of the band bending an i region, we attempt to understand it intuitively in the next two estions.

7. A simple approximation to a p–i junction is a p–n junction with $<< N_a$ (a p–n^- junction). Using the results of **1**, what can you say out the width of the space-charge zone in the n^- region compared with ɛ width of the corresponding zone in the p region? In which spatial region es most of the variation of the electrostatic potential occur?

8. Consider next the behavior of the space charge zone in an intrinsic niconductor. Write down in one dimension the Poisson equation and the ıation expressing the charge density as a function of the electrostatic po- ıtial, and deduce the differential equation satisfied by the potential. This ıation has no analytic solution. Nonetheless, one can estimate the order magnitude of the characteristic distance over which the bands curve. To ıs end, rewrite the equation in an approximate form in the limit of very ıall potential variation and give a solution in terms of a characteristic ıtance λ. Express the characteristic distance λ (the "screening length") a function of the intrinsic equilibrium carrier density $2n_i$. Calculate λ · silicon at room temperature ($\epsilon_r = 11.7$; $n_i = 1.4 \times 10^{16}$ m^{-3}).

9. For typical concentrations N_a, N_d (see the main text), what is the ler of magnitude of the width of the space-charge regions for the n and parts of the junction? Assuming that the results of **8** remain valid for $> kT/e$, give the form of the energy level scheme for an equilibrium *pin* ɔde.

10. When the diode is reverse biased ($-V_0 >> kT/e$), the i region is al- ɔst completely depleted of carriers. Deduce a very simple approximation · the form of the electric field in the three regions p, i, and n. The differ- tial capacitance of the diode then reduces to that of a plane capacitor. ve without calculation the expression for this capacitance per unit area.

11. Consider a forward biased diode ($V_0 >> kT/e$), and assume for e moment $\delta V = 0$. Because of the bias of the p–i and i–n junctions, a

significant number of carriers is injected into the i region: holes from the p region and electrons from the n region. In which sense does the screening length λ (found in **8**) vary? Deduce that the i region is neutral except for two layers of width λ, and that the electric field is either uniform or zero. We wish to show now that it is zero. For this we assume as in the main text that the electrons and holes are each separately in equilibrium in the space-charge regions and in the intrinsic region, and also that the recombination time is infinite. Applying Eq. (8.35) to the electrons and its equivalent to the holes, show that the potential difference between $x = 0_+$ and $x = \Delta_-$ is zero.

Sketch the shape of the energy bands as a function of x.

12. Assuming complete symmetry between electrons and holes ($N_c = N_v, N_a = N_d$), calculate the carrier density in the various regions. Show that the current is still given by the expression found for J_0 in **4** (Shockley's law).

13. The response of a *pin* diode to an ac excitation δV is essentially dominated by the i region. (You may assume this without proof.) This region is electrically equivalent to a resistance R_i in series with a capacitance C_i. We calculate C_i in this question and R_i in the next one.

If the diode is biased forward with a voltage V_0, the i region can store free carriers, electrons, or holes, which flow to the n side for electrons (and to the p side for holes) when the voltage V_0 is reduced to zero. We thus have a rather peculiar capacitor, which stores a hole charge Q_i and an electron charge $-Q_i$ in the same region of space. Calculate the value of Q_i per unit area, and the differential capacitance $C_i = dQ_i/dV_0$. Compare C_i with the capacitance found in **10** for a reverse bias.

14. For forward bias the i region does not behave simply as a capacitance C_i but also as a resistance R_i whose value is determined by the stored charge density. Give the value of R_i per unit surface as a function of V_0. Above what frequency ω is the impedance of a *pin* diode dominated by R_i rather than by the impedance of C_i? Show that the corresponding time can be regarded as a diffusion time for the i region.

Practical Importance of the *p-i-n* Diode

For a sufficiently high forward bias the impedance of the diode is small. This is therefore a short circuit controllable by the applied voltage V_0. The *pin* diode is thus an electronic switch controlled by the voltage. On the other hand if we apply a large ac voltage, since the impedance varies very rapidly with the voltage we have a non-linear device suitable for harmonic generation.

. In the presence of an applied voltage V, the potential difference be-
en the n and p sides becomes $\phi - V$. The charge density is $-eN_a$ on
p side and $+eN_d$ on the n side. The respective widths d_p and d_n of
se regions obey $(N_a d_p^2 + N_d d_n^2)e/2\epsilon_r\epsilon_0 = \phi - V$ (from continuity of the
ential) and $N_a d_p = N_d d_n$ (field continuity). We deduce

$$d_p\ N_a = d_n\ N_d = \left[\frac{2\ \epsilon_r\ \epsilon_0\ (\phi - V)}{e}\frac{N_a\ N_d}{N_a + N_d}\right]^{1/2}. \tag{8.70}$$

: charge stored in the junction is $Q_p = -eN_a d_p$ on the p side, and the
osite on the n side, with

$$|Q| = Q_n = -Q_p = [2\ e\ \epsilon_r\ \epsilon_0\ (\phi - V)\ N_a\ N_d/(N_a + N_d)]^{1/2}. \tag{8.71}$$

2. Counting the current as positive in the direction $p \to n$ we have

$$\delta J_1\ (t) = \frac{d}{dt}Q_p = \text{Re}\ (i\omega\ \delta V\ e^{i\omega t})\ \left.\frac{dQ_p}{dV}\right|_{V=V_0}, \tag{8.72}$$

$$= \text{Re}\ (i\omega\ \delta V\ e^{i\omega t}) \times \left[\frac{e\ \epsilon_r\ \epsilon_0}{2(\phi - V_0)}\frac{N_a\ N_d}{N_a + N_d}\right]^{1/2}, \tag{8.73}$$

ch is of the form

$$\text{Re}\ (iC\omega\ \delta V\ e^{i\omega t}) = C\ dV/dt.$$

s gives a differential capacitance per unit surface:

$$C = \left[\frac{e\ \epsilon_r\ \epsilon_0}{2\ (\phi - V_0)}\frac{N_a\ N_d}{N_a + N_d}\right]^{1/2}, \tag{8.74}$$

ɔciated with a complex impedance $Z_1 = [i\omega C]^{-1}$.

3. For the p region the transport equation for the carriers is

$$\frac{\partial n}{\partial t} = D_e\frac{\partial^2 n}{\partial x^2} - \frac{n - n_p}{\tau_n}. \tag{8.75}$$

The boundary conditions are $\begin{cases} x \to -\infty & n \to n_p \\ x \to 0_- & n \to n_p \exp(eV/kT). \end{cases}$ (8.76)

th $V = V_0 + \text{Re}\ (\delta V e^{i\omega t})$, and setting $n = n_0 + \text{Re}(\delta n e^{i\omega t})$ we have to
ɔ order in δV:

$$0 = D_e\frac{\partial^2 n_0}{\partial x^2} - \frac{n_0 - n_p}{\tau_n}. \tag{8.77}$$

The boundary conditions are $\begin{cases} x \to -\infty & n_0 \to n_p \\ x \to 0_- & n_0 \to n_p \exp\dfrac{eV_0}{kT}. \end{cases}$ (8.78)

To first order in δV:

$$i\omega\delta n = D_e\frac{\partial^2\delta n}{\partial x^2} - \frac{\delta n}{\tau_n} \tag{8.79}$$

$$\text{with boundary conditions } \begin{cases} x \to -\infty & \delta n_0 \to 0 \\ x \to 0_- & \delta n \to \dfrac{e\delta V}{kT} n_p \exp\dfrac{eV_0}{kT}. \end{cases} \tag{8.80}$$

Similarly, setting $p = p_0 + \mathrm{Re}(\delta p e^{i\omega t})$ in the n region, we have, to zero order in δV:

$$0 = D_h\frac{\partial^2 p_0}{\partial x^2} - \frac{p_0 - p_n}{\tau_p} \tag{8.81}$$

$$\text{with boundary conditions } \begin{cases} x \to +\infty & p_0 \to p_n \\ x \to 0_+ & p_0 \to p_n \exp\dfrac{eV_0}{kT} \end{cases} \tag{8.82}$$

and to first order in δV:

$$i\omega\delta p = D_h\frac{\partial^2\delta p}{\partial x^2} - \frac{\delta p}{\tau_p} \tag{8.83}$$

$$\text{with boundary conditions } \begin{cases} x \to +\infty & \delta p \to 0 \\ x \to 0_+ & \delta p \to \dfrac{e\delta V}{kT} p_n \exp\dfrac{eV_0}{kT}. \end{cases} \tag{8.84}$$

4. The solutions of the differential equations written down in **3** are: to zero order in δV:

$$p \text{ side: } n_0(x) = n_p\left\{1 + \left[\exp\left(\frac{eV_0}{kT}\right) - 1\right]\exp\left(\frac{x}{L_e}\right)\right\}, \tag{8.85}$$

$$n \text{ side: } p_0(x) = p_n\left\{1 + \left[\exp\left(\frac{eV_0}{kT}\right) - 1\right]\exp\left(-\frac{x}{L_h}\right)\right\} \tag{8.86}$$

with $L_e = \sqrt{D_e\,\tau_n}$ and $L_h = \sqrt{D_h\,\tau_p}$

to first order in δV:

$$p \text{ side: } \delta n(x) = n_p\frac{e\delta V}{kT}\exp(eV_0/kT)\exp(x\sqrt{1 + i\omega\tau_n}/L_e), \tag{8.87}$$

$$n \text{ side: } \delta p(x) = p_n\frac{e\delta V}{kT}\exp(eV_0/kT)\exp(-x\sqrt{1 + i\omega\tau_p}/L_h). \tag{8.88}$$

This gives the following currents at $x = 0$:

$$\begin{aligned} J_0 &= e\left(D_e\frac{\partial n_0}{\partial x} - D_h\frac{\partial p_0}{\partial x}\right)\Bigg|_{x=0} \\ &= e\left(\frac{D_e\,n_p}{L_e} + \frac{D_h\,p_n}{L_h}\right)\left[\exp\left(\frac{eV_0}{kT}\right) - 1\right], \end{aligned} \tag{8.89}$$

$$\delta J_2 = e\left(D_e\frac{\partial \delta n}{\partial x} - D_h\frac{\partial \delta p}{\partial x}\right)\Bigg|_{x=0}$$
$$= \frac{e^2\delta V}{kT}\exp\frac{eV_0}{kT}\times$$
$$\mathrm{Re}\left[\left(\frac{D_e\, n_p\,\sqrt{1+i\omega\tau_n}}{L_e} + \frac{D_h\, p_n\,\sqrt{1+i\omega\tau_p}}{L_h}\right)e^{i\omega\tau}\right]. \qquad (8.90)$$

The complex impedance Z_2 is the ratio of δV to δJ_2, both taken as complex:

$$Z_2 = \frac{\delta V}{\delta J_2} = \frac{kT}{e^2}\exp\left(-e\frac{V_0}{kT}\right)\left(\frac{D_e n_p\sqrt{1+i\omega\tau_n}}{L_e} + \frac{D_h p_n\sqrt{1+i\omega\tau_p}}{L_h}\right)^{-1} \qquad (8.91)$$

If $\omega << 1/\tau_n, 1/\tau_p$, the junction has the time to reach a steady state over a time of order half a period; Z_2 is thus real and close to dV_0/dJ_0.

If $\omega >> 1/\tau_n, 1/\tau_p$, a steady state is never reached: diffusion only affects a thickness of order $(D/\omega)^{1/2}$ (the distance over which carriers diffuse in a time of order $1/\omega$), giving the dependence of δJ_2 on $\omega^{1/2}$.

Remark: an impedance with $(i\omega)^{-1/2}$ dependence is typical of diffusion processes. Here it simultaneously describes the resistance and capacitance associated with the injected charge.

5. $\delta J_{\mathrm{total}} = \delta J_1 + \delta J_2$. The impedances Z_1, Z_2 are thus combined in parallel:

$$Z = \left[\frac{1}{Z_1} + \frac{1}{Z_2}\right]^{-1}$$
$$= \left\{i\omega\left[\frac{e\,\epsilon_r\,\epsilon_0}{2(\phi - V_0)}\frac{N_a\,N_d}{N_a + N_d}\right]^{1/2} + \right.$$
$$\left.\frac{e^2}{kT}\exp\frac{eV_0}{kT}\left(\frac{D_e\, n_p\,\sqrt{1+i\omega\tau_n}}{L_e} + \frac{D_h\, p_n\,\sqrt{1+i\omega\tau_p}}{L_h}\right)\right\}^{-1} \qquad (8.92)$$

In practice the impedance of a real diode does not tend to zero for infinite ω as the resistance R_S of the semiconductor bulk is in series with Z:

$$Z_{\mathrm{total}} = Z + R_S. \qquad (8.93)$$

6. In the region in which Z is dominated by the capacitance of the space-charge region (that is, typically for $eV_0/kT \leq 0$ for reverse bias), the linear approximation is acceptable for $\delta V << \phi - V_0 \sim 1$ V, typically. For a forward bias of the junction, Z_2 becomes dominant. The linear approximation means to replace the factor $\exp(e\delta V/kT)$ of the boundary conditions (8.76) by $1 + (e\delta V/kT)$ in the boundary conditions (8.80). This is acceptable only for $\delta V << kT/e \sim 25$ mV at room temperature.

7. From **1** we have $d_n/d_p = N_a/N_d >> 1$, and the space-charge region is much larger on the weakly doped side (n^-). The variations of electrostatic potential similarly compare as

$$\frac{(e/2\epsilon_r\epsilon_0)\ N_d\ d_n^2}{(e/2\epsilon_r\epsilon_0)\ N_a\ d_p^2} = \frac{d_n}{d_p} >> 1. \tag{8.94}$$

Most of the potential variation occurs on the less-doped side (n^-).

8. The Poisson equation is

$$\frac{d^2V(x)}{dx^2} = -\frac{\rho}{\epsilon_r\ \epsilon_0}. \tag{8.95}$$

In the presence of a potential difference V with respect to the intrinsic region, the densities of electrons and holes are given by

$$n\ (x) = n_i \exp[eV(x)/kT], \tag{8.96}$$

$$p\ (x) = n_i \exp[-eV(x)/kT]. \tag{8.97}$$

Hence

$$\rho\ (x) = e(p - n) = -2\ e\ n_i\ \mathrm{sh}[eV(x)/kT]. \tag{8.98}$$

The differential equation satisfied by the electrostatic potential is thus

$$\frac{d^2V(x)}{dx^2} = \frac{2\ e\ n_i}{\epsilon_r\ \epsilon_0}\mathrm{sh}[eV(x)/kT]. \tag{8.99}$$

If $eV(x) << kT$ we get the approximate equation

$$\frac{d^2V}{dx^2} = \frac{2\ n_i\ e^2}{\epsilon_r\ \epsilon_0\ kT}V, \tag{8.100}$$

whose general solution is

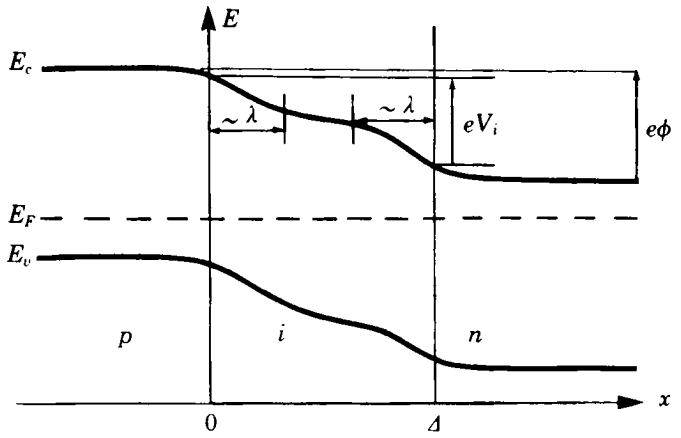
$$V = \alpha \exp(x/\lambda) + \beta \exp(-x/\lambda) \tag{8.101}$$

with

$$\lambda = (\epsilon_r\ \epsilon_0\ kT/2\ n_i\ e^2)^{1/2}. \tag{8.102}$$

The characteristic length λ is the screening or Debye length. This length characterizes the screening effect, i.e., the rapid spatial attenuation of the effect of an applied electrostatic perturbation on a conductor. This length is shorter the higher the free carrier density (whatever the sign of their charge). Here the total free carrier density is $2n_i$. Equation (8.102) gives $\lambda \simeq 24\ \mu$m for silicon at room temperature. This length is of the same order as Δ.

9. For typical dopings N_a, N_d the width of the space-charge regions on the n and p sides (see the text) is about $0.1 - 1\ \mu$m, and thus much smaller



8.19. Band profile of an equilibrium *pin* diode.

λ and Δ. Similarly as in **7**, we deduce that the largest part V_i of the ıtion of the potential ϕ occurs in the intrinsic region (Fig. 8.19).

0. Under reverse bias, since the i region is almost completely depleted of ers, we have $\rho \simeq 0$ there, giving $d^2V/dx^2 = 0$ from Poisson's equation. electric field is constant in the i region. The only space charges are ıarrow zones (cf. **7**) at the edges of the n and p regions. The potential ıtion is negligible in these zones, and we can approximate the electric as

$$' = 0 \ (n \text{ and } p \text{ regions}), \tag{8.103}$$

$$' = \frac{\phi - V_0}{\Delta} \ (i \text{ region}). \tag{8.104}$$

diode's capacitance is identical to that of a plane capacitor whose plates he p–i and i–n boundaries, hence

$$' = \frac{\epsilon_r \, \epsilon_0}{\Delta}. \tag{8.105}$$

1. If there are many injected carriers, the carrier concentration clearly eds $2n_i$ (strong injection regime), and from **8** the screening length be- es much smaller than Δ. The space charge in the i region concentrates ıe edges, and most of the i region becomes neutral ($n = p$).

ince the i region is neutral, $\mathcal{E}$ is constant. We assume that $\mathcal{E}$ is non-zero, show that this is contrary to the condition $n = p$.

Jnder the hypothesis that λ is small and $\mathcal{E}$ is constant, the band profile ld have the form shown in Fig. 8.20 for forward bias V_0.

n the approximation where the electrons and holes are separately in ibrium on either side of the space charge around $x = 0$, we have

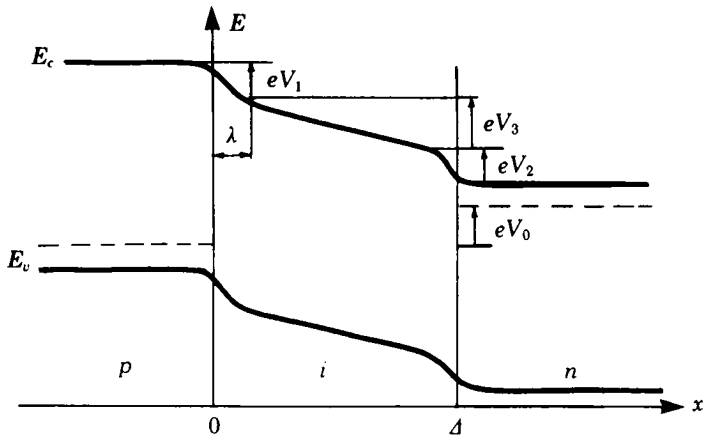


Fig. 8.20. Band profile in a forward biased *pin* diode, assuming non-zero electric field in the i region.

$$p\,(\lambda) = N_a \exp\left(-\frac{e\,V_1}{kT}\right). \tag{8.106}$$

Similarly

$$n\,(\Delta - \lambda) = N_d \exp\left(-\frac{eV_2}{kT}\right). \tag{8.107}$$

Further we also have equilibrium across the i region, so that

$$n\,(\lambda) = N_d \exp\left[-\frac{e\,(V_2 + V_3)}{kT}\right], \tag{8.108}$$

$$p\,(\Delta - \lambda) = N_a \exp\left[-\frac{e\,(V_1 + V_3)}{kT}\right]. \tag{8.109}$$

Charge neutrality implies the equality of Eqs. (8.106) and (8.108), and of Eqs. (8.107) and (8.109). We deduce

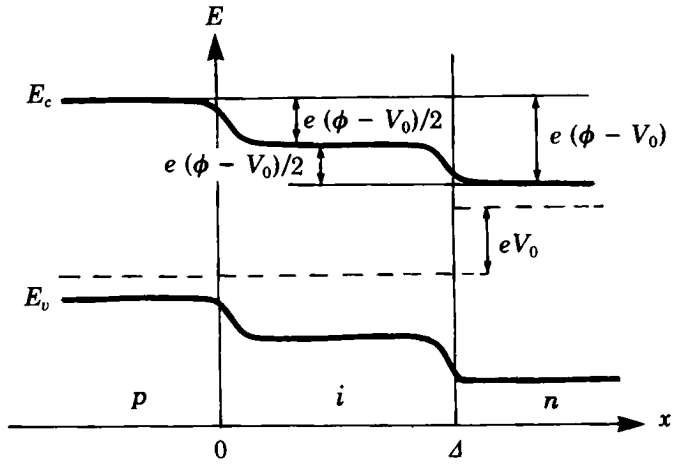
$$-\frac{eV_1}{kT} = -\frac{e(V_2 + V_3)}{kT} + \log\frac{N_d}{N_a}, \tag{8.110}$$

$$-\frac{eV_2}{kT} = -\frac{e(V_1 + V_3)}{kT} + \log\frac{N_a}{N_d}. \tag{8.111}$$

Adding Eqs. (8.111) and (8.110) gives $V_3 = 0$, showing that the bands are flat in the i region. The electron and hole concentrations are constant and equal in this region, and the electric field vanishes (Fig. 8.21).

12. Setting $V_3 = 0$ and equating (8.106) and (8.108) we have

$$N_a \exp\left(-\frac{eV_1}{kT}\right) = N_d \exp\left(-\frac{eV_2}{kT}\right). \tag{8.112}$$



8.21. Real band profile in a forward biased *pin* diode.

:her

$$e(V_1 + V_2) = e(\phi - V_0). \tag{8.113}$$

This fixes V_1 and V_2. If we assume for simplicity that $N_a = N_d$, then $= V_2 = (\phi - V - 0)/2$. In this case the electron and hole concentrations he i region are equal, and

$$n = p = N_d \exp[-e(\phi - V_0)/2\ kT], \tag{8.114}$$

$$= n_i \exp(eV_0/2\ kT). \tag{8.115}$$

Charge neutrality holds again, but with charge densities higher than in ilibrium: the product np differs from n_i^2.
The minority carriers in the n and p regions are governed by the same ations and the same boundary conditions as in **3.** Then we have, as in

$$p \text{ region: } n = n_p[1 + (\exp(eV_0/kT) - 1)\exp(x/L_e)] \quad \text{and } p \approx N_a \tag{8.116}$$

$$n \text{ region: } p = p_n[1 + (\exp(eV_0/kT) - 1)\exp(-x/L_h)] \quad \text{and } n \approx N_d. \tag{8.117}$$

As there is no recombination in the i region, we can find the current as l and in the main text:

$$J_0 = e\left(\frac{D_e\ n_p}{L_e} + \frac{D_h\ p_n}{L_h}\right)[\exp(eV_0/kT) - 1] \text{ (Shockley's law).} \tag{8.118}$$

13. The total hole charge stored per unit area in the i region is

$$Q_i \simeq e\ n_i\ \Delta\ \exp(eV_0/2\ kT). \tag{8.119}$$

The total electron charge stored in this region is $-Q_i$. The variation of Q_i with V gives a capacitance per unit area:

$$C_i = \left.\frac{dQ_i}{dV}\right|_{V=V_0} = \frac{e^2\, n_i\, \Delta}{2\, kT} \exp(eV_0/2kT). \tag{8.120}$$

This capacitance can be much larger than that found in **10**. The transition between the two regimes occurs at

$$V_0 = \frac{2kT}{e} \operatorname{Ln} \frac{2kT\, \epsilon_r\, \epsilon_0}{e^2\, n_i\, \Delta^2} = \frac{4kT}{e} \operatorname{Ln} \frac{2\lambda}{\Delta} \tag{8.121}$$

which is of the order of kT/e. In practice the capacitance of the junction will be dominated by this effect once $V_0 > 0$.

The existence of a capacitance when the i region is essentially neutral everywhere ($n = p$) may seem surprising at first sight. In contrast to a normal capacitor, electrons and holes are stored in the same spatial region, but with different chemical potentials (the quasi-Fermi levels E_{Fe}, E_{Fh}). They do not react because recombination is considered to be infinitely slow.

14. The series resistance of the i region per unit area is

$$R_i = \frac{\Delta}{n_i \exp(eV_0/2kT)\; e\; (\mu_e + \mu_h)}. \tag{8.122}$$

This resistance dominates the impedance of the diode if $R_i >> 1/C_i\omega$, i.e., for frequencies such that

$$\omega >> 1/R_iC_i = \frac{2kT\; (\mu_e + \mu_h)}{e\Delta^2} = \frac{2\; (D_e + D_h)}{\Delta^2}. \tag{8.123}$$

The associated response time $\Delta^2/2(D_e + D_h)$ represents the mean time an electron and a hole initially placed at opposite ends of the i region take to meet if they travel by diffusion. It is also the characteristic time taken by the i region to reach a steady state by diffusion. By imagining the inverse process (dissociation of an electron-hole pair into an n-region electron and a p-region hole) we can also regard this time as an effective lifetime for the i region.

›plications of the p–n nction and Asymmetrical ›vices

–n junction has many applications which we sketch below. There course other asymmetrical devices such as the metal–semiconductor t. We encounter this latter structure as soon as we try to use the cal properties of semiconductors in circuits, and we must thus under- its properties. The metal–semiconductor contact properties are also basis of the Schottky diode operation.

ıny of the properties of these devices depend directly on the surface ·ties of the semiconductors, which we shall summarize. These proper- ɔo allow the use of semiconductors as light detectors. Finally we shall ɔe junctions between chemically distinct semiconductors, or "hetero- ɔns."

\pplication of p–n Junctions

are many applications of the p–n junction. The most obvious is their ac rectifiers, which produce the dc current required for motors, elec- s, and all types of electronic devices. These applications make direct Shockley's law

$$= J_s \left[\exp\left(\frac{eV_e}{kT}\right) - 1\right]. \tag{9.1}$$

applications involve the physics we discussed in Chap. 8.

9.1a Photovoltaic Cells and Solar Cells

Consider an unbiased p–n junction. The junction is in equilibrium and there is an energy barrier $e\phi$ between the n side and the p side. The electrical field of the barrier is $\mathcal{E}$. If we illuminate the semiconductor with light of sufficient photon energy $\hbar\omega$ we create electron-hole pairs. Excess minority carriers are then accelerated by the internal field $\mathcal{E}$, which separates them, repelling the electrons to the n region and the holes to the p region. This is shown in Fig. 9.1(a). The separated carriers create an electric field $\mathcal{E}'$ which opposes the field $\mathcal{E}$. The resultant field is then $\mathcal{E} - \mathcal{E}'$, meaning that the potential drop between the p and n sides is reduced from ϕ to $\phi - V_f$ as shown in Fig. 9.1(b).

The effect is the same as a forward bias V_f applied to the junction, the net result being the creation of a potential difference V_f at the extremities of the junction. The appearance of this potential difference at the edges of an illuminated junction is called the "photovoltaic effect." The maximum value of V_f is ϕ, which is in turn less than E_g/e (Eq. (8.17)).

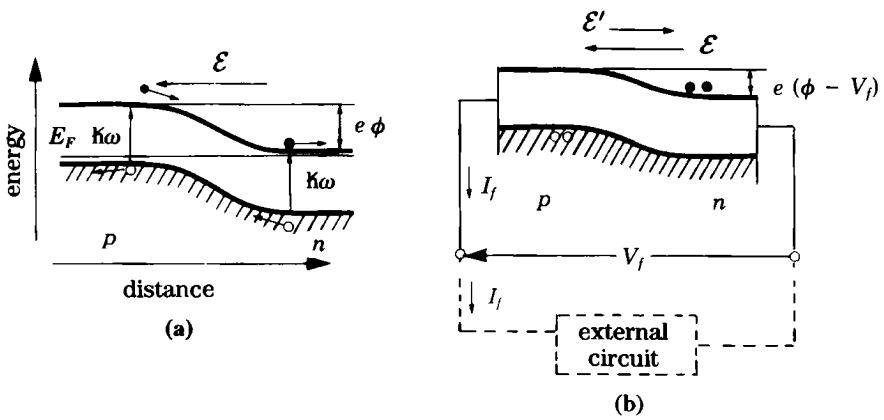


Fig. 9.1. Band profile of an illuminated p–n junction. (a) The electric field $\mathcal{E}$ of the junction separates the carriers created by the illumination. (b) The photocreated carriers induce an electric field $\mathcal{E}'$, and thus a voltage V_f usable in an external circuit.

When the junction is illuminated we may then say that a photocurrent $-I_f$ appears, in the opposite sense from the direct current. Connecting an external circuit to the illuminated diode allows us to measure this current. If the I–V characteristic of the junction is of the form (9.1), and if P is the photon flux and η the quantum efficiency, i.e., the number of pairs collected for one photon, then the photocurrent I_f is

$$I_f = 2\eta e P, \tag{9.2}$$

the total current under illumination is

$$I = I_s \left[\exp\left(\frac{eV_f}{kT}\right) - 1 \right] - I_f \tag{9.3}$$

and the voltage V_f is given by

$$V_f = \frac{kT}{e} \log \left[\frac{I + I_s + I_f}{I_s} \right]. \tag{9.4}$$

The current–voltage characteristic of the illuminated junction is shown in Fig. 9.2. The junction behaves as a generator in the region $V_f > 0, I < 0$.

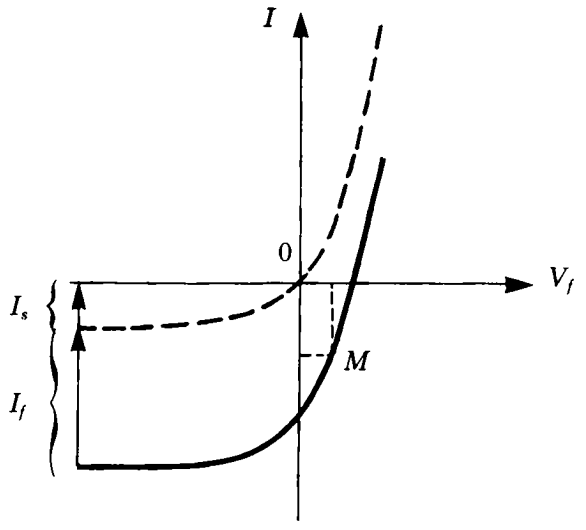


Fig. 9.2. Current–voltage characteristic of a photovoltaic cell. The full curve is that obtained under illumination, and the dashed curve without illumination. The distance between the two curves increases with I_f, and thus with illumination.

A junction diode can be used as a photoelectric cell and as a light detector, but its most important application is the direct conversion of light energy into electrical energy in a solar cell. The power of such a cell, i.e., the product $V_f \times I$, is maximal when the area of the rectangle OM is maximal.

The global yield of a solar cell depends on the fraction of radiation absorbed by the semiconductor and on the quantum efficiency. The solar spectrum has its maximum intensity at $\lambda = 0.5\ \mu$m, and the spectral energy is still about half of this maximum for $\lambda = 1\ \mu$m. This means that a diode with band gap E_g of about 2 eV is efficient in transforming the high-energy part of the solar spectrum but does not collect the infrared energy. By contrast a semiconductor with a band gap of 0.5 eV absorbs most of the

photons arriving from the sun but the quantum efficiency decreases: solar photons whose energy greatly exceeds E_g create very energetic electron-hole pairs which can lose their energy by exciting crystal vibrations, i.e., by heating the crystal, and this energy is lost. Crystalline silicon is a good compromise; the theoretical quantum efficiency is 20% and cells have been constructed with efficiencies of 15%. Of course if the recombination time is too short the current I_f is smaller, the pairs recombining on the spot before the junction field separates them.

The power supplied by a solar cell is proportional to its surface: the flux of solar energy at the Earth is about 1 kW/m^2, so a 1 cm^2 cell will yield about 10 mW. It is expensive to produce large areas of crystalline silicon. This is why solar cells in everyday use (e.g., in calculators) use amorphous silicon (see Sect. 3.3), which is cheaper to produce. One can make junctions out of it, and the efficiency of such solar cells can reach 10%. These cells are currently used in pocket calculators.

9.1b Electroluminescent Diodes and Lasers ✗

Electroluminescent diodes are devices that directly convert electrical energy into luminous energy via radiative recombination. This is the inverse transformation to that occurring in a photovoltaic cell. We have seen in Chap. 6 that in the presence of excess electron-hole pairs recombinations will occur, and some of these will be radiative. In a forward biased junction one can cause injection of excess carriers and thus light emission. The emission will be intense if the radiative efficiency is relatively large, i.e., if the efficiency of the non-radiative processes is not too large. This requires pure semiconductors, but also, from the formula of Appendix 6.1, semiconductors whose absorption coefficient is large at energies near the band gap. Semiconductors with direct gaps (see Sect. 2.6a), i.e., where the minimum of the conduction band and the maximum of the valence band are at the same point of the Brillouin zone, have large radiative efficiencies. This is the case for GaAs which emits in the very near infrared. The main materials now used for electroluminescent diodes are, besides GaAs, GaP and alloys $GaAs_{1-x}P_x$ which have band gaps between 1.4 and 2.2 eV. We thus get emissions from the near infrared to the green according to the amount of arsenic in the alloy.

To use electroluminescent diodes for electronic displays, we seek emission in a range where the human eye is most sensitive, i.e., in the green region of the spectrum.

These diodes are also used in optical telecommunications for converting an electrical signal into an optical one. The signal is then conveyed by optical fibers and transformed into an electrical signal by a photovoltaic cell. One wants to use the wavelength where the attenuation in the fiber is a minimum, in general 1.5 μm. For this one uses the compound $Ga_{0.47}In_{0.53}As$. The quantum efficiency in the best cases lies between 10^{-3} and 10^{-2}. All

of this shows the importance of the study of recombination processes for practical optimization of such devices.

The radiative recombinations referred to above are spontaneous recombinations, i.e., incoherent. However we have seen in Appendix 6.1 that they are fairly monochromatic, the relative emission width being of order kT/E_g.

Besides these processes there exists a completely different emission process, stimulated emission which occurs when a photon of energy E_g meets an electron-hole pair of equal energy. There is then emission of a second photon in phase with the first one. If the number of pairs is large enough this process can be cumulative and lead to the emission of coherent monochromatic radiation, i.e., to a laser (Light Amplification by Stimulated Emission of Radiation). A threshold current must be reached to start the phenomenon in a diode. Semiconducting lasers are small and can work from a very simple energy source, an ordinary battery. Quantum well lasers made of GaAs between barriers of $Al_x\ Ga_{1-x}\ As$, or of $Ga_{0.47}\ In_{0.53}\ As$ between InP barriers (cf. Appendix 3.2; Sect. 5.4 and Sect. 9.6) have particularly low threshold currents and are used in fiber optic telecommunications.

9.2 The Metal–Semiconductor Contact in Equilibrium

An understanding of the properties of a junction between a metal and a semiconductor, also named "Schottky diode," becomes necessary as soon as we weld metallic contacts on a semiconductor in order to use it in an electrical circuit. Will the contact be a rectifying one or an ohmic one, i.e., behaving as a pure resistance?

First we study the equilibrium junction and then the effect of applying a voltage. We must use the same energy scale for the quantum states of the metal and the semiconductor. We choose as energy reference the energy of a static electron at infinity (this is the energy zero in the hydrogen atom). We call this the "vacuum level."

Let us begin by some definitions. In a metal we call the work function ϕ_m the energy required to remove an electron that is close to the surface and at the metal Fermi energy E_{Fm}, and leave it in vacuum with zero velocity (see Fig. 9.3). In truth the energy of an electron close to the surface is not the same as that of a static electron at infinity as there is an interaction between the moving electrons of the metal and this electron. This "image charge effect" is small and we shall neglect it in the following.

An analogous definition applies to the semiconductor, and allows us to define its work function ϕ_s. However in a non-degenerate semiconductor the Fermi level E_{Fs} is in the band gap, and there are therefore no electrons at this level. Also the position of E_{Fs} relative to the maximum E_v of the valence band and the minimum E_c of the conduction band depends on the

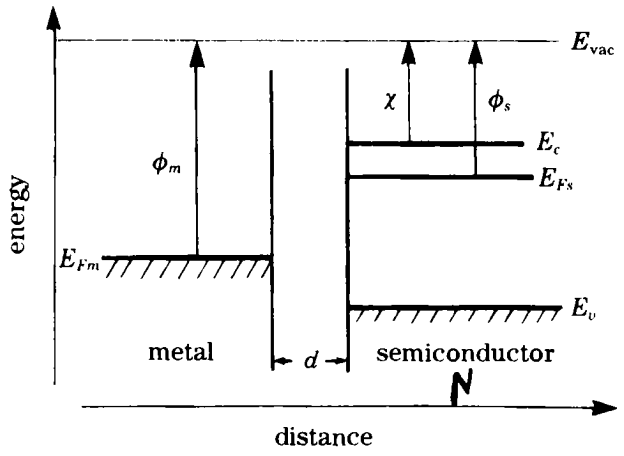


Fig. 9.3. Band profile of separated metal (left) and semiconductor (right). We have assumed that the work function ϕ_m of the metal exceeds that of the semiconductor (ϕ_s). The electron states occupied at $T = 0$ K are hatched.

doping. The work function ϕ_s is therefore not an intrinsic quantity and we prefer another quantity, which is intrinsic, the affinity χ which is the difference between E_c and the vacuum level E_{vac}. We have the following relations:

$$\phi_m = E_{vac} - E_{Fm}, \tag{9.5}$$

$$\phi_s = E_{vac} - E_{Fs}, \tag{9.6}$$

$$\chi = E_{vac} - E_c = \phi_s - (E_c - E_{Fs}). \tag{9.7}$$

In the usual metals and semiconductors ϕ_m, ϕ_s, and χ are a few eV. To form a metal–semiconductor junction we imagine bringing the two solids closer, i.e., we reduce the distance d of Fig. 9.3. We assume in the following that the semiconductor is of type n, and in Sect. 9.2a we take $\phi_m > \phi_s$ as in Fig. 9.3. The other case is treated in Sect. 9.2b.

9.2a $\phi_m > \phi_s$

As d becomes very small the electrons from the semiconductor can pass into the metal by tunnelling until the Fermi levels line up (Fig. 9.4). There is then in the semiconductor a space-charge region of width W, empty of electrons, where there are only the fixed positive charges of the uncompensated donors. As the number of available electron states in the metal is very large, the screening is very strong and only the metal surface is affected. The electron current forms a negative surface charge layer, which, with the positive space charge of the semiconductor, creates an electric field confined almost entirely within the space-charge region.

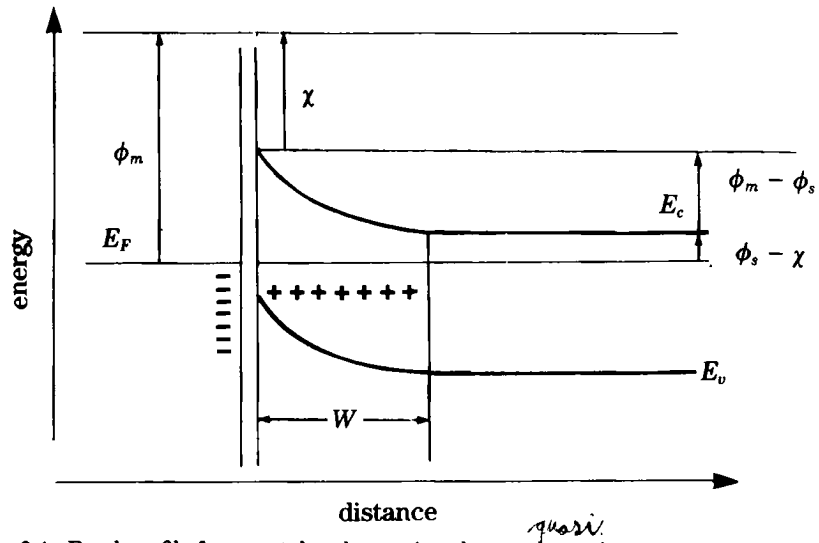


Fig. 9.4. Band profile for a metal and a semiconductor almost in contact.

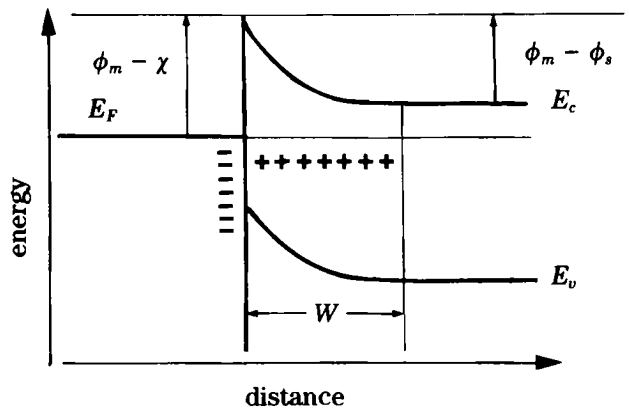


Fig. 9.5. Metal–n-type semiconductor junction in contact ($\phi_m > \phi_s$).

An electric potential and thus a bending of the bands results. In the semiconductor the energy of an electron at the bottom of the conduction band is greater at the surface, where this electron is an amount χ below E_{vac}, than in the interior, where it is $\phi_s - \chi$ above the Fermi level, and thus at $\phi_m - (\phi_s - \chi) = \chi + (\phi_m - \phi_s)$ below E_{vac}. The energy of the band bending is thus $(\phi_m - \phi_s)$. When the contact is made the band profile has the form shown in Fig. 9.5.

As the figure shows, the barrier $\phi_m - \chi$ for crossing from the metal to the semiconductor is larger than the one separating the semiconductor from the metal, which is $\phi_m - \phi_s$. The width of the band bending is found by solving the Poisson equation: in the space-charge region the charge density is eN_d, where N_d is the donor concentration.

From

$$\frac{d^2V}{dx^2} = -\frac{eN_d}{\epsilon_0\epsilon_r} \tag{9.8}$$

and the boundary conditions for $x = W$:

$$\frac{dV}{dx} = 0, \tag{9.9}$$

$$-e[V(W) - V(0)] = -(\phi_m - \phi_s), \tag{9.10}$$

we deduce that the bending is parabolic and that

$$\frac{e^2 N_d W^2}{2\,\epsilon_0\,\epsilon_r} = \phi_m - \phi_s \tag{9.11}$$

or

$$W = [2\,\epsilon_0\,\epsilon_r(\phi_m - \phi_s)/e^2 N_d]^{1/2}. \tag{9.12}$$

This expression is very similar to that for the half-width of a p–n junction (Eq. (8.27)). The accumulated charge per unit area in this region is

$$Q_{sc} = eW N_d = [2\epsilon_0\epsilon_r N_d(\phi_m - \phi_s)]^{1/2}. \tag{9.13}$$

9.2b $\phi_m < \phi_s$ ✗

We now turn to the case where the metal work function is less than that of the semiconductor. When the two solids are far apart the energy level diagram is that of Fig. 9.6, which we should compare with Fig. 9.3. We imagine the two solids brought close together.

When the distance d becomes very small, electrons pass from the metal to the semiconductor: this creates a positive charge layer at the metal surface and a negative one near the semiconductor surface, until there is only a single common Fermi level. At the surface the conduction band remains a distance χ from E_{vac}, as shown in Fig. 9.7.

Because of the existence of these charges there is an electric field and hence a bending of the bands. The transferred electrons occupy the region where E_c is below the Fermi level, and constitute an accumulation layer whose thickness W' is much less than the width W of the band-bending region. Because of the large density of states in the conduction band, a small W' is enough to accommodate the electrons transferred from the metal. Once contact is made, the energy profile becomes as in Fig. 9.8.

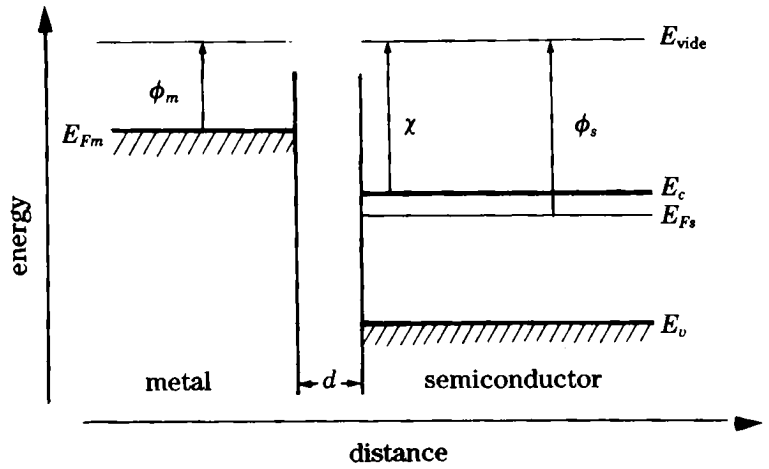


Fig. 9.6. Band profiles of separated metal (left) and an n-type semiconductor (right), in the case $\phi_m < \phi_s$.

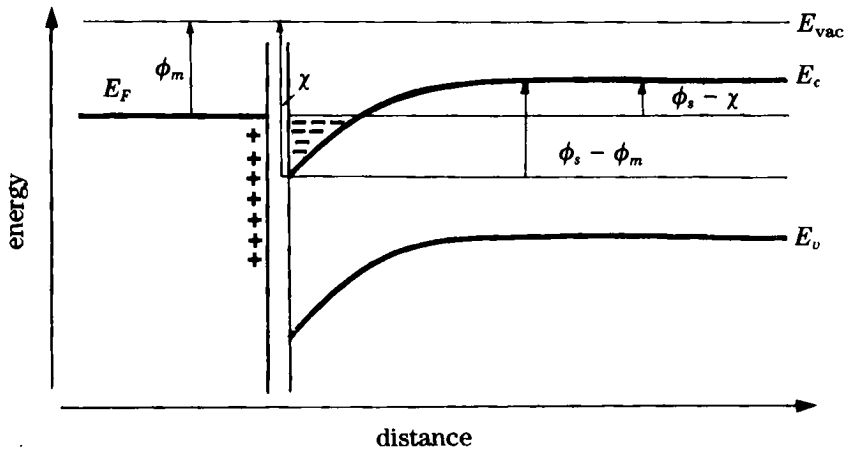


Fig. 9.7. Metal–n-type semiconductor with contact almost made ($\phi_m < \phi_s$).

9.3 Non-Equilibrium Metal–Semiconductor Diode

9.3a $\phi_m > \phi_s$

In this case there is a very resistive space-charge region, and the applied external potential V_s appears exclusively across this region; in the rest of the system the bands remain horizontal on the energy diagram, and are simply shifted. The structure of Fig. 9.5 becomes that of Fig. 9.9(a) for forward bias (potential more negative on the semiconductor side, $V_e > 0$) and that of Fig. 9.9(b) for reverse bias ($V_e < 0$).

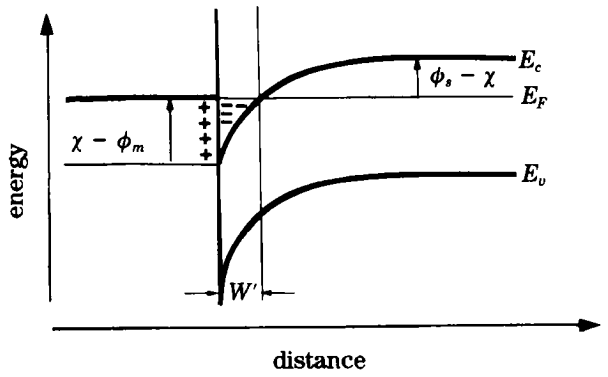


Fig. 9.8. Metal–n-type semiconductor junction in contact ($\phi_m < \phi_s$).

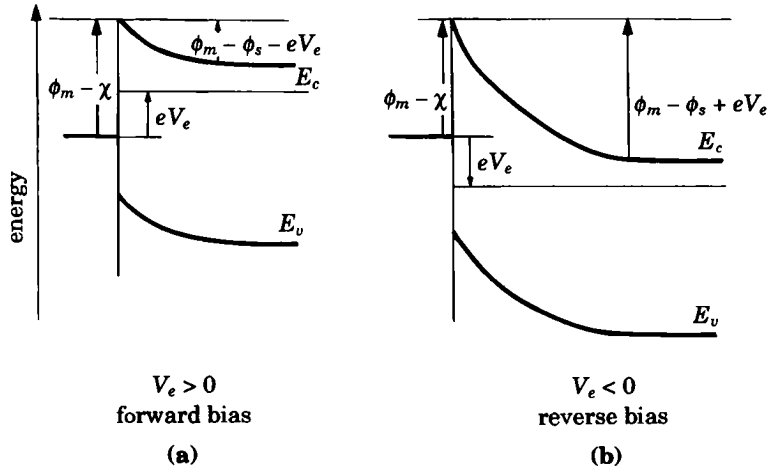


Fig. 9.9. Polarized metal–semiconductor junction ($\phi_m > \phi_s$): (a) $V_e > 0$, semiconductor negatively biased with respect to the metal, forward sense; (b) $V_e < 0$, reverse bias.

In the absence of V_e the current vanishes, so the current J_1 from the metal to the semiconductor exactly balances the current J_2 from the semiconductor to the metal. Applying V_e changes the barrier between the semiconductor and the metal by an amount $-eV_e$; the corresponding current becomes J_2'. However the current J_1 stays the same, since the barrier in the metal–semiconductor direction is unchanged. As the current J_2' corresponds to an activation process, the number of electrons able to cross the barrier is given by a Boltzmann factor, such that

$$J_2' = J_1 \exp \frac{eV_e}{kT}. \tag{9.14}$$

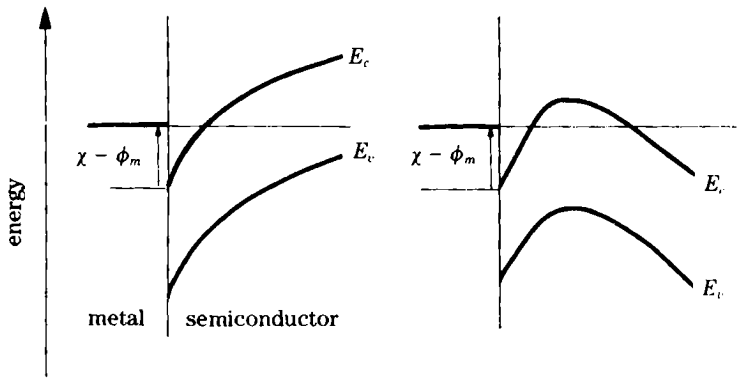
The resulting current $J = J_1 + J_2'$ is

$$J = J_1 \left[\exp\left(\frac{eV_e}{kT}\right) - 1 \right] \tag{9.15}$$

The current–voltage characteristic $J(V_e)$ is analogous to that of a p–n junction and gives rise to rectifying behavior. The saturation current is to first approximation the thermionic emission current of the metal below the $\phi_m - \chi$ barrier.

9.3b $\phi_m < \phi_s$ X

The width of the space-charge region is small, so this region has very low resistance. Thus the voltage V_e appears over the bulk semiconductor and produces a "sloping" band profile. The structure of Fig. 9.8 becomes Fig. 9.10(a) or that of Fig. 9.10(b) according to the sign of V_e.



(a) forward bias (b) reverse bias

Fig. 9.10. Metal–semiconductor junction ($\phi_m < \phi_s$): (a) forward biased, $V_e > 0$; (b) reverse biased, $V_e < 0$. A way of remembering the shape of the figure is to imagine that the band structure is made of rubber bars attached to the metal side and one seeks to bend them up or down to the right.

For forward bias, there is no barrier to cross from the semiconductor to the metal since near to the contact all of the electron states are full.

For reverse bias, there is a barrier to cross of order $\phi_s - \chi$, the equilibrium distance of the Fermi level from the conduction band. In a semiconductor of typical n doping, this barrier is no more than one hundred meV; it is

thus easily crossed by the bias electrons. In consequence the current passes in a similar way for both: one observes an approximately ohmic behavior (non-rectifying).

In summary, a metal–n-type semiconductor junction behaves very differently, depending on whether the work function ϕ_m of the metal is less than or greater than that of the semiconductor (ϕ_s). If we want to make a simple ohmic contact on a given semiconductor we must choose a metal whose work function ϕ_m is less than ϕ_s. By contrast if we wish to make a diode we take a metal with ϕ_m greater than ϕ_s.

We can easily show that these conclusions reverse if we consider metal contact with a p-type semiconductor. Note that the above discussion neglected the presence of surface states in the semiconductor. We will consider their effect in Sect. 9.4.

We note finally that in contrast to the p–n junction, the currents crossing metal–semiconductor junctions are always majority carrier currents.

9.4 The Semiconductor Surface

A free surface constitutes a rupture of the periodicity of the solid. The crystal potential vanishes outside the solid, thus, *a priori*, we would expect to represent it as the dashed curve in Fig. 9.11 near the surface. In reality quantum mechanics predicts a non-zero probability of finding the electron in the vacuum near the surface. There is then a lack of negative charge, thus a positive charge, in the first atomic layers of the solid. We thus have a surface dipole, whose effect is to create a potential jump. The shape of the potential is therefore that of the full curve in Fig. 9.11(b), the surface potential jump being directly related to the work function of the solid.

Moreover the presence at the surface of dangling (unsaturated) bonds or atomic rearrangements induces surface electron states. These states can be acceptors or donors. In usual semiconductors, e.g., Si or GaAs, the density N_S of these states is very high, of the order of a fraction of a state per surface atomic site. For example, at the surface (110) of GaAs where the total atom density is 8.8×10^{14} cm^{-2}, the density N_S of quantum surface states is of order 10^{13} cm^{-2}.

Let us consider the effect on the band profile of the introduction of surface states. We start with a semiconductor with no surface states: its bands are flat (Fig. 9.12(a)). We then consider the introduction of a few surface states, e.g., surface acceptors whose energy level is Δ below the conduction band, in an n-doped semiconductor. Electrons from the bulk will occupy these states since their energy is below the Fermi level.

There will remain an uncompensated charge on the semiconductor, and thus a depletion zone of height ϕ_b as in Fig. 9.12(b). Charge neutrality requires these surface charges Q_S to be compensated by the total volume charge:

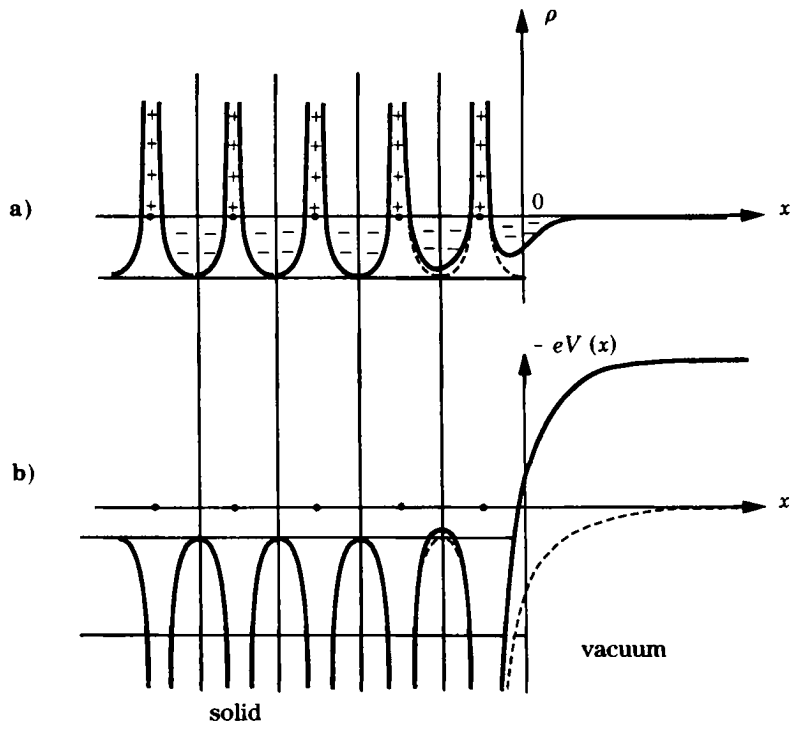


Fig. 9.11. (a) Charge density ρ and (b) electrostatic energy near a solid–vacuum interface. The dashed curves take no account of the probability of finding an electron just outside the solid, while the full curves correspond to the real situation. The dashed and solid curves differ only within two interatomic distances of the surface.

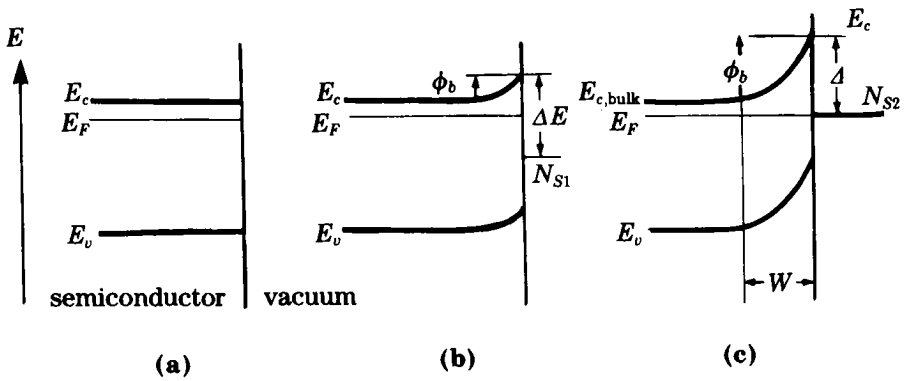


Fig. 9.12. Band profile at a semiconductor–vacuum interface: (a) semiconductor without surface states; (b) low density N_{S1} of surface states; (c) high density N_{S2} of surface states. The Fermi level is then said to be pinned by these states.

$$Q_{\text{vol}} = (2\epsilon_0\epsilon_r N_d \phi_b)^{1/2} = -Q_S. \tag{9.16}$$

For average doping, $N_d \sim 10^{22}$ m^{-3} and $\phi_b \sim 1$ eV, the charge accumulated in the space-charge region is of the order of 3×10^{11} electrons per cm^2. **We obtain a remarkable result: the charge which can be accumulated in the volume, for fixed doping, is limited, as ϕ_b is always smaller than E_g.** Consequently the surface charge Q_S is limited, independently of the number of surface states N_S. These surface states are thus only partially filled if their number exceeds $(\epsilon_0\epsilon_r N_d \Delta)^{1/2}/e$: the Fermi level is thus a few kT from the surface quantum levels. This is the situation depicted in Fig. 9.12(c). We say that the Fermi level is **pinned** by the surface states. In fact the Fermi level of the solid is still determined by the concentrations of dopants and not by the surface. As the surface states must be within a few kT from the Fermi level, it is the band bending ϕ_b that adjusts its value to fulfill this condition. The Fermi level is at a distance ΔE (to within a few kT) of the surface conduction band. We have

$$\Delta E = \phi_b + E_{c,\text{bulk}} - E_F, \tag{9.17}$$

where $E_{c,\text{bulk}}$ denotes the position of the bulk Fermi level. If we make a junction between a metal and this semiconductor the height of the barrier from the semiconductor side will be ϕ_b, independent of the work function of the metal.

The possible existence of surface states may have a considerable influence on properties of semiconductor interfaces, junctions. In particular the discussion of Sect. 9.3 on the metal–semiconductor diode has to be modified to account for their influence.

9.5 Photoemission from Semiconductors

In the photoemission process, or photoelectric effect, a photon gives energy to a bound electron and releases it from the solid. Let us consider a p-type semiconductor. If the electron is in the valence band (Fig. 9.13) the required energy is $E_g + \chi$, where χ is the affinity defined in Sect. 9.2. For p doping the bands bending caused by the pinning of the Fermi level is downwards, so that the surface affinity is $\chi_{\text{surf}} = \chi_{\text{volume}} + \phi_b$. For a clean surface $\chi_{1\text{bulk}}$ and $\chi_{1\text{surface}}$ are of order 5 eV, and only an ultraviolet photon has enough energy $h\nu_1$ to liberate electrons.

We have seen in Sect. 9.4 that the value of the work function or the affinity is determined by the size of the surface charge dipole. This dipole can be reduced by the surface deposition of a monolayer of cesium, an easily ionized alkaline metal. The vacuum level is shifted relative to the semiconductor bands, and we can even reach a situation in which the vacuum level $E_{2\text{vac}}$ is lower than the level E_c of the bottom of the conduction band in

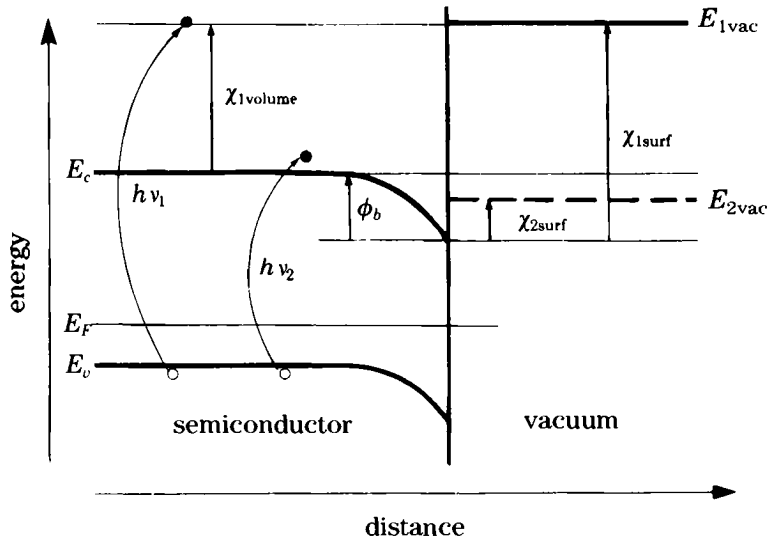


Fig. 9.13. Vacuum–p-type semiconductor interface and the lowering of the affinity. For an excitation energy $h\nu_2$ near the forbidden band, photoemission does not occur unless the affinity has been reduced so that the vacuum level E_{2vac} is lower than E_c.

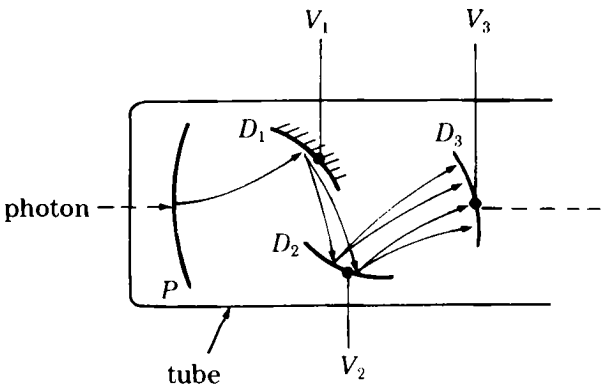


Fig. 9.14. Schematic view of a photomultiplier tube. The photoemissive surface P is followed by dynodes $D_1, D_2, D_3,...$ held at potentials $0 < V_1 < V_2 < V_3,...$ which multiply the electrons. (From Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

the bulk of the solid. The affinity is then $\chi_{2surface}$. To remove an electron it therefore suffices to excite it with a photon of energy $h\nu_2$ just exceeding E_g, and thus in the visible range. If the electron is excited into the conduction band close enough to the surface it may be emitted into vacuum: this is possible if the electron reaches the surface after its random walk in the bulk, before recombination.

This is the principle of the photocathodes in GaAs photomultipliers. These are very sensitive light detectors, whose quantum efficiency can reach 30%, or 0.3 electrons per photon. After being extracted from the cathode by photoemission, the electron passes through several multiplier stages, called

dynodes, held at increasing positive potentials: at each stage an electron gives rise to several secondary electrons. This is shown in Fig. 9.14. These detectors, which can be used in the visible or the ultraviolet range (for any photons with energy larger than the band gap of the semiconductor) are very sensitive and can count single photons.

9.6 Heterojunctions

In Chap. 8 we studied the $p-n$ junction, made of two semiconductor samples of the same chemical composition but different dopings ("homojunction"). If we have two semiconductors A and B differing in their chemical composition and possibly their doping, but with sufficiently similar crystalline lattices that epitaxy, i.e., a continuity of the crystal lattices of A and B, is possible, we can form a "heterojunction."

We take the example of a junction formed from p-type GaAs and n-type $Al_x\ Ga_{1-x}\ As$. Figure 9.15 shows the band structure when the two semiconductors are apart.

The material with the larger band gap is $Al_x\ Ga_{1-x}\ As$. The vacuum level, i.e., that of a free electron at rest outside the solid, is the same for the two materials, and the affinities are $e\chi_1$ and $e\chi_2$. The difference between the two band gaps $E_{g2}-E_{g1}$ is divided unequally between the valence bands (ΔE_v) and the conduction bands (ΔE_c). In non-degenerate semiconductors these quantities do not depend on the doping. Once the contact is made the Fermi levels are aligned. An internal potential difference $e\phi = E_{F2} - E_{F1}$ appears here too. Figure 9.16 corresponds to a heterojunction made from the materials depicted in Fig. 9.15.

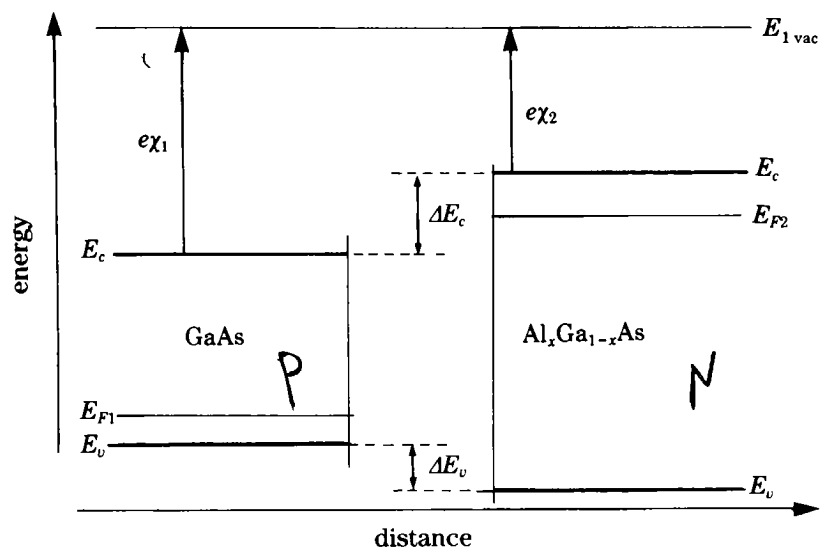


Fig. 9.15. Band profiles of two different semiconductors when far apart from each other: on the left GaAs (p type); on the right $Al_x\ Ga_{1-x}\ As$ (n type).

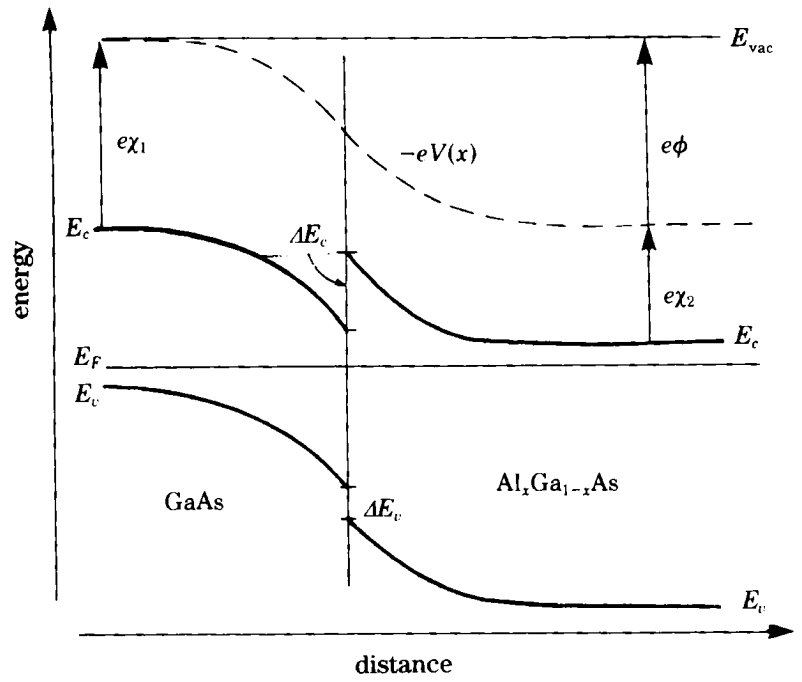


Fig. 9.16. Band profile of a heterojunction formed from the materials of Fig. 9.15. The dashed curve shows the behavior of electrostatic potential energy.

We note from Fig. 9.16 the asymmetry between the electrons and holes for such a band profile: in the conduction band we have a potential well and a barrier, and in the valence band a simple discontinuity, which is increased by ΔE_v with respect to the homojunction. We can guess that this increase of the height of the potential barrier will modify the transport of holes in the p–n direction, and that moreover the GaAs electrons will tend to accumulate in the (triangular) conduction well. The width of this well is a fraction of the width of the space charge W, i.e., a few tens of nanometers, depending on the doping. We thus obtain a quasi-two-dimensional electron layer.

Heterostructures are in current use. We noted in Sect. 5.4b that by selective doping of superlattices, which are periodic repetitions of heterojunctions, one can obtain extremely high electron mobilities, the electrons and ionized donors being spatially separated.

10.

The Principles of Some Electronic Devices

In this last chapter we apply the concepts introduced in this book to explain the operation of several electronic devices which make use of junctions: the junction transistor, the field-effect transistor (FET), the junction FET, and the MOSFET. Some problems in Appendices 10.1, 10.2, and 10.3 give a more quantitative description. We shall describe the principle of integration and planar technology, and we shall discuss the concepts and hopes of band structure engineering, which allows the manufacture of semiconductors conceived on paper for particular uses. Miniaturization of circuits and memories has been the great technological revolution of the last 40 years. The physical limitations of this revolution are briefly discussed at the end of this Chapter.

10.1 The Junction Transistor

A junction transistor is a single crystal containing two junctions (in principle monocrystalline) of opposite polarities in series. We can thus have p–n–p or n–p–n transistors. Here we describe the operation of a p–n–p transistor. The results apply without restriction to n–p–n structures if we change the directions of the currents and voltages and permute the symbols p and n. These transistors are also called bipolar transistors, as their operation relies on the existence of two types of carriers.

In Fig. 10.1 three regions are shown: the p^+-doped emitter, the weakly n-doped base, and the moderately p-doped collector.

The base is smaller than the diffusion length in a recombination time. A forward voltage is applied to the emitter-base junction, and reverse voltage to the base-collector junction. We shall confine ourselves to a semi-quantitative description of the working of this junction transistor.

The applied emitter-base voltage is positive and the applied base-collector voltage is positive and large. Thus:

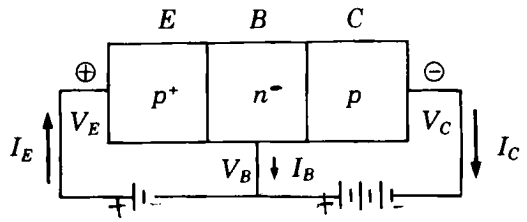


Fig. 10.1. Bias of a p–n–p transistor and sign conventions for the currents.

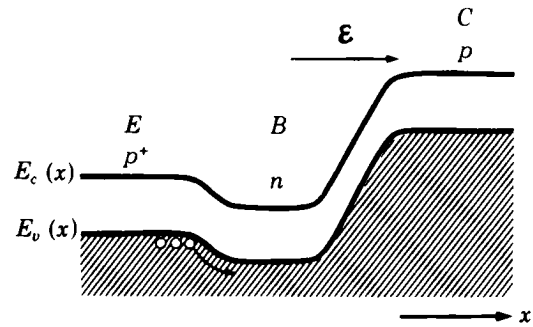


Fig. 10.2. Schematic band diagram for a biased p–n–p transistor: E = emitter; B = base; C = collector. (From Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

$$V_B - V_C >> V_E - V_B \geq 0. \tag{10.1}$$

Figure 10.2 shows the band diagram for a transistor biased in this way. The detailed equations governing the working of a junction transistor are given in Appendix 10.1. Consider a flux of particles within the transistor as shown in Fig. 10.3. Since the emitter-base junction is forward biased, there is an injection of holes into the base (flux 1 of the figure). The holes diffuse (flux 2) without recombining if the base is small. However at the base-collector junction there is a strong field that sweeps up the holes reaching the base-collector space-charge region (flux 3). In fact, a few holes diffusing into the base will recombine with electrons, the majority carriers within the base (which is weakly n-doped); this is the flux 4 of holes, which do not reach the collector. Electrons must therefore be supplied by the external circuit to replace electrons recombining with the holes. This flux is represented by the dashed arrow (flux 5). Also there are electrons injected from the base to the emitter in the forward biased emitter-base junction, but this injection is small as the base is weakly n-doped. This is flux 6.

Overall the relation between the emitter-base current and the voltage differs little from the characteristic equation for a junction, i.e.,

$$I_E = I_s \left[\exp\left(\frac{V_E - V_B}{kT} \right) - 1 \right]. \tag{10.2}$$

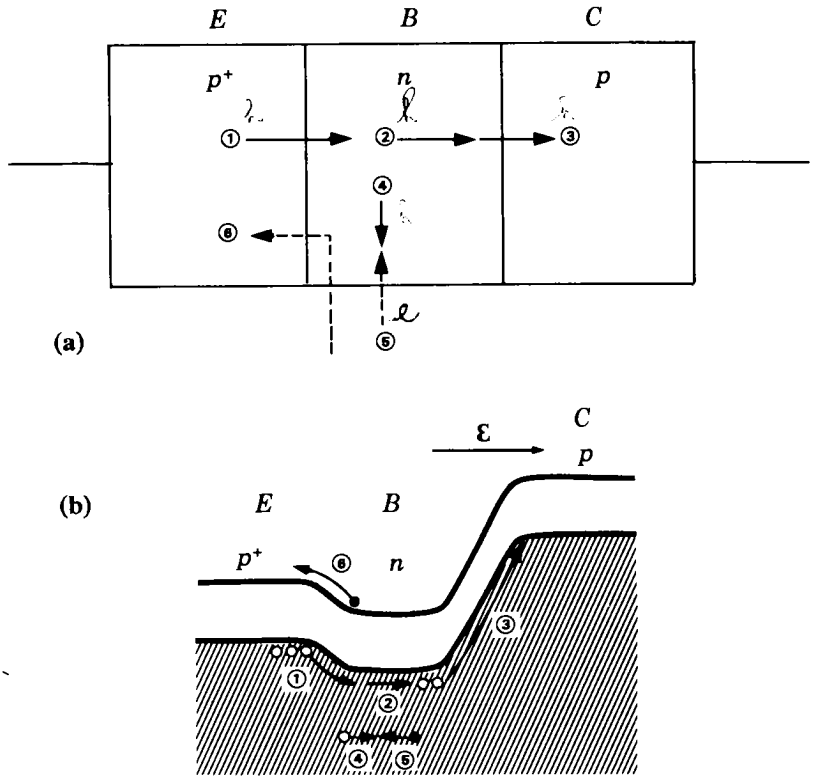


Fig. 10.3. Schematic view of particle fluxes in a p–n–p transistor: (a) across the leads; (b) on the band diagram. The arrows denote the particle currents (thick for electrons and dashed holes). We note that the electric currents corresponding to (5) and (6) add, respectively, to (4) and (1). (After Dalven, "Introduction to Applied Solid State Physics," Plenum Press, 1980.)

On the other hand, we see that all the fluxes are proportional and that I_C is a little smaller than I_E. We set

$$I_C = \beta\, I_B. \tag{10.3}$$

Since

$$I_E = I_C + I_B, \tag{10.4}$$

$$\frac{I_C}{I_E} = \frac{\beta}{1+\beta}. \tag{10.5}$$

The factor β is called the current gain of the transistor. This factor is typically 100 and shows that the current I_C is only slightly less than I_E. A large current gain requires a very thin base, and the diffusion coefficient and

recombination time must be large. The effective width of the base depends on the technology used in manufacturing the transistor. But it also depends on the doping of the base. We have seen that if the base is weakly doped the space-charge region will extend further on each side of the base, leaving at the center a narrower neutral zone for the holes to cross. The lifetime of the minority carriers will be longer in a semiconductor with an "indirect gap" where the recombination probability is lower (cf. Appendix 6.2) as the optical absorption is weak. Silicon is a good example.

Application: Transistor Amplifier

Here we shall describe a very simple example of the use of a p–n–p transistor as an amplifier, with the further aim of introducing planar technology, the basis of integration at both small and large scales (cf. Sect. 10.4). The symbols shown in Fig. 10.4 are used to represent p–n–p and n–p–n transistors in circuit diagrams; the base is the connection to the left, and the emitter is represented by the arrow, whose sense is the forward sense of the base-emitter junction. The simplest amplifier then consists of the arrangement in Fig. 10.5. The transistor is suitably biased by the dc voltages V_1 and V_2, and we can use Eqs. (10.2) and (10.3). The generator produces an input signal of small amplitude s which may, for example, be sinusoidal. We can write down Ohm's law for the base loop

$$s + V_2 = (V_B - V_E) - R_E\, I_E. \tag{10.6}$$

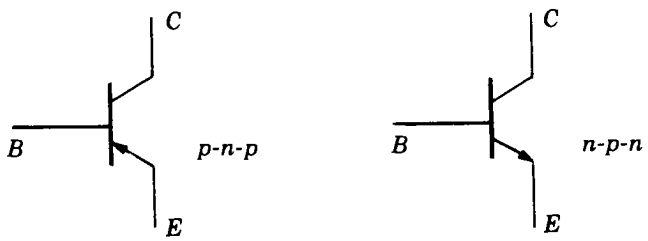
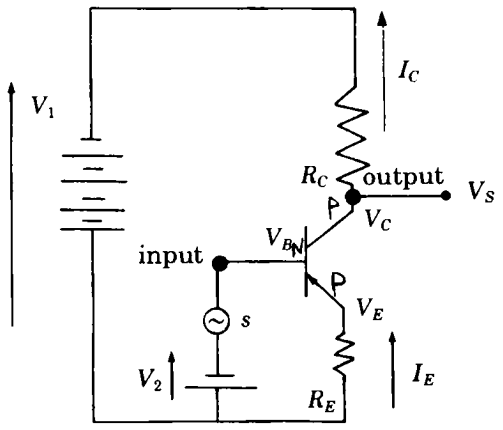


Fig. 10.4. Conventional representation of p–n–p and n–p–n junction transistors. The sense of the arrow is that of the direct electric current across the emitter-base junction.

In this expression $(V_B - V_E)$ is always small compared with the other voltages in the circuit as the junction is forward biased. Thus using Eq. (10.5),

$$I_C = \frac{-\beta}{\beta+1}\frac{(s+V_2)}{R_E} \sim \frac{-(s+V_2)}{R_E}. \tag{10.7}$$

$V_e \simeq I_e R_e$



AMPLIFICATORE

Fig. 10.5. Arrangement of a p–n–p transistor as an amplifier.

There is therefore a small sinusoidal current $\sim -s/R_E$ in the collector and thus an output voltage signal $-sR_C/R_E$ at the collector, the output terminal of the amplifier. The voltage gain is therefore

$$|G| = \frac{R_C}{R_E}. \tag{10.8}$$

One can easily obtain gains of the order of 100. An n–p–n transistor is (or rather was) made by pulling a large crystal from an n-doped bath whose composition is changed by adding acceptors in sufficient numbers, and then again by adding donors in still larger numbers. If necessary this operation can be repeated several times, since as we saw in Chap. 4, what matters is $N_d - N_a$ in the n material and $N_a - N_d$ in the p material. We thus obtain a cylindrical bar alternately n and p doped, which we cut off and arrange as in Fig. 10.6.

In the device studied above the intrinsic limit to the operation speed of the transistor is the transit time of the minority carriers in the base, i.e., the time taken for the injected holes to diffuse to the base collector junction. This time is given by

$$\tau = \frac{d^2}{D}, \tag{10.9}$$

where d is the base thickness. With a very thin base of 10^{-6} m this time is about 10^{-9} s for silicon. If we wish to make very high-frequency transistors we must use a material such as gallium arsenide GaAs where the mobility, hence the diffusion coefficient, is higher than in silicon and a thin base.

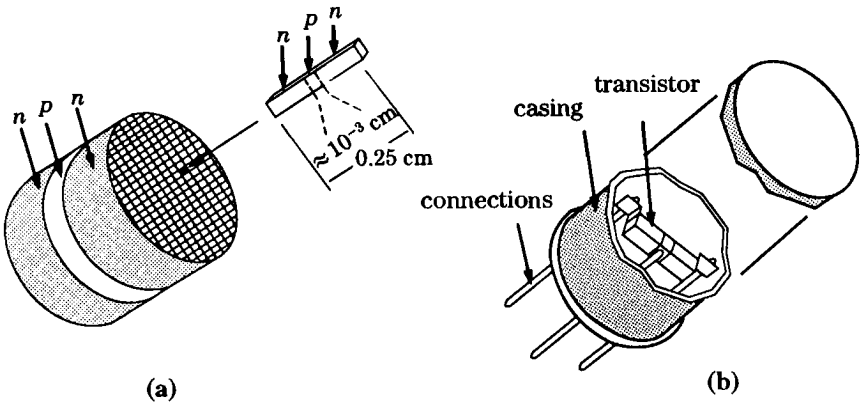


Fig. 10.6. Manufacture (a) and mounting (b) of an n–p–n transistor. (From Marmstadt and Enke, "Digital Electronics for Scientists," Benjamin, 1969.)

10.2 The Field-Effect Transistor

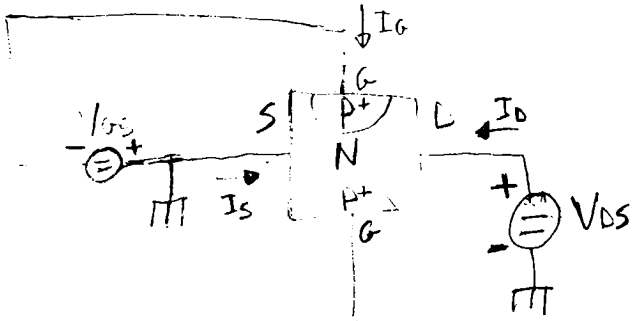
This frequency limitation does not exist in devices whose functioning relies only on the majority carriers, such as field-effect transistors (FETs). In a FET, the current is controlled by an applied voltage: such a transistor is essentially a resistance whose value is controlled by the applied voltage.

10.2a The Junction FET

Figure 10.7(a) shows a rectangular bar of n-type silicon with two metallic contacts at its ends, the source and drain. It acts as a resistor. Suppose now that we create above and below this resistor two p^+–n junctions as shown in Fig. 10.7(b). We reverse bias these junctions using a contact. Then the original conductance of the resistor

$$G = \frac{1}{R} = \frac{\sigma W l}{L} \tag{10.10}$$

is reduced as the width of the neutral region where the electrons can flow is decreased by the width of the space charge. This in turn is controlled by the voltage applied at the gate G. We thus obtain a resistor controlled by the voltage. The problems of Appendix 10.2 analyze the junction FET in greater detail.



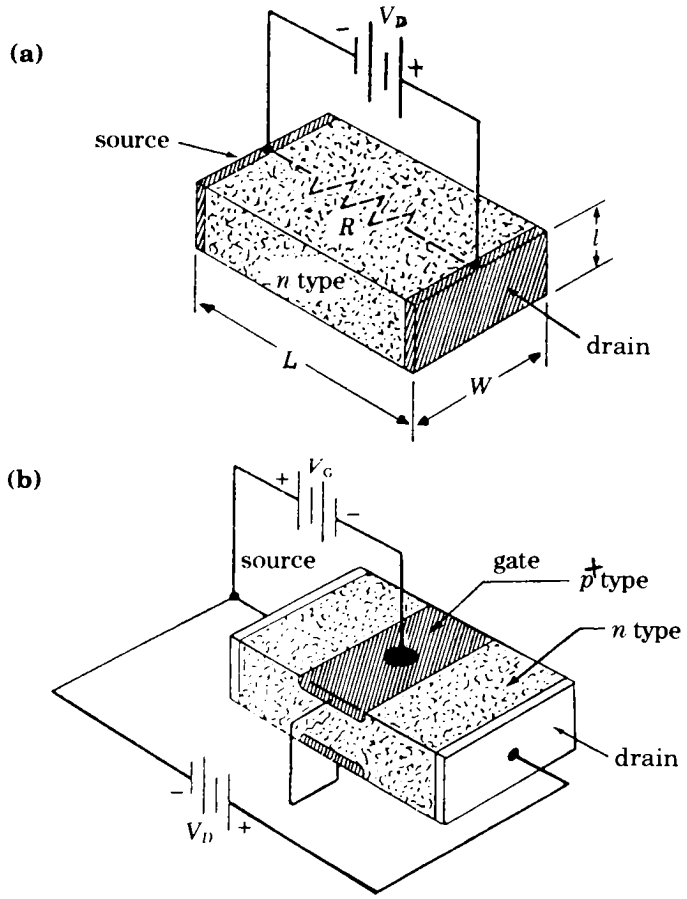


Fig. 10.7. Principle of a junction FET. (a) Silicon resistor; (b) schematic picture of a junction FET. (After Marmstadt and Enke, "Digital Electronics for Scientists," Benjamin, 1969.)

10.2b The MOSFET

The MOSFET (Metal-Oxide–Semiconductor Field-Effect Transistor) is a unipolar device where the resistance of a semiconducting channel is controlled electrostatically. The idea is to use a capacitor, one plate of which is metallic and the other being the semiconductor. The structure is shown in Fig. 10.8(a).

For a silicon structure the insulator is a layer of silica (SiO_2) obtained simply by oxidation of the semiconductor. When a voltage is applied across such a system a charge appears on the two plates of the capacitor. The car-

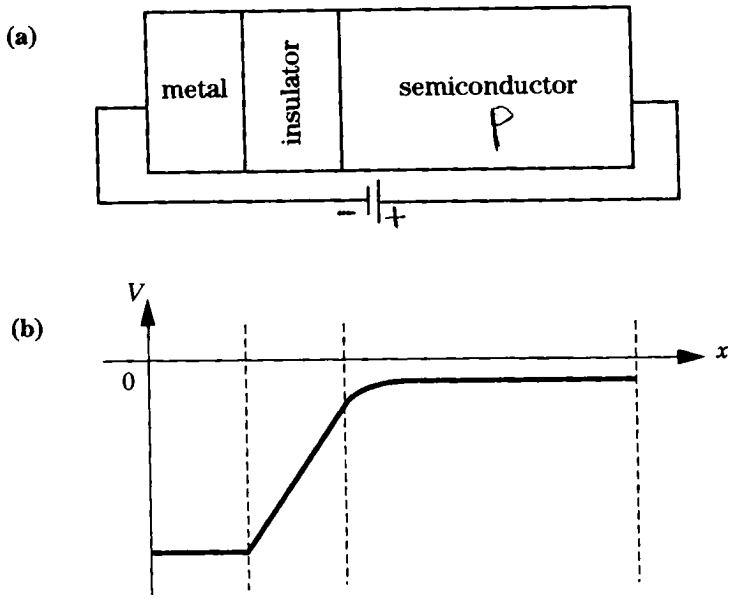


Fig. 10.8. (a) Structure of a MOSFET; (b) behavior of the potential across the structure.

riers distribute themselves so as to create a total charge opposed to that at the metal surface. There thus appears a space charge near the semiconductor surface. We shall see that depending on the sign and size of the charge, we may have accumulation or depletion of free carriers, or even inversion, i.e., the appearance of a layer of minority carriers at the surface. In this system the semiconductor carriers are in thermodynamic equilibrium, since there is no current in the semiconductor. This is possible, even though the Fermi levels in the metal and the semiconductor differ, because the presence of the insulator prevents charge transport perpendicular to the surface of the device.

The equations governing the charge distribution are thus the Poisson equation (Eqs. (8.2) and (8.3)) (8.8) and the law $np = n_i^2$. The problem is the same as that of the equilibrium junction. We confine ourselves here to a qualitative description. Appendix 10.3 gives a quantitative treatment.

Figure 10.8(b) shows the behavior of the potential in this structure for the case where the metal is negative with respect to the semiconductor ($V_m - V_s < 0$). In this case the negative potential of the metal attracts holes. The energy levels of the semiconductor increase near to the insulator. The band scheme is shown again in Fig. 10.9(a) for p-type material. We see that the Fermi level, which is constant since we are in equilibrium, is closer to the valence band at the surface and there will be accumulation or enhancement of hole concentration at the surface.

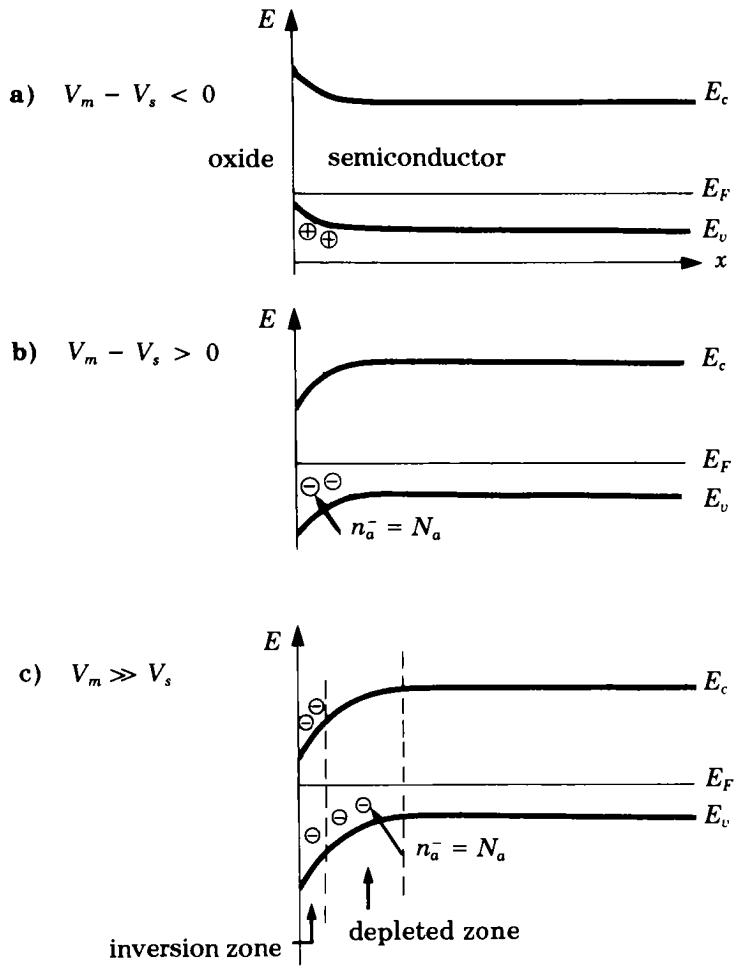


Fig. 10.9. Energy levels in the semiconductor of a MOSFET as a function of distance from the insulator: (a) negative voltage applied to the metal; (b) small positive voltage; (c) large positive voltage.

By contrast in the case where $(V_m - V_s)$ is positive, the holes are repelled and the band scheme is shown in Fig. 10.9(b). At the surface the Fermi level is further from the band edges, and there is depopulation, or depletion, of mobile carriers. Thus there remain only ionized acceptors near the surface with concentration N_a.

Above a certain threshold, for large positive voltages applied to the metal we obtain the band scheme shown in Fig. 10.9(c): there are three

zones. At the surface the Fermi level is close to the conduction band, and electrons appear. This is called an inversion layer. The electron number in this inversion layer depends directly on the voltage applied to the metal. Deeper in the semiconductor we encounter a depleted region where the only charges are ionized acceptors, before reaching the neutral semiconductor.

The operation of a MOSFET makes use of these phenomena. The structure of a MOSFET is shown in Fig. 10.10(a). Two contacts of types opposite from the chosen substrate (here p type) constitute the source and the drain. These contacts bound the active region of the MOSFET which is situated below the gate. The gate, oxide, and substrate constitute the structure (metal, oxide, semiconductor) discussed above. The source is linked to the substrate. In the absence of sufficient positive polarization at the gate, the substrate remains p type everywhere and there is no conduction between the source and the drain, as the source-substrate and substrate-drain junctions are back-to-back.

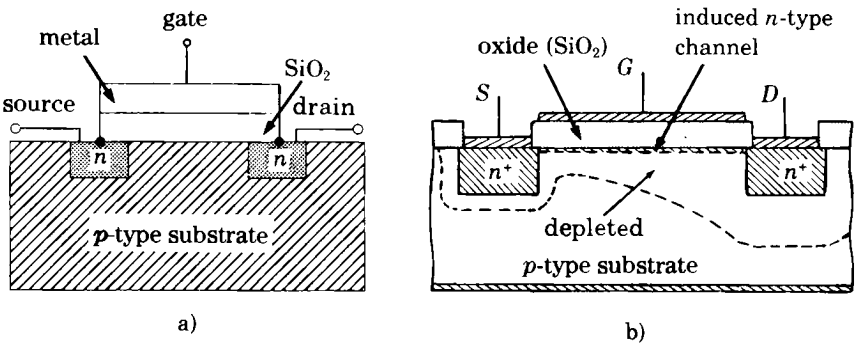


Fig. 10.10. MOSFET transistor with n-type channel: (a) structure; (b) effect of bias. (After Leturcq and Rey, "Physique des Composants Actifs à Semi-Conducteurs," Dunod, 1978.)

In contrast, if a sufficient positive voltage is applied to the gate, one can create a surface inversion layer, called the induced channel—see Fig. 10.10(b). There is then continuity of n-type conduction between source and drain and a source-drain current can flow. The resistance between source and drain is therefore controlled by the value of the applied voltage between gate and source. We can thus construct a device whose current is controlled by a voltage and not a current as in the junction transistor. This constitutes the n-MOS, since the induced channel is n type; the device operates in the enhancement mode, since applying $V_G > 0$ makes electrons appear.

Exercise: Can an n-MOS enhancement transistor work as an n-MOS depletion transistor under certain circumstances?

Present electronic logic provides the logical functions NO, AND, OR by use of inverters, which are just MOS transistors or combinations of

MOS, conducting or not, depending on the voltage applied to the gate. The advantage of these circuits over bipolar transistors is that, because of the oxide layer, no current flows between the gate and the conducting channel.

The complementary MOS inverter, called CMOS (see Appendix 10.3, end of the solutions), is the basic element of logic circuits. It consists of an electron-channel MOS transistor (n-MOS) and a complementary hole-channel transistor (p-MOS) which behaves exactly opposite to an n-MOS transistor under voltages applied to its gate. Whatever the applied voltage, one of the transistors is always blocked: in principle no electric current flows, implying very small dissipation.

10.3 An Application of the MOSFET: The Charge-Coupled Device (CCD)

We have just seen that in a MOSFET made of a p-type semiconductor we can create a potential well for the electrons (Fig. 10.9(c)) in the inversion layer if we apply a strong positive voltage to the metal. The accumulation and storage of minority carriers in a surface potential well are the basic principle of CCD cameras which are now the most common image converter in video systems.

The structure involves a p-type semiconductor covered by an insulating layer and a series of very close metal gates. When we apply to the system of Fig. 10.11(a) positive potentials $V_1 = V_3 = V_4, V_2 > V_1$, a potential well for the electrons appears under electrode 2.

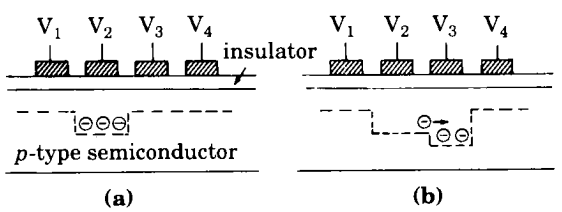


Fig. 10.11. Schematic figure of a CCD. The metallic gates are hatched. The potential profile of the electrons is shown dashed. The potentials applied in (a) are such that $V_2 > V_1 = V_3 = V_4$, and the potential well is under electrode 2. In (b), $V_3 > V_2 > V_1 = V_4$, and the electrons transfer from electrode 2 to 3.

Let us assume that some electrons are introduced under electrode 2, where they are stored: as there are no holes in the inversion layer there is no recombination. If the potentials are changed so that $V_1 = V_4$, $V_1 < V_2 < V_3$, the charges move towards the deeper potential well, ending up under electrode 3 (Fig. 10.11(b)).

By applying an appropriate sequence of voltages on the various gates, we can transfer the charges from one region to another of the surface, and scan it point by point and line by line. We can identify which charges come from which region by analyzing the resulting sequence of electrical signals.

The minority carriers (electrons in this figure) can be produced by luminous excitation. Each surface element or pixel provides a charge, and we thus have the conversion of an image consisting of a set of light or dark points into an electro-optical signal. This provides an electrical signal whose amplitude varies with the light intensity at various points of the original image.

Charge-coupled (CCD) devices are very sensitive light detectors. They have now replaced Vidicon tubes in cameras and kinescopes, and they are used in spectroscopy and astronomical detection, where, for example, arrays of 400×1200 pixels with a quantum efficiency of 0.5 electrons per photon are currently used.

10.4. Concepts of Integration and Planar Technology

The recent and spectacular development of microelectronics is related to the use of planar techniques and the consequent idea of an integrated circuit. Consider the amplifier whose circuit diagram is given in Fig. 10.12. It consists of a p–n–p transistor, resistors, and capacitors. We have seen that a homogeneous semiconductor can behave as a resistor and that a reverse biased junction behaves like a capacitor for ac current. Thus the circuit (Fig. 10.12(a)) can be constructed from fused elements, but all made from silicon as shown in Fig. 10.12(b). The passage to Fig. 10.12(c) is obvious as it is not necessary to link semiconductors of the same type by metallic wires. We are led to a circuit of three elements. Deforming these elements we can combine them into a single semiconducting layer containing n, p, and intrinsic regions, the latter being insulating (Fig. 10.12(d)). It is these same ideas that are currently used in the manufacture of electronic devices from audio sets to computer components: microprocessors, memories, etc.

These circuits are constructed on a substrate: a thin slice, or "wafer," of monocrystalline material of adequate resistivity. The wafers are several cm (typically 10) in diameter and a thickness of several 10^{-1} mm. Their manufacture involves the operations of oxidation, masking, diffusion, metal deposition *in vacuo*, and possibly epitaxy, i.e., the growth of a monocrystalline layer, of adjustable doping on a monocrystalline substrate.

An important step involves the ability to create by "lithography" an oxide layer of a given design on the silicon wafer. This layer acts either as an insulator or as protection for the substrate in a later operation, e.g., diffusion.

The first operation is to (a) oxidize the substrate uniformly (Fig. 10.13). Then a photoresist resin which polymerizes in light is deposited on the

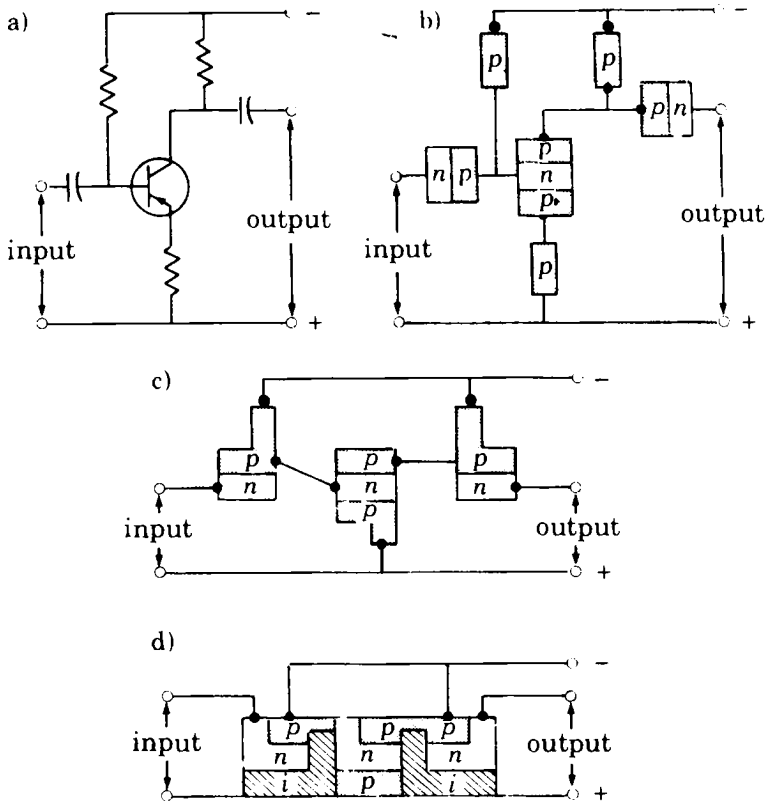


Fig. 10.12. (a) Schematic view of a p–n–p transistor used as an amplifier; (b) circuit made of fused elements; (c) suppression of connecting wires; (d) planar technology. (After Brophy, "Semiconductor Devices," George Allen and Unwin, 1964.)

oxide (b). The system is then illuminated through a mask (c) and the resin polymerizes at the places that receive light. The non-polymerized resin is then dissolved (d) and the silica attacked by an acid that does not affect the polymerized resin (e). Finally the resin is dissolved in an appropriate solvent (f).

The manufacture of an n–p–n transistor is illustrated in Fig. 10.14. An epitaxial n layer is grown on an n^+ substrate. (a) After oxidation, a window is removed by photoetching. Acceptors (e.g., boron) are then diffused through this window. (b) After reoxidizing, a new smaller window is removed, through which phosphorus is diffused to make the n^+ layer, the emitter (c). It remains to make the metallic contacts. The device is oxidized and windows cut over the active regions. A metallic layer is deposited over the device from which we eliminate superfluous zones by photoetching (d).

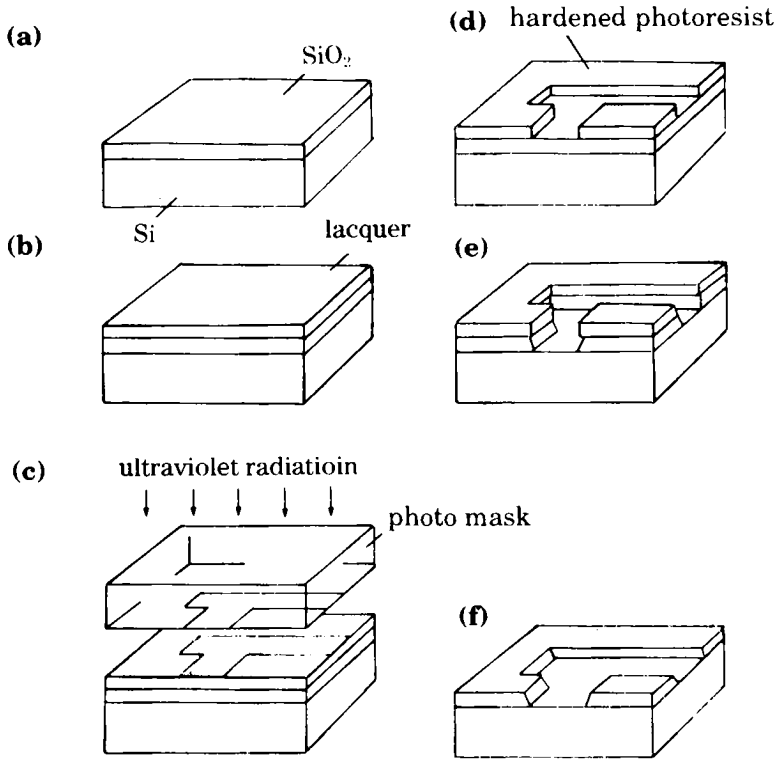


Fig. 10.13. The photolithography process (see the text). (After Tom Forrester, ed. "The Microelectronics Revolution," Basil Blackwell, 1980.)

Note that the exact shape of the base volume is not really limiting, because diffusion currents always follow the concentration gradients, whatever their geometrical distribution.

These techniques allow us to produce from a limited set of operations a considerable variety of logic circuits with complex functions, known colloquially as "chips." One can manufacture connected sets of transistors, resistors, and capacitors, but integration allows us to conceive devices without discrete equivalents, e.g., transistors with multiple emitters. This last technique is called I^2L (Integrated Injection Logic).

Two parameters govern progress in this domain: the size of an element and the speed of operation. The minimum size of an element is determined by the wavelength of the electromagnetic radiation used for the photolithography. Diffraction effects limit the definition of the image to around half a wavelength. The size of an element is currently of the order of 2 μm or about ten times the wavelength of the ultraviolet light used for the photolithography. A transistor and its contacts occupy between 10^{-4} and 10^{-3} mm^2. The use of light of shorter wavelength (UV or X-rays from synchrotron ra-

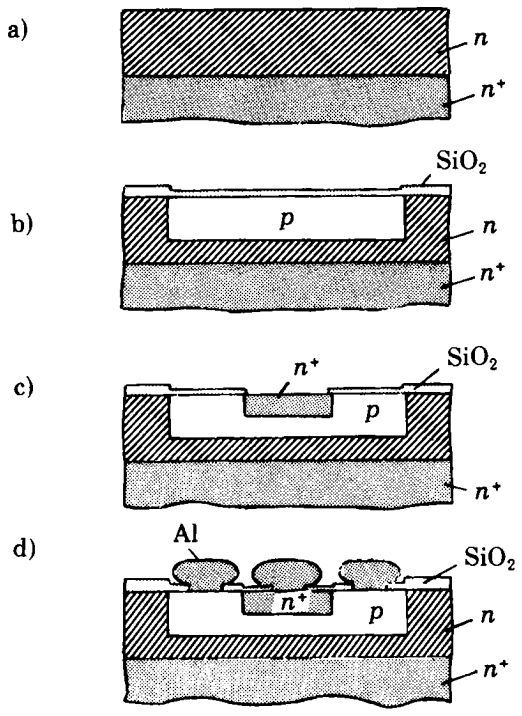


Fig. 10.14. Manufacture of an *n–p–n* transistor. (After Leturcq and Rey, "Physique des Composants Actifs à Semi-conducteurs," Dunod, 1978.)

diation) should allow the size to be reduced. The precision of positioning the masks in the successive operations will have to conform to this size reduction. A 0.3 μm scale technology is expected to be operative in the late nineties.

Another important parameter is the minimum time for an elementary operation. This is the time taken, for example, by the carriers to cross the channel of a MOSFET. If L is the length of the channel and v_d the drift speed

$$\tau = \frac{L}{v_d} = \frac{L}{\mu E} = \frac{L^2}{\mu V_{DS}}.$$

For $L = 10$ μm, $\mu = 0.1$ m$^2 \cdot$ V$^{-1} \cdot$ s^{-1}, and $V_{DS} = 2$ V, this time is of the order of 10^{-9} s. (Compare with (10.9).) The time taken by the whole system to complete a logical operation is not τ but proportional to τ. We see that the reduction in size corresponds not only to a greater integration density but also to increased speed. However, in strong electric fields of order 10^6 V/m the speed v_d saturates at around 10^5 m/s for electrons of

silicon and thus τ will in future vary as the size L and not as its square. There are however difficulties in increasing the number of components per chip, related to the fact that a single defect can prevent the functioning of the entire device. These defects can come from the substrate (scratches) but also from dust that may contaminate the surfaces at each stage of the process—the work must therefore be performed in dust-free environments ("clean rooms"). Also circuits that are too complex cannot be tested in all their states. It is now becoming necessary to design chips in which a part of the circuit is devoted to testing the subcircuits of the same chip.

10.5 Band Gap Engineering

The recent ability to manufacture semiconductors layer by atomic layer through molecular beam epitaxy allows one to produce structures that are sequences of heterojunctions. Depending on how successive layers are doped we obtain either a band profile with gaps (intrinsic or weakly doped semiconductors) or of the type shown in Fig. 9.16. This gives a new technique, "band gap engineering," so-called since the band profile can be modeled at will. It takes advantage of several effects: the possibility of confining and controlling the wave function culminates, for example, in the manufacture of a FET in which the conducting channel does not contain any dopant, increasing the mobility (cf. Sect. 5.4). The asymmetry of the electrons and holes in heterojunctions allows the design of circuits where only the majority carriers are active. The ability to stack circuits of several nanometers thickness leads to a miniaturization no longer limited by lithography.

Semiconducting heterostructures that make use of quantum effects have been used commercially for several years: for example, adjusting the energy of the transition between the conduction and the valence ground states through the choice of the well width (cf. Appendix 3.2), together with the large value of the absorption coefficient (cf. Appendix 6.3), allow the construction of efficient laser diodes, in which the rate of conversion of electrical energy into light is large. These systems operate in the red or near infrared range and are currently used in fiber optics communication.

Many other applications of band gap engineering are in progress. As an example we describe the principle of an infrared detector in the 5–10 μm range based on a dissymetrical quantum well structure (Fig. 10.15).

In such a system, the average position of an electron in the quantized level E^2 is shifted in space with respect to that in the quantized level E^1. This charge displacement is associated with a very large transition matrix element between E^1 and E^2. This ensures a very strong coupling of the well with the electromagnetic field. In a sense the well behaves as a giant molecule with a size of the order of nanometers.

If the well is n doped, in the dark electrons occupy the level E^1. By illumination with infrared photons, these electrons are promoted to E^2 (bound-

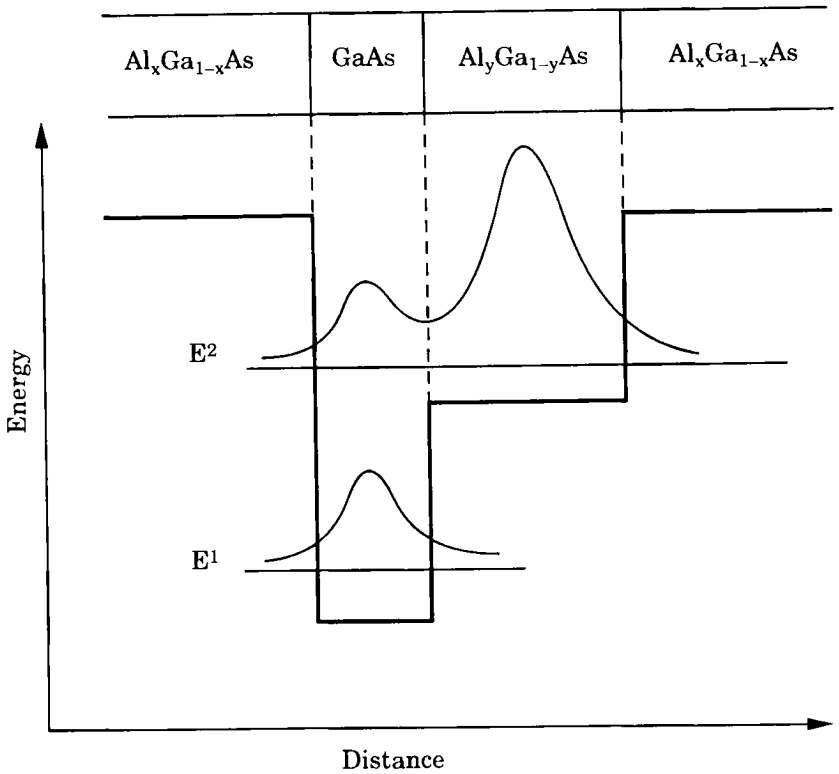


Fig. 10.15. Conduction band profile of a dissymmetrical quantum well structure.

to-bound transition) or to the delocalized levels above the Al_x Ga_{1-x} As barrier (bound-to-free transition). If the structure is submitted to a transverse electric field, the energy profile is as in Fig. 10.16.

When an infrared photon induces a bound-to-free (a) transition, like in the figure, i.e., if its energy is larger than the distance between E_1 and the barrier, the photoexcited electron is swept by the electric field, and produces a photocurrent. In the case of a photon exciting the E^1–E^2 bound-to-bound (b) transition, the excited electron can tunnel from the level E^2 through the barrier in presence of the electric field (Fig. 10.16). The barrier being triangular, it is easier to cross by tunneling when the field is strong. In both cases the absorption of an infrared photon leads to a photocurrent.

Moreover in presence of the electric field, the positions of E^1 and E^2 are slightly modified by the linear Stark effect (Fig. 10.17). Then the energy of the photons which are absorbed by the structure can be slightly tuned through this electric field.

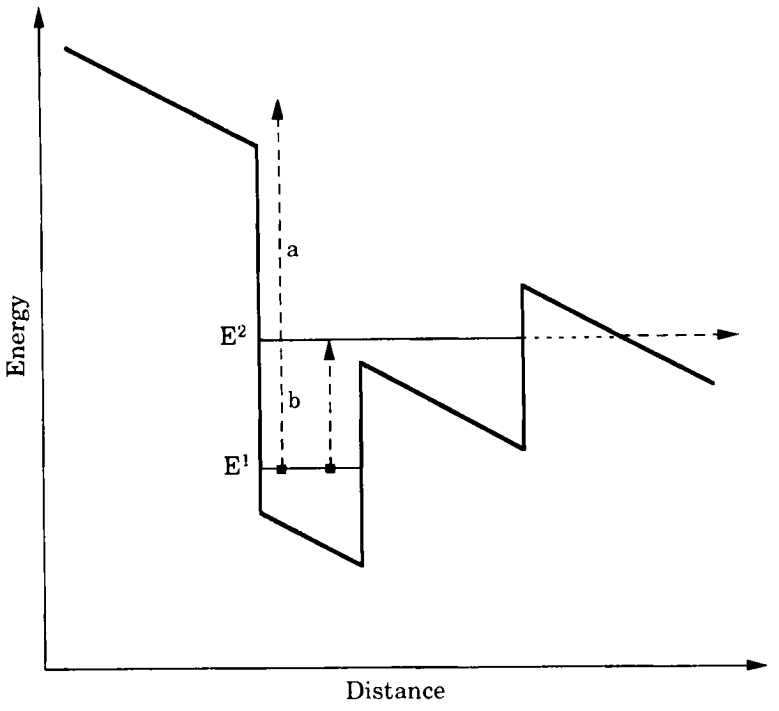


Fig. 10.16. Modification of the band profile of Fig. 10.15 by application of a transverse electric field.

10.6 Physical Limits in Digital Electronics

Since 1958, the beginning of the integrated circuit era, the size of the smallest circuits has decreased by an average of 13% per year. At this rate it will be around 0.1 μm by the year 2000. Assuming that technological progress allows the manufacture of submicron circuits, it is essential to understand the fundamental limits on miniaturization and integration.

We summarize below some of the physical limitations to the electronic treatment of information by solid-state electronics. We shall see that one of the main limitations arises from the need to treat numerical data digitally.

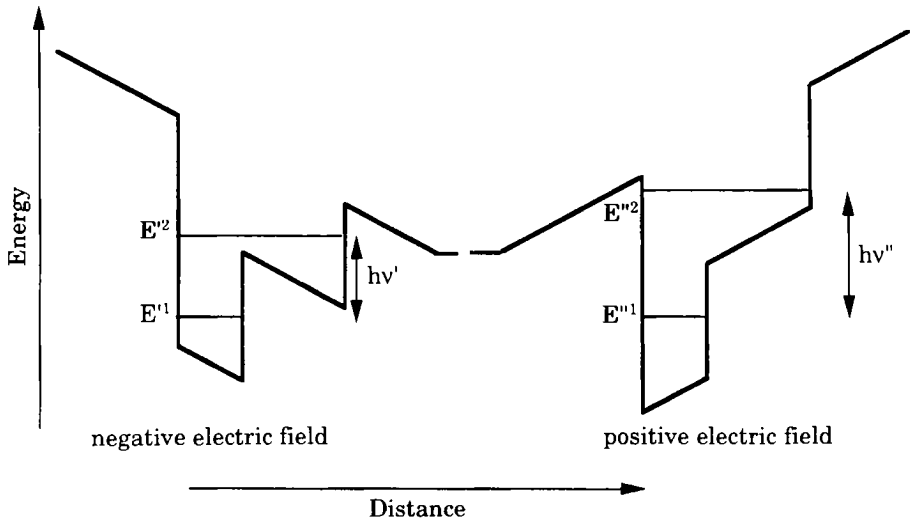


Fig. 10.17. Linear Stark effect on the band profile of Fig. 10.15. The modifications of the energy level positions and of the optical transitions depend on the sign and amplitude of the electric field.

The Need for a Digital Treatment

In electronic information processing or calculation, all data are manipulated in a large number of successive operations. A small error introduced into each operation can cause the complete loss of the original information. A random error of 1% at each step becomes an error of 100% after 10^4 operations. For this reason one uses digital, particularly binary, representation of the input data, which allows one to prevent the buildup of random errors by "restandardizing" each "digit" at each stage. Consider the simplest binary calculation, which associates an output $S = E + 0$ with each input E. If $E = 1$ V it may be that because of an imperfection or perturbation we have $S = 1.01$ V (a 1% error). At the end of this elementary process we recognize that the actual output value (1.01) is close to 1 and thus corresponds *with certainty* to the value 1 and not the value 0. We can thus assign it the value 1. Digital representation thus allows one to obtain arbitrarily high precision by increasing the number of binary digits of the representation.

Minimum Energy of a Logical Operation

We may also ask what is the limiting energy required for a logical operation (Landauer, 1961). The simplest physical system representing a binary state has the potential energy shown in Fig. 10.18.

This system is bistable in the sense that it can exist in either the state 0 or the state 1. Consider the operation of putting the system in state 0 whatever the initial state. If we wish to be certain to find the system in

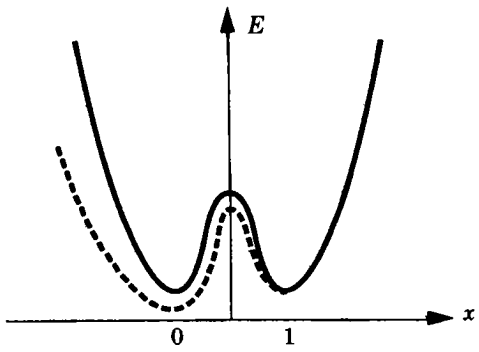


Fig. 10.18. Potential energy of a binary system. Dashed, the energy obtained by application of an external force.

state 0 we have to apply an external force so as to lower the energy of the left-hand potential well (dashed curve in Fig. 10.18). We thus increase the probability of finding the system on the left. For this probability to get very close to 1 we must lower the potential by several times kT. In other words, the system must dissipate an energy several times kT, i.e., $4 \cdot 10^{-21}$ J, in each operation of resetting to zero. In fact, for the best currently available complete circuits, the product of the power P times the duration τ of an operation is of the order of 10^{-10} J. It requires, for example, 10^{-1} W to perform an operation in 10^{-9} s, and we are thus very far above the fundamental limit described here.

The Quantum Limit

If we wish to localize a physical system between two distinct states after a delay τ, these states must differ by an energy $\Delta E \geq \hbar \tau^{-1}$. The dissipated power is thus $\hbar \tau^{-2}$. Here again this limit is remote from practical performances. For $\Delta E = 10^{-10}$ J, $\Delta t \sim 10^{-23}$ s while the best circuits presently work at $\Delta t \sim 10^{-9}$ s. It is therefore not the fundamental limits which determine the present performance of computers. We should nevertheless gather from this discussion that the order of magnitude of the energy necessary to run a logical system must be several times kT.

The "Natural" Voltage Scale for Semiconductor Electronics

The operation of restandardizing requires by definition the use of a non-linear device. Now, the basic non-linear relation available is Shockley's law, which gives the characteristic $J(V)$ of a p–n junction (Chap. 8). For an electronic device built using this law to be strongly non-linear we have to

apply voltages of a few times $kT/e = 25 \times 10^{-3}$ V. This is why the "natural" voltage scale used in digital electronics is of the order of a volt.

Transmission of Data

To move from one stage to the next the voltage must be carried by conductors with an impedance Z. The power required is thus at least of the order of $(kT/e)^2 Z^{-1}$. It is very difficult to produce lines where Z is very different from the impedance Z_0 of vacuum, which is 300 Ω. We find in this case a power $\sim 2 \times 10^{-6}$ W. This power corresponds to an energy $10^6 kT$ per nanosecond, showing why the energy dissipated per operation is much higher than the thermodynamic limit kT.

In practice the voltages used are 10 or 100 times larger than kT/e and this is the main source of dissipation in logic circuits. Given this power, one has to efficiently cool the most rapid circuits to prevent their deterioration. Heat removal at present sets the practical limit for the most powerful machines.

Miniaturization Limit

There are other limits which are caused by the nature of the manufacturing process itself (photolithography, chemical etching, and so forth). We mention the striking limit coming from the inherent fluctuation of small objects. A cube of semiconductor of size 0.2 μm contains on average only 80 impurities if the semiconductor is doped at a concentration of 10^{22} m^{-3}. The role of statistical concentration fluctuations becomes very important when we envisage several million circuits per chip as we do now, near the end of the 20th century.

Appendix 10.1

Problems on the n–p–n Transistor

The aim of these problems is to give a simple quantitative description of the operation of the bipolar transistor, which we have already discussed qualitatively. We consider the device shown schematically in Fig. 10.19, consisting of a semiconducting single crystal with three successive regions of different dopings. These regions are conventionally called emitter (here of type n^+, strongly doped), base (weakly p doped), and collector (normal n doping), and we shall denote physical quantities in these regions by indices E, B, and C. For example, for the emitter n_E and p_E are the equilibrium electron and hole concentrations; D_E, τ_E, and L_E the diffusion coefficient, recombination time, and diffusion length of the minority carriers (here the holes since the emitter is n doped).

The crystal has the form of a rectangular parallelepiped with its long side along Ox, and cross section S (Fig. 10.19): the emitter-base and base-collector junctions are in planes perpendicular to Ox, so the currents have non-zero components only along Ox. We assume further the lengths of the emitter and collector sections are infinite. The thickness of the base is Δ.

Problems

1. Show the band energy profile as a function of x when the system is in equilibrium. Note the appearance of two space-charge regions, bounding three neutral regions.

In the transistor's normal regime there is a forward bias on the base-emitter junction (B–E), and the base-collector junction is strongly biased in the reverse sense: if we adopt the convention $V_B = 0$, the applied voltages shown in Fig. 10.19 are $V_E < 0, |V_E| >> (kT/e) = 25$ mV at the emitter and $V_C >> |V_E|$ at the collector. We assume that the voltage drops in the bulk of the semiconductor are negligible. Show the band scheme as a function of x for these conditions.

2. Write down the evolution equations for the minority carrier density in the three neutral regions taking account of diffusion and recombination, and simplify them in the stationary case.

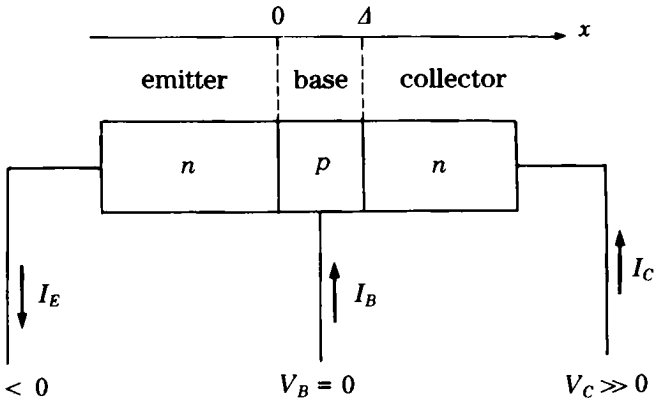


Fig. 10.19. Schematic representation of the structure of an n–p–n transistor with the standard sign conventions for currents.

3. We will assume as for the biased p–n junction that the widths of the space-charge regions are negligible compared with the diffusion lengths and that in the space-charge regions the electrons and holes are in thermal equilibrium (cf. Eq. (8.35)). Deduce the ratio $n(0_+)/n_B$, where $n(0_+)$ is the electron concentration in the base at the edge of the space-charge region. Similarly calculate $p\ (0_-)/p_E, n(\Delta_-)/n_B$, and $p(\Delta_+)/p_C$.

4. Find the hole concentration in the emitter by solving the differential equation corresponding to **2** with the boundary conditions of **3**. Deduce the current density $J_h(0_-)$ of injected holes at the base-emitter junction. Assume that the current densities are conserved across the junctions, and that the points $0_+, 0_-$ have the same abscissa, 0, in the algebraic expressions obtained.

5. The same question for the holes in the collector. Deduce $J_h(\Delta_+)$, the hole current density at the base-collector junction.

6. The same question for the electrons in the base. Because of boundary conditions at $x \simeq 0$ and at $x \simeq \Delta$ it is convenient to introduce hyperbolic functions $\sinh(x/L_B)$ and $\sinh[(x-\Delta)/L_B]$ rather than exponentials. Deduce the current densities $J_e(0_+)$ and $J_e(\Delta_-)$ at the two junctions. For ease of writing set $eV_E/kT = v_E, eV_C/kT = v_C$.

7. Deduce from questions **4** and **6** the total intensities I_E crossing the emitter and I_C crossing the collector. Take the positive senses as in Figs. 10.19 and 10.20.

8. Using the assumptions $V_E < 0, V_C >> 0$ show that if we also have $\Delta << L_B$, then I_C and I_E are of the same order, and thus $I_B << I_C$. Interpret this physically.

9. In the limit $V_C >> 0$ the currents I_E and I_C depend only on V_E; calculate the differential current gain $\beta = dI_C/dI_B$ of the transistor. How should the transistor be made so that β is as large as possible? Calculate

β for $n_E = 10^{25}$ m^{-3}, $p_B = 3 \cdot 10^{22}$ m^{-3} $n_C = 5 \cdot 10^{21}$ m^{-3}, $\Delta = 3$ μm, $D_E = 10^{-4}$ m^2/s, $D_B = 3.5 \times 10^{-3}$ m^2/s, $L_E = 1$ μm, $L_B = 100$ μm.

10. Draw the characteristics of the transistor, i.e., the set of curves $I_C = f(V_C - V_E)$ for different values of V_E.

11. Why are users of transistors advised not to interchange the collector and emitter leads?

Remark: As seen in this problem, the diffusion of the carriers across the base (here the electrons) determines the operation of the transistor. The geometry of the emitter or collector has no influence. This justifies the choice of simple parallelepiped geometry in this problem.

Solutions

1. In equilibrium the Fermi level is constant throughout the system. We know that each p–n or n–p junction has a space-charge region whose width is determined by the doping as represented in Fig. 10.20(a). Figure 10.20(b) shows the band profile of a polarized transistor.

2. The carrier conservation equation for holes in the emitter is

$$\frac{\partial p}{\partial t} = \frac{1}{e}\frac{\partial J_h}{\partial x} - \frac{p - p_E}{\tau_E}, \tag{10.11}$$

where J_h is the hole electric current.

For the minority carriers, the current is essentially the diffusion term, so for holes in the emitter we have

$$J_h \simeq -e\, D_E \frac{\partial p}{\partial x}. \tag{10.12}$$

Combining these equations we get

$$\frac{\partial p}{\partial t} = D_E \frac{\partial^2 p}{\partial x^2} - \frac{p - p_E}{\tau_E} = 0 \text{ in the steady state.} \tag{10.13}$$

Similarly, in the steady state we write for the base electrons:

$$D_B \frac{\partial^2 n}{\partial x^2} - \frac{n - n_B}{\tau_B} = 0 \tag{10.14}$$

and for the collector holes:

$$D_C \frac{\partial^2 p}{\partial x^2} - \frac{p - p_C}{\tau_C} = 0. \tag{10.15}$$

3. The equations above involve the diffusion lengths

$$L_E = \sqrt{D_E\, \tau_E},\ L_B = \sqrt{D_B\, \tau_B},\quad \text{and } L_C = \sqrt{D_C\, \tau_C}. \tag{10.16}$$

At the edge of the space charge of the emitter-base junction minority carriers are injected into the base since the junction is forward biased. For

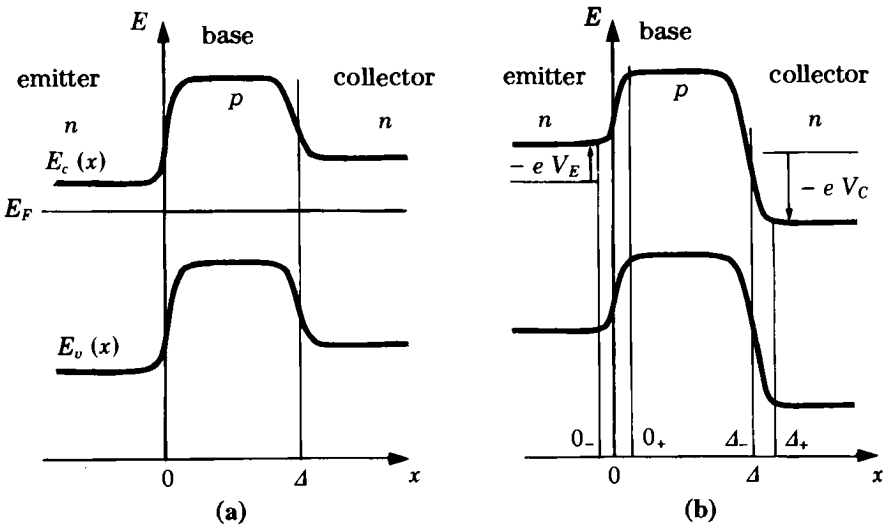


Fig. 10.20. Band profile in an n–p–n transistor: (a) zero bias; (b) profile when biased ($V_E < 0, V_C > 0$). The arrows show the energy shifts, giving the algebraic values on the energy axis.

example, electrons are injected from the emitter to the base. The assumption of thermal equilibrium for the carriers in the space-charge region allows us to write

$$n\,(0_+) = n_B \exp(-e\,V_E/kT), \tag{10.17}$$

$$p\,(0_-) = p_E \exp(-e\,V_E/kT). \tag{10.18}$$

As $V_E < kT/e$, we have

$$n\,(0_+) >> n_B,\ p\,(0_-) >> p_E. \tag{10.19}$$

There are thus more minority carriers than in equilibrium.

The base-collector junction is reverse biased, resulting in extraction of minority carriers, so that

$$n\,(\Delta_-) = n_B \exp(-e\,V_C/kT), \tag{10.20}$$

$$p\,(\Delta_+) = p_C \exp(-e\,V_C/kT). \tag{10.21}$$

As $V_C >> 0$ we have

$$n\,(\Delta_-) << n_B,\ p\,(\Delta_+) << p_C. \tag{10.22}$$

Remark: We can use the language of quasi-Fermi levels introduced in Eqs. (8.48) and (8.49) for the polarized junction. As we are in the low injection regime, the quasi-Fermi levels of the majority carriers in the neutral regions are very close to those in equilibrium. The widths of the space-charge

regions are small compared with the diffusion lengths, so the quasi-Fermi levels are little changed in these regions. The position of the quasi-Fermi levels E_{Fe}, E_{Fh} for electrons and holes is shown in Fig. 10.21 for the band profile of Fig. 10.20.

4. The hole concentration in the emitter is

$$p = p_E + \lambda_1 \exp\left(\frac{x}{L_E}\right) + \lambda_2 \exp\left(-\frac{x}{L_E}\right), \tag{10.23}$$

where λ_1, λ_2 are constants determined by the boundary conditions

$$\begin{aligned} &\text{for } x \to -\infty \qquad p \to p_E \text{ so that } \lambda_2 = 0, \\ &\text{for } x \to 0_- \qquad p = p_E \exp(-e\,V_E/kT). \end{aligned} \tag{10.24}$$

Hence, for $x < 0$,

$$p = p_E \left[\exp\left(-\frac{e\,V_E}{kT}\right) - 1\right] \exp\left(\frac{x}{L_E}\right) + p_E. \tag{10.25}$$

The corresponding current is

$$J_h\,(x) = -e\,D_E \frac{\partial p}{\partial x}, \tag{10.26}$$

$$J_h\,(x) = -e \frac{D_E\,p_E}{L_E} \left[\exp\left(-\frac{e\,V_E}{kT}\right) - 1\right] \exp\left(\frac{x}{L_E}\right) \tag{10.27}$$

The current of holes injected at $x = 0_-$ from the base is

$$J_h\,(0_-) = -\frac{e\,D_E\,p_E}{L_E} \left[\exp\left(-\frac{e\,V_E}{kT}\right) - 1\right]. \tag{10.28}$$

5. Following the method of **4**, we get

$$J_h\,(x = \Delta_+) = \frac{e\,D_C\,p_C}{L_C} \left[\exp\left(-\frac{e\,V_C}{kT}\right) - 1\right]. \tag{10.29}$$

6. The electron concentration in the base is found from

$$n - n_B = \lambda_1'' \sinh \frac{x}{L_B} + \lambda_2'' \sinh \frac{x - \Delta}{L_B}, \tag{10.30}$$

$$\text{at } x = 0_+ \qquad n_B \exp\left(-\frac{e\,V_E}{kT}\right) = n_B - \lambda_2'' \sinh\left(\frac{\Delta}{L_B}\right), \tag{10.31}$$

$$\text{at } x = \Delta_- \qquad n_B \exp\left(-\frac{e\,V_C}{kT}\right) = n_B + \lambda_1'' \sinh\left(\frac{\Delta}{L_B}\right). \tag{10.32}$$

For ease of writing we set

$$e\,V_E/kT = v_E, \quad e\,V_C/kT = v_C. \tag{10.33}$$

We deduce

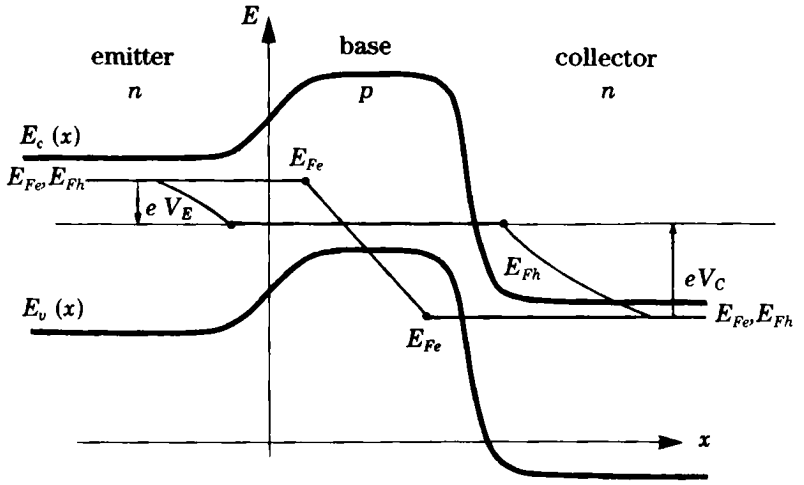


Fig. 10.21. Quasi-Fermi levels E_{Fe}, E_{Fh} in a polarized n–p–n transistor ($V_E < 0$, $V_C > 0$).

$$n - n_B = n_B \left\{ \sinh \frac{\Delta}{L_B} + \left[\sinh \frac{x}{L_B} \right] [\exp(-v_C) - 1] - \left[\sinh \frac{x - \Delta}{L_B} \right] [\exp(-v_E) - 1] \right\} / \sinh \frac{\Delta}{L_B}, \tag{10.34}$$

$$J_e(x) = e D_B \frac{\partial n}{\partial x} = \frac{e D_B n_B}{L_B \sinh(\Delta / L_B)} \left\{ \left[\cosh \frac{x}{L_B} \right] [\exp(-v_C) - 1] - \left[\cosh \frac{x - \Delta}{L_B} \right] [\exp(-v_E) - 1] \right\}, \tag{10.35}$$

so that

$$J_e(0_+) = \frac{eD_B\, n_B}{L_B} \times \left\{ \frac{1}{\sinh(\Delta/L_B)} [\exp(-v_C) - 1] - \left[\coth \frac{\Delta}{L_B} \right] [\exp(-v_E) - 1] \right\}, \tag{10.36}$$

$$J_e(\Delta_-) = \frac{eD_B\, n_B}{L_B} \times \left\{ \left[\coth \frac{\Delta}{L_B} \right] [\exp(-v_C) - 1] - \frac{1}{\sinh(\Delta/L_B)} [\exp(-v_E) - 1] \right\}. \tag{10.37}$$

7. In the space-charge region of the emitter-base junction the total current is conserved. As in the text, we use the expression for the electron and hole currents in the region where the current is purely diffusive. The currents I_E, I_C flow towards $x < 0$; also the current is the flux of the current density across the section S, so that

$$I_E = -S\, [J_e\, (0_+) + J_h\, (0_-)], \tag{10.38}$$

$$I_C = -S\, [J_e\, (\Delta_-) + J_h\, (\Delta_+)], \tag{10.39}$$

$$\frac{I_E}{eS} = \frac{D_B\, n_B}{L_B} \left\{ -\frac{1}{\sinh\, (\Delta/L_B)} [\exp(-v_C) - 1] + [\coth \frac{\Delta}{L_B} (-v_E) - 1] \right\} + \frac{D_E\, p_E}{L_E} [\exp(-v_E) - 1], \tag{10.40}$$

$$\frac{I_C}{eS} = \frac{D_B\, n_B}{L_B} \left\{ -\coth \frac{\Delta}{L_B} [\exp(-v_C) - 1] + \frac{1}{\sinh\, (\Delta/L_B)} [\exp(-v_E) - 1] \right\} - \frac{D_C\, p_C}{L_C} [\exp(-v_C) - 1]. \tag{10.41}$$

8. The assumptions about V_E, V_C imply that $\exp(-eV_E/kT) >> 1$ and $\exp(-eV_C/kT) << 1$. Further, if the base is narrow enough that $\Delta/L_B << 1$, we can write $\sinh(\Delta/L_B) \simeq \Delta/L_B$, $\coth(\Delta/L_B) \simeq L_B/\Delta >> 1$ and

$$I_E \simeq eS \left(\frac{D_B\, n_B}{\Delta} + \frac{D_E\, p_E}{L_E} \right) \exp(-v_E), \tag{10.42}$$

$$I_C \simeq \frac{eS\, D_B\, n_B}{\Delta} \exp(-v_E). \tag{10.43}$$

Then the term in $(1/\Delta)$ dominates the terms in $1/L_E$ or $1/L_C$, and I_E, I_C effectively reduce to their first term, and are thus of the same order. As $I_E = I_B + I_C$, this implies $I_B << I_E$.

Physically, if the base is narrower than the diffusion length, and if it is rather weakly doped, so that $(p_E/n_B) \cdot (\Delta/L_E) << 1$, almost all the electrons injected from the emitter towards the base manage to cross it without recombining, and reach the collector.

9. The differential current gain β is defined by

$$\beta = \frac{d\,I_C}{d\,I_B} = \frac{d\,I_C/d\,V_E}{d\,(I_E - I_C)/d\,V_E}. \tag{10.44}$$

Taking the exact expressions (10.40) and (10.41) for I_E and I_C we get

$$\begin{aligned} I_E - I_C = eS\left[D_B \frac{n_B}{L_B}\frac{\cosh\,(\Delta/L_B) - 1}{\sinh\,(\Delta/L_B)} + \frac{D_E\,p_E}{L_E}\right][\exp(-v_E) - 1] + \\ eS\left[\frac{D_B\,n_B}{L_B}\frac{\cosh\,(\Delta/L_B) - 1}{\sinh\,(\Delta/L_B)} + \frac{D_C\,p_C}{L_C}\right] \times \\ [\exp(-v_C) - 1]. \end{aligned} \tag{10.45}$$

For $\Delta << L_B$,

$$d\,I_C/d\,V_E \simeq \frac{-e^2\,S\,D_B\,n_B}{kT\,\Delta}\exp(-v_E), \tag{10.46}$$

$$d\,(I_E - I_C)/d\,V_E \simeq \frac{-e^2 S}{kT}\left(\frac{D_B\,n_B}{L_B}\frac{\Delta}{2\,L_B} + \frac{D_E\,p_E}{L_E}\right)\exp(-v_E), \tag{10.47}$$

$$\beta \simeq \frac{1}{(\Delta/L_B)\,[(\Delta/2)\;L_B + (p_E/n_B)\,(D_E/D_B)\,(L_B/L_E)]}. \tag{10.48}$$

To increase β, one adjusts the doping to have $p_E << n_B$, or equivalently $n_E >> p_B$ (since $n_E \cdot p_E = n_B \cdot p_B = n_i^2$). The emitter will be strongly doped, the base weakly doped. An indirect consequence of these different dopings is an increase in the lifetime in the base, and hence L_B.

For the values given in the question β is about 1500.

10. The characteristics (Fig. 10.22) are horizontal straight lines ("saturated" transistor) except when $V_C - V_E$ is small. I_C then decreases abruptly. For $V_E = 0$ and any V_C, the current I_C is also very small ("blocked" transistor).

11. The symmetry of the n–p–n transistor is only superficial. The emitter is strongly doped and the collector is not, so reversing these connections leads to poor performance (for example, a lower β). Moreover, because of the strong doping of the emitter, the base-emitter junction would not in general stand large voltages and would "break down."

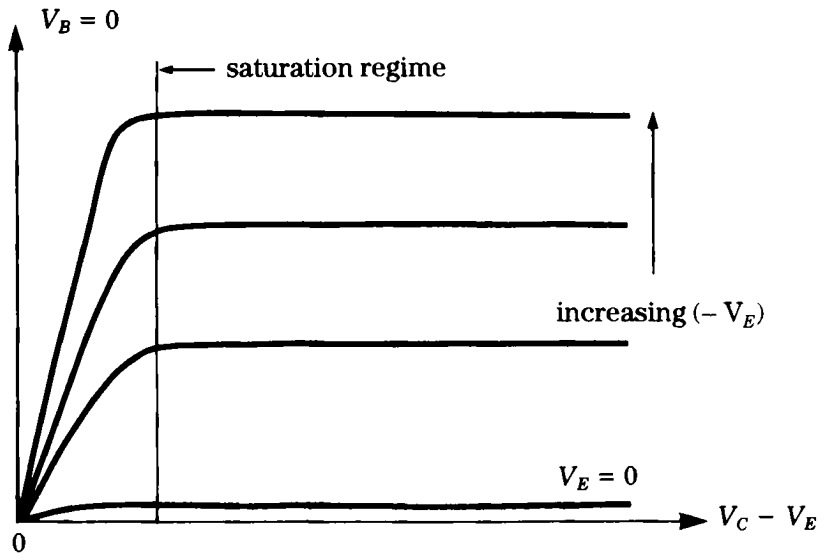


Fig. 10.22. Characteristics $I_C = f(V_C - V_E)$ with the base at zero potential, for different values of the emitter potential.

Appendix 10.2

Problems on the Junction Field-Effect Transistor (JFET)

In 1952 W. Shockley conceived an active device consisting of a slice of semiconductor between two reverse-biased p–n junctions. The principle is to modify the size of the conducting region called the channel by varying the reverse bias of the junctions. The device is shown in Fig. 10.23. We shall study the simplified geometry shown in Fig. 10.24, concentrating on the active part between the junctions.

The p^+ regions (gates) are negatively biased with respect to the contact S, called the source, and taken as the origin of the potentials. The contact D is called the drain. We apply to the drain a voltage positive with respect to the source. The thickness of the n region is $2a$, its length is L, and its width b. The set of characteristics measured for such a device (curves of drain current I_D as a function of the drain voltage V_D for various gate voltages V_G) has the form shown in Fig. 10.25. The aim of these problems is to understand this figure. The various notations are introduced below.

Problems

1. We are first of all interested in the case of infinitesimal drain potentials. In this limit the channel can be regarded as an equipotential, and because of the inverse polarization applied to the gates a depleted zone appears on each side, with uniform width d. Express d as a function of the internal potential ϕ of the junction, the applied gate voltage V_G, the dielectric constant $\epsilon_0\epsilon_r$ of the semiconductor, and the donor concentration N_d. (Consider the case where $N_d << N_a$, the acceptor concentration in the p^+ region, and neglect N_d/N_a compared with 1.)

2. If the electron mobility is μ_e, give the channel conductance G as a function of the voltage applied to the gate. Express the result as a function of $G_0 = 2e\mu_e N_d ab/L$, the channel conductance in the absence of any space charge.

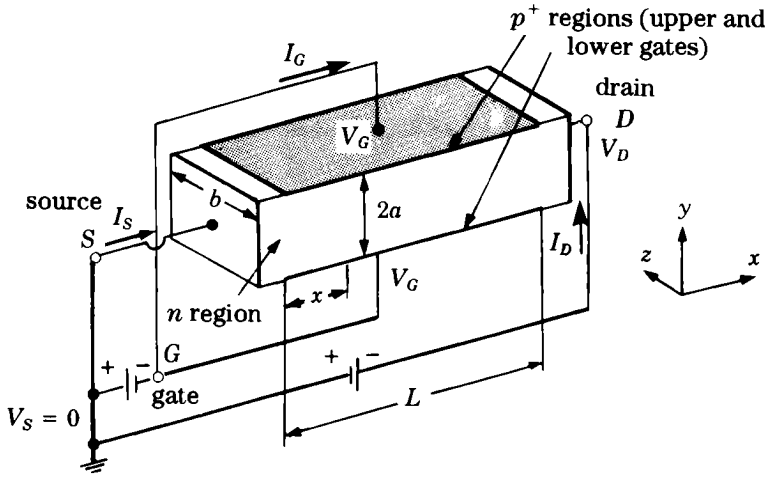


Fig. 10.23. Perspective view of the active part of a junction field-effect transistor. (After Millman and Halkias, "Integrated Electronics: Analog and Digital Circuits and Systems," McGraw–Hill, 1972.)

3. Give the minimum thickness $2a_m$ of the n-doped slice required for the conductance to be non-zero in the absence of a potential. Give this explicitly in the case $N_d = 10^{21}$ m^{-3}; $\epsilon_0\epsilon_r = 10^{-10}$ F $\cdot$ m^{-1}: take $\phi = 0.8$ V.

4. In the same regime where V_D is very small, consider the case where there is non-zero conduction, i.e., $a > a_m$. The device behaves as a variable resistance controlled by the gate. Show that the conductance vanishes for a particular value V_P of the gate voltage, which we call the pinch voltage, and give the expression for it. Find this voltage for $a = 2.65 \cdot 10^{-6}$ m.

5. The gate current does not appear in the results. Why?

6. In the same approximation (V_D small), write the equations of the characteristics $I_D = f(V_D)$ as functions of the parameters $G_0, (\phi - V_G)$, and $\psi = (\phi - V_P)$. Compare briefly with Fig. 10.25.

7. Now discard the assumption that V_D is infinitesimal. The larger V_D is, the less can the channel be regarded as equipotential. For a positive drain voltage the junction is more strongly reverse biased near the drain than near the source, and the width $d(x)$ of the depopulated zone on each side varies with x as shown in Fig. 10.26.

We assume $L >> 2a$. Then the current lines are effectively parallel to Ox and $J_y << J_x$, implying $\mathcal{E}_y << \mathcal{E}_x$ (Fig. 10.27). The equipotentials in the conducting channel are to a first approximation planes orthogonal to Ox and $\mathcal{E}_x$ is then independent of y.

We count the current positive when it flows from the drain to the source. Using the functions $d(x), V(x)$, the width and voltage in the depleted region, find the current at x in the useful section.

8. The current $I(x)$ is independent of x. Why?

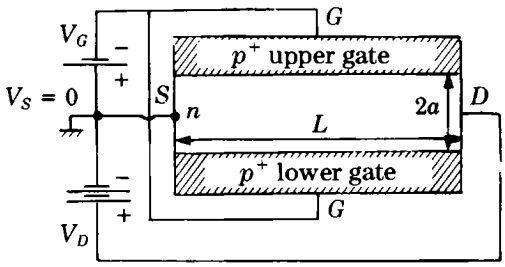


Fig. 10.24. Schematic view of the same transistor showing the doping of various regions. (After Leturcq and Rey, "Physique des Composants Actifs à Semi-conducteurs," Dunod, 1978.)

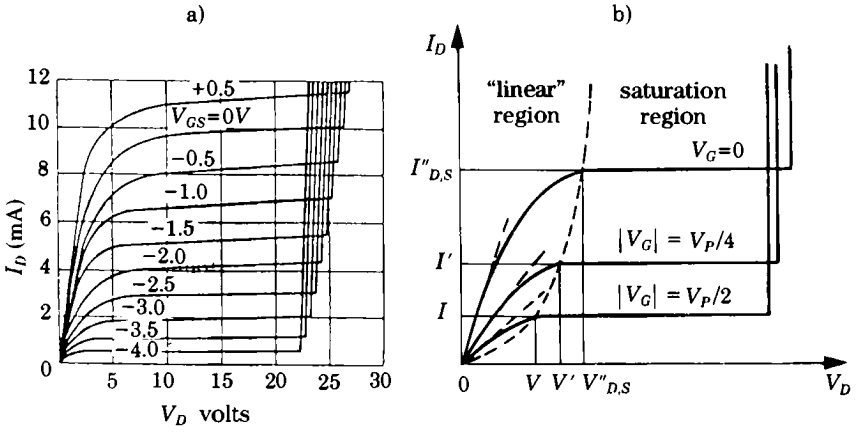


Fig. 10.25. Characteristics of the 2N3278 transistor (Fairchild Semiconductor Company). (a) Experimental curves; (b) simplified version of the same curves: I, I', I'', V, V', V'' denote saturation drain currents and voltages. The pinch voltage V_P is defined in problem 4.

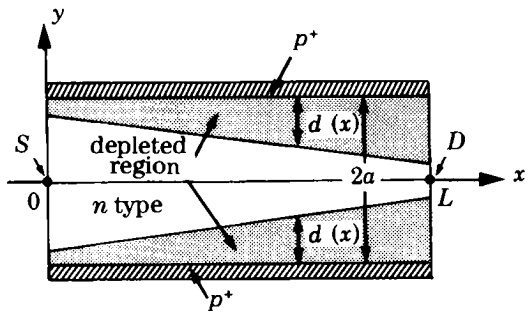


Fig. 10.26. Depleted regions and the conducting channel in a polarized JFET.

9. Deduce a differential equation relating $V(x)$ and $dV(x)/dx$, in which the drain current I_D is a parameter. For convenience use G_0 and ψ.

10. Integrate this differential equation between $x = 0$ and $x = L$ so as to obtain I_D as a function of G_0, V_D, ϕ, ψ, and V_G. Assume the channel is nowhere pinched.

11. What are the conditions on the applied voltages for this equation to hold? For what value of $V_D = V_{D,\text{Saturation}} = V_{D,S}$ is the channel pinched at $x = L$?

12. Under these conditions express the saturation current $I_{D,S}$ as a function of G_0, ψ and $V_{D,S}$. This relation bounds in Fig. 10.25(b) a region called "linear," to the left of the dashed curve. Show that this curve is a parabola with the equation

$$I_{D,S} = \frac{G_0}{4} \frac{(V_{D,S})^2}{\psi} \tag{10.49}$$

for $V_{D,S} << \psi$.

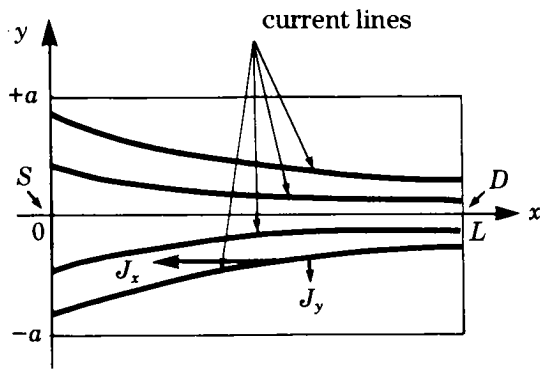


Fig. 10.27. Current lines in the channel of a polarized JFET. (After Leturcq and Rey, "Physique des Composants Actifs à Semi-conducteurs," Dunod, 1978.)

13. One might imagine that for $V_D > V_{D,S}$ the conducting region is closed off at the pinch and the current I_D drops. But experimentally one sees (Fig. 10.25(a)) that this does not happen. The current is stable above $V_{D,S}$ (in fact it rises very slightly). We investigate this regime here.

The problem is as follows: in the pinch region the electric field is very large. But in the presence of such a field the usual transport equations, derived for situations close to thermodynamic equilibrium, are no longer valid. In particular Ohm's law fails. Show first that in the above description the electric field grows in modulus and tends to infinity for $x = L$ when $V_D \to V_{D,S}$.

14. We thus have to consider the properties of a semiconductor in a very strong field and we may wonder if the electron velocity tends to infinity as we increase the field. It does not. In the presence of a field, the electrons gain between collisions an energy which is no longer negligible compared with the thermal energy kT, where T is the lattice temperature. If the electron collisions with the lattice were strictly elastic the mean energy of the electron gas would increase. But the collisions are not elastic: the electrons lose energy at each collision and an equilibrium is established. We call this a "hot" electron gas, which we describe by an effective temperature T_e which differs from that of the lattice (T), and by a distribution of electron energies E obeying

$$f = A \exp\left(-\frac{E}{kT_e}\right). \tag{10.50}$$

The effective temperature of the electrons is an increasing function of the electric field intensity. By substituting this distribution function in the expression for the mobility deduced from the Boltzmann equation, show that for collisions with crystal vibrations (cf. Sect. 5.4a) the mobility is proportional to $T^{-1}T_e^{-1/2}$, and is thus a decreasing function of the electric field.

15. In fact one observes that in silicon the mobility follows a law which can be modeled as

$$\mu = \frac{\mu_e}{1 + \mu_e \left|dV/dx\right| / v_l}. \tag{10.51}$$

The velocity v_l appearing in this equation is a limiting velocity, which is 10^5 m $\cdot$ s^{-1} in silicon. The behavior of the drift velocity as a function of electric field is shown in Fig. 10.28.

The description of the fields and currents in this case is more complex. In particular the approximation of neglecting the y dependence of the component $\mathcal{E}_x$ (and the $\mathcal{E}_y$ component) no longer holds throughout the device.

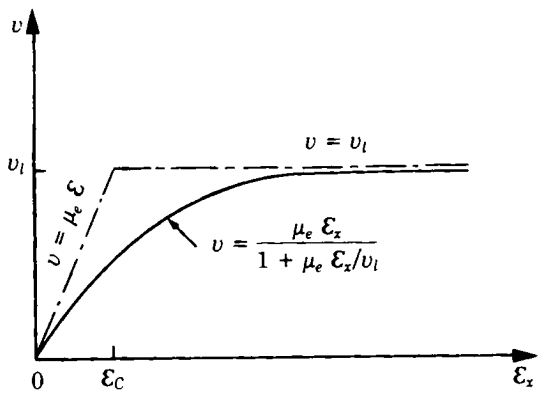


Fig. 10.28. Variation of the drift velocity with the electric field $\mathcal{E}_x$.

We retain the approximation however. Show from the equation obtained in 7 that for a given current the channel is no longer completely pinched in the strong electric field region, the width of the conducting region remaining approximately constant in this region. Give an order of magnitude of a "very strong" field.

16. To understand this saturation region we can schematically represent the problem by decomposing the crystal into two regions: one with a weak field and constant mobility, of length L_1, and a region with very strong field and constant velocity with length L_2, with $L = L_1 + L_2$. We can assume that $V(x = L_1) = V_{D,S}$ and thus that $V(x = L) - V(x = L_1) = V_D - V_{D,S}$. Give a maximum value for L_2. Evaluate this in the case $V_D - V_{D,S} = 10$ V.

17. Assume that $L = 10^{-4}$ m. Summarize the behavior of the characteristics in the saturation region.

18. What causes the abrupt rise of I_D for very large V_D?

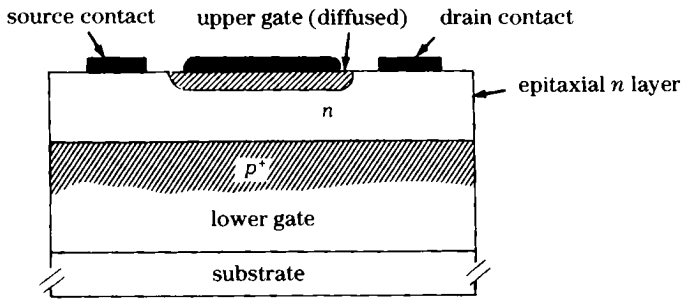


Fig. 10.29. Section of a JFET.

19. A real device, manufactured using planar techniques, is shown in section form in Fig. 10.29. What other effects, neglected here, can modify its behavior?

Solutions

1. From Eq. (8.60),

$$d = [2\ \epsilon_0 \epsilon_r\ (\phi - V_G)/eN_d]^{1/2}. \tag{10.52}$$

2. We have for a conducting parallelepiped of cross section $s = 2(a-d)b$:

$$G = \frac{1}{R} = \sigma \frac{s}{L} = N_d\ e\ \mu_0 \frac{2\ b}{L}(a - d) \tag{10.53}$$

$$= G_0 \left(1 - \frac{d}{a}\right) \tag{10.54}$$

$$= G_0 \left\{1 - \left[\frac{2\ \epsilon_r \epsilon_0\ (\phi - V_G)}{e\ N_d\ a^2}\right]^{1/2}\right\}. \tag{10.55}$$

3. For $d = a_m$ the conductance is zero, because the depleted region blocks all the conducting space. From Eq. (10.52), with $V_G = 0$:

$$a_m = \left(\frac{2\epsilon_0 \epsilon_r}{e} \frac{\phi}{N_d} \right)^{1/2} \tag{10.56}$$

$$a_m = \left(\frac{2 \cdot 10^{-10}}{1.6 \cdot 10^{-19}} \cdot \frac{0.8}{10^{21}} \right)^{1/2} = 1 \mu\text{m} \tag{10.57}$$

4. From Eq. (10.54), $G = 0$ when

$$\frac{2\epsilon_r \epsilon_0}{e\, N_d\, a^2} (\phi - V_P) = 1, \tag{10.58}$$

or

$$V_P = \phi - \frac{e\, N_d\, a^2}{2\epsilon_r \epsilon_0}, \tag{10.59}$$

$$V_P = -4.8 \text{ V}. \tag{10.60}$$

5. The gate current is that of a reverse biased junction and is negligible.

6. Using Eq. (10.56) in the form $\psi = eN_d a^2 / 2\epsilon_r \epsilon_0$, we get

$$I_D = G_0 \left[1 - \left(\frac{\phi - V_G}{\psi} \right)^{1/2} \right] V_D. \tag{10.61}$$

This gives a family of straight lines through the origin. The slopes increase when the magnitude of V_G rises from V_P.

7. The current at x is the flux of the current vector $J(x)$ in the depleted section, with modulus

$$I\,(x) = \int_{-a+d(x)}^{a-d(x)} b |J\,(x)| dy. \tag{10.62}$$

$J(x) = N_d e \mu_0 (dV/dx)$ is independent of y in the present approximation, so

$$I\,(x) = 2[a - d\,(x)] b\; e\; N_d\; \mu_e \frac{dV}{dx}. \tag{10.63}$$

8. If no current flows in the gate, all the electrons supplied by the source must reach the drain.

9. Equation (10.52), evaluated at the point x where the applied voltage at the junction is $V_G - V(x)$, can be written

$$d\,(x) = \left[\frac{2\epsilon_0 \epsilon_r [\phi - V_G + V\,(x)]}{eN_d} \right]^{1/2} \tag{10.64}$$

or using Eq. (10.59),

$$d\,(x) = a \left(\frac{\phi - V_G + V\,(x)}{\psi} \right)^{1/2}. \tag{10.65}$$

Substituting in Eq. (10.59),

$$I(x) = I_D = 2abeN_d\ \mu_e \left[1 - \left(\frac{\phi - V_G + V(x)}{\psi}\right)^{1/2}\right] \frac{dV}{dx} \tag{10.66}$$

$$I_D = G_0\ L \left[1 - \left(\frac{\phi - V_G + V\ (x)}{\psi}\right)^{1/2}\right] \frac{dV}{dx}. \tag{10.67}$$

10. From Eq. (10.67),

$$\int_0^L I_D\ dx = I_D\ L = G_0\ L \int_0^{V_D} \left[1 - \left(\frac{\phi - V_G + V}{\psi}\right)^{1/2}\right] dV, \tag{10.68}$$

$$I_D = G_0 \left[V_D - \frac{2}{3} \frac{(\phi - V_G + V_D)^{3/2} - (\phi - V_G)^{3/2}}{\psi^{1/2}}\right]. \tag{10.69}$$

11. For Eq. (10.69) to hold we must have reverse bias, the gate current must be negligible, and the channel must be open everywhere, or

$$V_G < 0 \text{ and } |V_G| < V_P,\ V_D - V_G > 0,\ V_D - V_G < -V_P. \tag{10.70}$$

The channel is pinched at $x = L$ when $V_{D,S} = V_G - V_P$.

12. The current $I_{D,S}$ is found by substituting $V_{D,S}$ into Eq. (10.69),

$$I_{D,S} = G_0\ \psi \left[\frac{V_{D,S}}{\psi} - \frac{2}{3} + \frac{2}{3}\left(\frac{\psi - V_{D,S}}{\psi}\right)^{3/2}\right]. \tag{10.71}$$

For $V_{D,S}/\psi << 1$,

$$I_{D,S} \sim G_0\psi \left[\frac{V_{D,S}}{\psi} - \frac{2}{3} + \frac{2}{3}\left(1 - \frac{3}{2}\frac{V_{D,S}}{\psi} + \frac{3}{8}\left(\frac{V_{D,S}}{\psi}\right)^2 + \cdots\right)\right]$$

$$\sim \frac{G_0}{4} \frac{(V_{D,S})^2}{\psi}. \tag{10.72}$$

13. From Eq. (10.67),

$$\left(\frac{dV}{dx}\right)_{x=L} = \frac{I_D}{G_0\ L} \frac{1}{1 - [(\phi - V_G + V_D)/\psi]^{1/2}} \tag{10.73}$$

is infinite for $V_D = V_{D,S} = V_G - V_P$.

We can also say that the area crossed by the constant current vanishes. This requires the electric field to be infinite.

14. We showed in Chap. 5 that

$$\mu = e\frac{< \tau >}{m} \quad \text{with} \tag{10.74}$$

$$< \tau > = \frac{\int_0^\infty \tau(E)\ E^{3/2}\ f(E)\ dE}{\int_0^\infty E^{3/2}\ f(E)\ dE}. \tag{10.75}$$

Replacing $f(E)$ by $A\exp(-E/kT_e)$ and τ by $E^{-1/2}T^{-1}$ for collisions with crystal vibrations for a lattice temperature T (see Sect. 5.4a), we get

$$<\tau> \sim T^{-1}\, T_e^{-1/2}. \tag{10.76}$$

15. Replacing μ in Eq. (10.63) we have

$$I_D = 2[a - d(x)]b\; e\; N_d\; \mu_e \frac{dV/dx}{1 + \mu_e/v_l\,|dV/dx|}. \tag{10.77}$$

For $\dfrac{dV}{dx} >> \dfrac{v_l}{\mu_e}$,

$$I_D = 2[a - d(x)]be\; N_d\; v_l. \tag{10.78}$$

For a very strong field, d becomes independent of x since I_D is conserved. As this constant value is less than a ($I_D > 0$) the channel is no longer pinched. The critical value of the field is $\mathcal{E}_C = v_l/\mu_e$. For $v_l \sim 10^5\ \mathrm{m \cdot s^{-1}}$, $\mu_e \sim 0.1\ \mathrm{m^2 \cdot V^{-1} \cdot s^{-1}}$, a "strong field" will be greater than $\mathcal{E}_C$, which is $10^6\ \mathrm{V \cdot m^{-1}}$.

16. In the strong-field region of length L_2, $\dfrac{dV}{dx} > \dfrac{v_l}{\mu_e}$, and

$$V_D - V_{D,S} = \int_{L_1}^{L} \frac{dV}{dx}dx > \frac{v_l}{\mu_e}L_2, \tag{10.79}$$

$$L_2 < \frac{(V_D - V_{D,S})\,\mu_e}{v_l} \sim 10^{-5}\ \mathrm{m}. \tag{10.80}$$

17. In the saturation regime the conductance of the low-field region is given by Eq. (10.71), where we replace L in G_0 by L_1, and thus G_0 by $G_0L/(L - L_2)$. As L_2 is much smaller than L the current is almost constant (in reality it increases slightly), and the strong-field region does not contribute significantly to the total conductance.

18. This is the reverse breakdown of the gate–drain junction. A current therefore flows in the gate. We note that the more negative V_G is, the lower the breakdown voltage at the drain: the important quantity here is the difference between the gate and drain potentials, which must be kept below the breakdown voltage.

19. We have neglected the y dependence of the electric field and the edge effects, among others arising from the finite width of the device. The coefficients appearing in the calculation are slightly modified without changing the behavior. The main effect is the existence of non-zero resistances between the source contact and the active region and between the active region and the drain contact.

Appendix 10.3

Problems on MOS (Metal-Oxide-Semiconductor) Structure

We study here the MIS (metal–insulator–semiconductor) and MOS (metal–oxide–semiconductor) structures, which are the building blocks of MOSFET transistors, just as p–n junctions are for bipolar transistors. To this end, we consider the plane capacitor shown in Fig. 10.30(a), formed of a semiconducting substrate covered with an insulating oxide layer of thickness d and a metallic layer. The insulating layer prevents a current from flowing between the metal and the semiconductor, and thus allows control of the electric field in the semiconductor by the voltage applied to the metal electrode without there being injection of carriers, unlike the case of a p–n junction.

Figure 10.30(b) gives the energy diagram of the MOS structure when the two ends of the structure are not electrically connected. Conditions are assumed uniform in planes parallel to the interfaces, and thus depend only on the variable x orthogonal to the interfaces, as in the notation of Fig. 10.30.

The energy origin is arbitrary, but a convenient way of relating the semiconductor and metal energies in the present case is to refer the energies of the two media to the energy E_0 of the bottom of the oxide conduction band. The Fermi levels in the metal and semiconductor are E_{Fm}, E_{Fsc}. We denote by ϕ_m, ϕ_{sc} the quantities $E_0 - E_{Fm}, E_0 - E_{Fsc}$. In general the Fermi levels do not coincide since the metal and the semiconductor are isolated from each other and cannot exchange electrons. We set $\phi_m - \phi_{sc} = \Delta\phi$. The case $\Delta\phi > 0$ is shown schematically in Fig. 10.30(b). This figure also shows the bottom E_c of the conduction band, the top of the valence band E_v ($E_c - E_v = E_g$, the semiconductor band gap). We call the interface affinity χ the difference between the energy of the oxide conduction band and the energy of the conduction band of the semiconductor at the oxide–semiconductor interface:

$$\chi = (E_0 - E_c)_{x=0}.$$

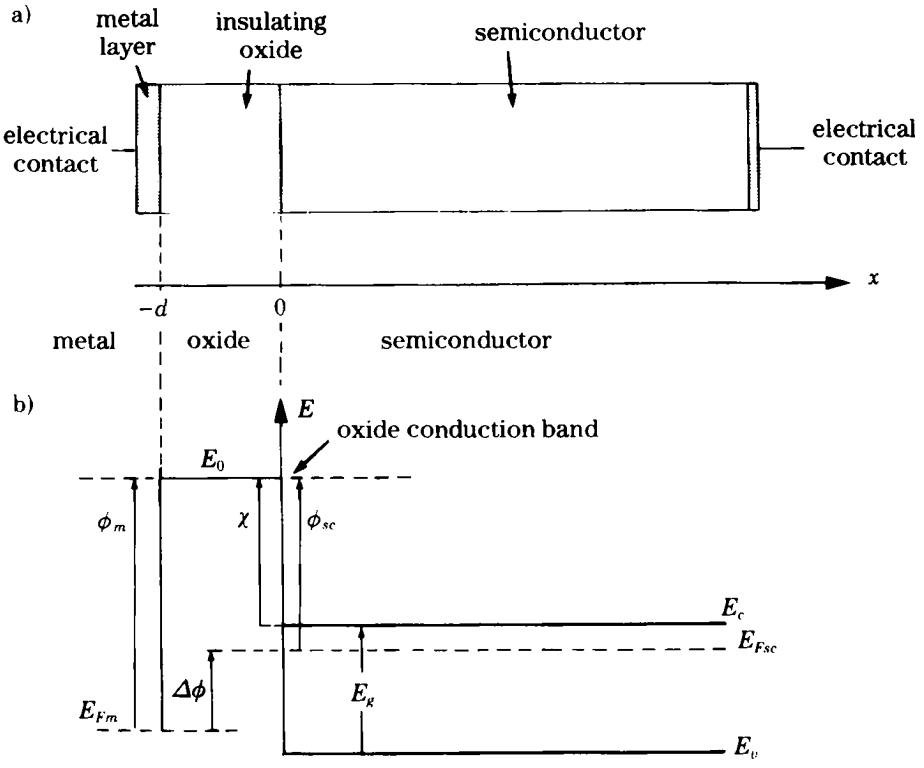


Fig. 10.30. (a) Schematic view of a MOS structure; (b) energy diagram of this structure when the end contacts are not electrically connected.

MOS Capacitor without Applied Voltage

In Part 1 of this discussion we connect the two electric contacts by an external lead without any applied voltage (short circuit). Figure 10.31 shows the resulting energy diagram for this structure.

1. Explain how the charges move to establish equilibrium, defined by the equality of Fermi levels in the metal and semiconductor. Justify the new energy diagram of Fig. 10.31 quantitatively, in particular the bending of the bands at the surface. Can you explain briefly why the interface affinity has not changed?

2. Deduce that there is now an electric field between the two plates of the capacitor formed by the MOS structure. Assume that there are no charges in the oxide and thus that the field is constant over the width of the oxide. Give the sign of the electric field as a function of the sign of $\Delta\phi$.

3. Under the conditions of Fig. 10.31 the energy of an electron depends on x, since there is an electrostatic potential $V(x)$ in the semiconductor and in the oxide caused by the charge distribution induced by the electrical contact between the metal and the semiconductor. The energy of an electron

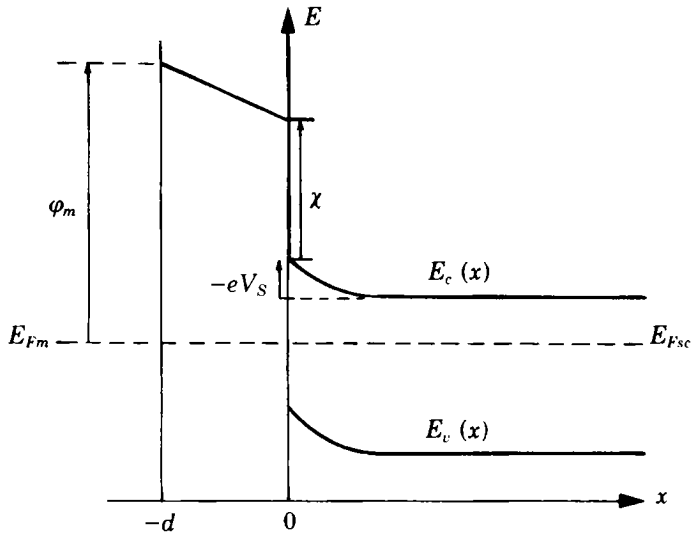


Fig. 10.31. Energy diagram of the structure shown in Fig. 10.30 when the metal and semiconductor are short-circuited by an external lead.

at the bottom of the semiconductor conduction band is $E_c(x) = E_c(+\infty) - eV(x)$, where $-e$ is the electron charge, and we assume that the induced electrostatic potential vanishes far from the interface, i.e., for x large and positive.

Write down the continuity of the electric displacement vector $\mathcal{D} = \epsilon\mathcal{E}$, where ϵ is the dielectric constant of the medium ($\epsilon = \epsilon_0\epsilon_r$) and $\mathcal{E}$ the electric field. Deduce a relation between the electrostatic potential $V_S = V(x = 0)$ at the oxide–semiconductor interface and its derivative at $x = 0$ for $x > 0$, $(\partial V/\partial x)_{0_+}$ involving $\Delta\phi, e, d$ and the dielectric constants $\epsilon_{ox}, \epsilon_{sc}$ of the oxide and semiconductor. Sketch the form of $V(x)$ in the case $\Delta\phi > 0$ for $-d \leq x < +\infty$.

4. From now on we study an n-type semiconductor, uniformly doped with donor concentration N_d. We assume further that the Fermi level remains deep enough in the band gap for all the donors to be ionized. Express the densities $n(x), p(x), \rho(x)$ of electrons, holes, and charges as functions of $eV(x)/kT$, N_d, and the intrinsic density n_i. Write down the differential equation obeyed by $V(x)$.

MOS Capacitor with Applied Voltage

To simplify this study (and particularly the notation), we henceforth consider an "ideal" MOS structure, where $\Delta\phi = 0$. The energy level scheme when the metal and semiconductor are short-circuited by an external lead, is given in Fig. 10.32.

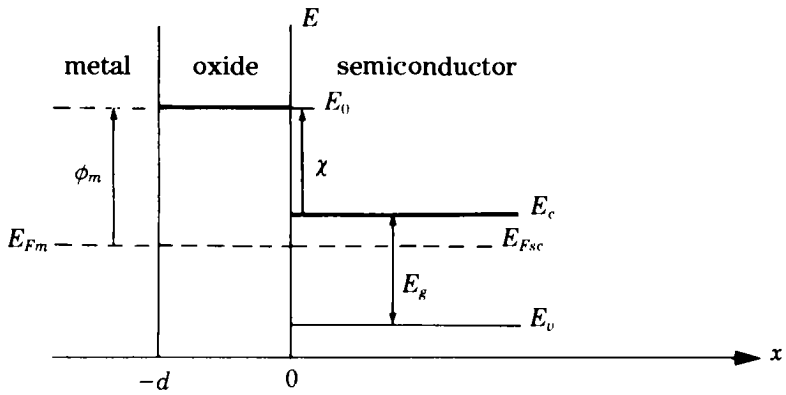


Fig. 10.32. Energy diagram of an "ideal" MOS structure. The figure should be compared with Fig. 10.31, which shows a "non-ideal" structure under the same conditions.

We now consider the MOS capacitor with a voltage V_G applied at the metal layer called the gate, the right-hand region of the n-type semiconductor being now at zero potential (Fig. 10.33).

5. The oxide is a perfect insulator, so, apart from a transient interval during the establishment of the voltage V_G, no current flows in the capacitor. What can we deduce about the Fermi level in the semiconductor and the relative positions of the Fermi levels in the metal and the semiconductor?

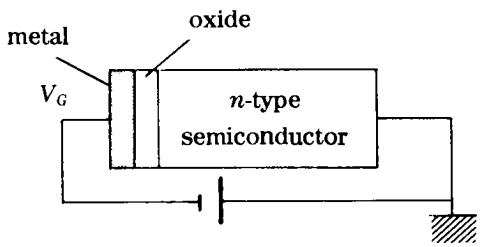


Fig. 10.33. MOS capacitor biased by the voltage V_G.

6. For the moment we limit ourselves to the case where V_G is negative and not too large in modulus ($|V_G|$ < about $E_g/2e$).

(a) Integrate the equation giving $V(x)$, assuming that the donors are ionized and using for $\rho(x)$ an approximation copied from that in the abrupt p–n junction (cf. Chap. 8). State this approximation and show that

$$V(x) = V_S\left(1 - \frac{x}{W}\right)^2 \quad \text{for } 0 < x < W, \tag{10.81}$$

$$V(x) = 0 \quad \text{for } x > W, \tag{10.82}$$

$$V(x) = V_S + (V_S - V_G)\frac{x}{d} \quad \text{for } -d < x < 0, \tag{10.83}$$

where the width W of the space charge depends on V_S. Give an expression for V_S as a function of N_d and W.

(b) Writing the continuity of the electric displacement vector $\epsilon\mathcal{E}$ across the oxide–semiconductor interface, give a relation between V_G, V_S, and W (V_G). Show that

$$W\ (V_G) = d\frac{\epsilon_{sc}}{\epsilon_{ox}}\left[\sqrt{1 - \frac{2V_G\ \epsilon_{ox}^2}{d^2\ e\ N_d\ \epsilon_{sc}}} - 1\right]. \tag{10.84}$$

Applications:

Calculate W for $V_G = -1$ V; $e = 1.6 \cdot 10^{-19}$ C; $\epsilon_{ox} = 3.2 \cdot 10^{-11}\ \mathrm{F \cdot m^{-1}}$; $\epsilon_{sc} = 10^{-10}\ \mathrm{F \cdot m^{-1}}$; $d = 150$ nm; $N_d = 10^{22}\ \mathrm{m^{-3}}$.

The maximum field that can be supported by silicon without breakdown is 10^9 V/m. What is the maximum value of W for n-type silicon ($N_d = 10^{22}\ \mathrm{m^{-3}}$)?

7. Under the preceding approximations, give an expression for the total charge appearing in the semiconductor when we apply a voltage V_G to the gate:

$$Q_S\ (V_G) = \int_0^\infty \rho\ (x)\ dx.$$

A charge $Q_G = -Q_S$ appears at the metal–oxide interface, and we define the differential capacitance per unit area of the MOS structure by the relation $C = dQ_G/dV_G$.

Show that C can be regarded as the resultant capacitance when the capacitance $C_{ox} = \epsilon_{ox}/d$ of the oxide layer and the capacitance C_{sc} of the space-charge region are placed in series, where

$$C_{sc} = -\frac{dQ_S}{dV_S} = \frac{\epsilon_{sc}}{W(V_G)}. \tag{10.85}$$

Calculate C_{ox} and C_{sc} for the data given in **6**.

8. Sketch the forms of the energy levels in the following cases:

(a) $V_G \leq 0$; $|V_G| \lesssim E_g/2e$;

(b) $V_G << -E_g/e$;

(c) $V_G > 0$.

Explain briefly why the cases (a), (b), (c) are known, respectively, as depletion, inversion, and enhancement. Sketch the curves giving $V(x)$ for $-d \leq x \leq +\infty$ in the various cases.

9. Find the value $V_G = V_{\text{threshold}}$ for which there is complete depletion of a region of the semiconductor of thickness $l = 200$ nm doped with $N_d = 10^{23}\ \mathrm{m^{-3}}$, when the oxide layer has thickness $d = 100$ nm.

For what value of V_G do we reach inversion? Take the intrinsic concentration as $n_i = 1.6 \times 10^{16}\ \mathrm{m^{-3}}$.

Applications of the MOSFET

In practice the semiconductor is used as a resistor whose value can be controlled by use of the applied gate voltage V_G. We thus consider a square of semiconductor in the y–z plane, with total thickness l in the x direction. On one of its faces is an oxide layer of thickness d and a metal layer. On the other face is a metallic contact, through which we can apply a voltage V_G to the MOS structure as described in Part 2. We denote as the transverse resistance $R(V_G)$ the resistance measured for a current flowing in the z direction while a voltage V_G is applied across the two sides of the structure (Fig. 10.34).

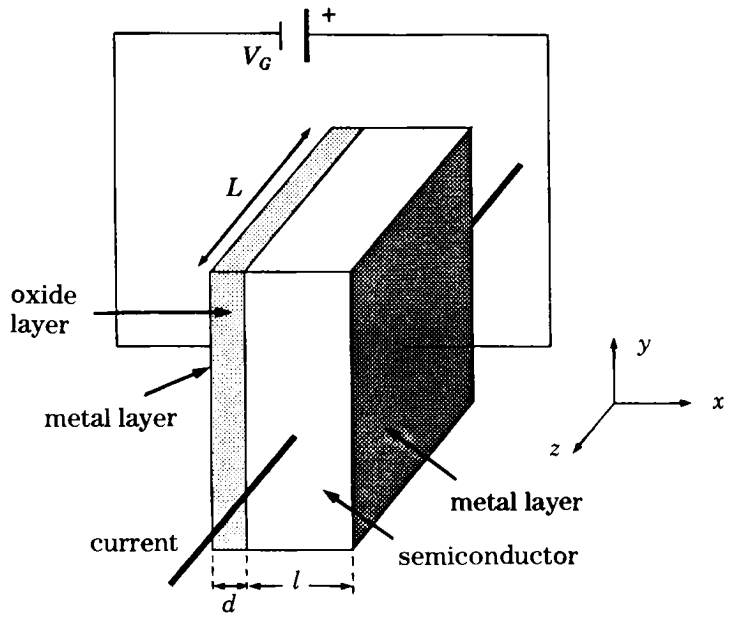


Fig. 10.34. Schematic view of a MOSFET.

10. Show that $R(V_G)$ does not depend on the surface area of the square and express $R(V_G)$ as a function of $R(V_G = 0) = R_0$, l, and $W(V_G)$ in the depletion regime (a) of **8**.

The interesting quantity for applications is the differential conductance $G = d(1/R)/dV_G$. Show that $G(V_G)$ is proportional to $C(V_G)$ and that the constant of proportionality is the mobility of the charge carriers in the semiconductor.

11. We observe that when $|V_G|$ is increased with $V_G < 0$, the resistance R first grows monotonically, and for sufficiently small l it reaches effectively

infinite values. On the other hand if l is large enough there is a value of V_G above which R decreases abruptly. Can you explain this?

Solutions

1. The alignment of the Fermi levels required for equilibrium is achieved by a flow of electrons from the semiconductor to the metal, since $E_{Fsc} - F_{Fm} = \Delta\phi > 0$, thus charging the semiconductor positively and the metal negatively. The upward bending of the conduction band results from an electron depletion near the interface. The bands stay parallel because the width of the band gap is an intrinsic property of the semiconductor; similar considerations hold for $\chi = (E_0 - E_c)_{x=0}$, the energy required to transfer an electron from a quantum level in the interior of the semiconductor to a well-defined level at the exterior, near the interface.

2. As charges have now appeared on both sides of the oxide layer, there is an electric field from the semiconductor to the metal if $\Delta\phi > 0$ and in the opposite direction if $\Delta\phi < 0$.

3.

$$\epsilon_{ox}\mathcal{E}_{ox} = \epsilon_{sc}\,\mathcal{E}_{sc}; \quad \mathcal{E}_{ox} = -\frac{\epsilon_{sc}}{\epsilon_{ox}}\left(\frac{dV}{dx}\right)_{0+}. \tag{10.86}$$

The quantity $\Delta\phi$ is now the energy shift between the two regions with flat bands in the metal and the semiconductor. This shift is split between the oxide and the region where the bands bend in the semiconductor. The energy jump χ at the semiconductor surface remains the same. The electrostatic potential $V(x)$ shown in Fig. 10.35 varies continuously.

We have

$$\begin{aligned}\Delta\phi &= [-eV(-d)] - [-eV(\infty)] \\ &= -e[(V(-d) - V(0)) + (V(0) - V(\infty))] \\ &= -e\,(d\,\mathcal{E}_{ox} + V_S),\end{aligned}$$

so that

$$\frac{\Delta\phi}{e} = \frac{\epsilon_{sc}}{\epsilon_{ox}} d\left(\frac{dV}{dx}\right)_{0+} - V_S. \tag{10.87}$$

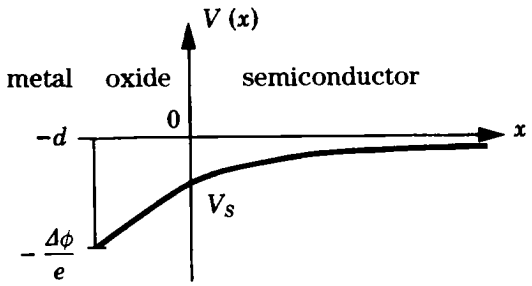


Fig. 10.35. Electrostatic potential across a MOS capacitance in the case $\Delta\phi > 0$.

4.

$$n\,(x) = N_c \exp\left(-\frac{E_c(x) - E_F}{kT}\right)$$

$$= N_c \exp\left(-\frac{[E_c(\infty) - eV(x)] - E_F}{kT}\right),$$

$$n\,(x) = N_c \exp\left(-\frac{E_c(\infty) - E_F}{kT}\right) \exp\frac{eV(x)}{kT}$$

$$= N_d \exp\frac{eV(x)}{kT}, \tag{10.88}$$

$$p\,(x) = \frac{n_i^2}{n(x)} = \frac{n_i^2}{N_d} \exp\left(-\frac{eV(x)}{kT}\right). \tag{10.89}$$

The potential $V(x)$ obeys the Poisson equation

$$\frac{\partial^2 V}{\partial x^2} = -\frac{\rho\,(x)}{\epsilon_{sc}} \quad \text{with} \quad \rho\,(x) = e\,(N_d - n + p) \tag{10.90}$$

or

$$\frac{\partial^2 V(x)}{\partial x^2} = -\frac{eN_d}{\epsilon_{sc}}\left[1 - \exp\frac{eV(x)}{kT} + \frac{n_i^2}{N_d^2} \exp\left(-\frac{eV\,(x)}{kT}\right)\right]. \tag{10.91}$$

5. There is no current, so E_F is constant in the semiconductor.

$$E_{Fm} - E_{Fsc} = -eV_G. \tag{10.92}$$

6. In this case the Fermi level is always within the gap. We assume, as for the p–n junction (Sect. 8.2), that the electron density $n(x) = N_d \exp(eV(x)/kT)$ varies rapidly enough so that we can define a depletion length $W(V_G)$ such that

$$\rho\,(x) = N_d\, e \quad \text{for} \quad 0 < x < W(V_G),$$
$$\rho\,(x) = 0 \quad \text{for} \quad x > W(V_G),$$
$$\rho\,(x) = 0 \quad \text{for} \quad -d < x < 0. \tag{10.93}$$

The distance over which n changes from N_d to nearly zero is much smaller than $W(V_G)$ and we thus take $\rho(x)$ as varying discontinuously, as in Chapt. 8, so that

$$\frac{\partial^2 V}{\partial x^2} = -\frac{eN_d}{\epsilon_{sc}} \quad \text{for } 0 < x < W. \tag{10.94}$$

(a) For $0 < x \leq W$, $V(x) = V_S\left(1 - \frac{x}{W}\right)^2$, (10.95)

$$\text{with } V_S = -\frac{e\, N_d\, W^2}{2\, \epsilon_{sc}} \tag{10.96}$$

$$\text{for } x > W, \quad V\,(x) = 0, \tag{10.97}$$

$$\text{for } -d < x < 0, \quad \frac{\partial^2 V}{\partial x^2} = 0 \tag{10.98}$$

with boundary conditions $V(-d) = V_G; V(0) = V_S$, so that

$$V\,(x) = V_S + (V_S - V_G)\frac{x}{d}. \tag{10.99}$$

(b) We have continuity of the displacement vector at $x = 0$:

$$\epsilon_{ox}\, \mathcal{E}_{ox}(0) = \epsilon_{sc}\, \mathcal{E}_{sc}(0) \tag{10.100}$$

or, using the expressions for $\mathcal{E} = -dV/dx$ in the two materials,

$$-\epsilon_{ox}\frac{V_S - V_G}{d} = \frac{2\,\epsilon_{sc}}{W} V_S \tag{10.101}$$

giving

$$V_G = V_S\left(\frac{\epsilon_{sc}}{\epsilon_{ox}}\frac{2d}{W} + 1\right). \tag{10.102}$$

Substituting expression (10.96) for V_S in Eq. (10.102):

$$V_G = -\frac{e\, N_d}{2\, \epsilon_{sc}} W^2 \left(\frac{2\, \epsilon_{sc}}{\epsilon_{ox}}\frac{d}{W} + 1\right), \tag{10.103}$$

$$W^2 + 2d\frac{\epsilon_{sc}}{\epsilon_{ox}} W + \frac{2\epsilon_{sc}}{eN_D} V_G = 0 \tag{10.104}$$

$$W = \frac{\epsilon_{sc}}{\epsilon_{ox}} d \left[\sqrt{1 - \frac{2\, V_G\, \epsilon_{ox}^2}{d^{\,2}\, e\, N_d\, \epsilon_{sc}}} - 1\right]. \tag{10.105}$$

Numerical applications: for the given values $W(V_G) \simeq 120$ nm. The maximum field in the oxide is $(\epsilon_{sc}/\epsilon_{ox})\mathcal{E}_{\max}$, where $\mathcal{E}_{\max}$ is the field in the semiconductor. Integrating $dE/dx = \rho/\epsilon_{sc}$ we get $eN_d W_{\max}/\epsilon_{sc} = \mathcal{E}_{\max}$, implying $W_{\max} = 20\ \mu$m.

7. The charge Q_S appears in the space-charge region, of width $W(V_G)$:

$$Q_S\,(V_G) = \int_0^W e\,N_d\,dx = e\,N_d\,W(V_G), \tag{10.106}$$

$$Q_G = -Q_S = -e\,N_d\,W, \tag{10.107}$$

$$dQ_G = -e\,N_d\,dW.$$

Differentiating Eq. (10.98) with respect to $W(V_G)$ we get

$$dV_G = -e\,N_d\left(\frac{d}{\epsilon_{ox}} + \frac{W}{\epsilon_{sc}}\right)dW \tag{10.108}$$

giving the differential capacitance per unit area C:

$$\frac{dV_G}{dQ_G} = \frac{1}{C} = \frac{d}{\epsilon_{ox}} + \frac{W}{\epsilon_{sc}} = \frac{1}{C_{ox}} + \frac{1}{C_{sc}} \tag{10.109}$$

with

$$C_{ox} = \epsilon_{ox}/d,\ C_{sc} = \epsilon_{sc}/W. \tag{10.110}$$

For the values given $C_{ox} = 2.1 \times 10^{-4}$ F/m^2; $C_{sc} = 8.3 \times 10^{-4}$ F/m^2.

8. (a) $E_g/2e < V_G < 0$.

We are in the depletion region since $n(0) < N_d$. (See Fig. 10.36.)

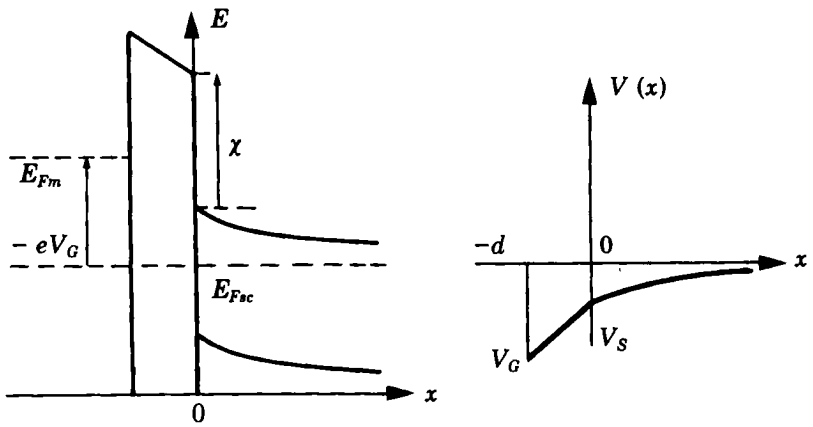


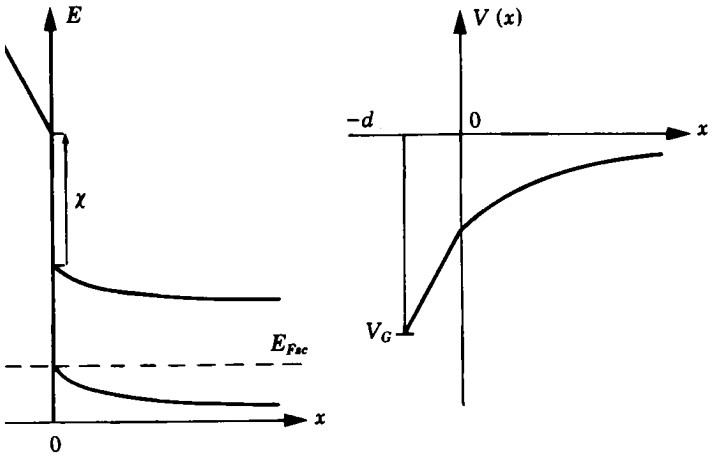
Fig. 10.36. Band profile and electrostatic potential in the depletion regime.

(b) $V_G << -E_g/e$.

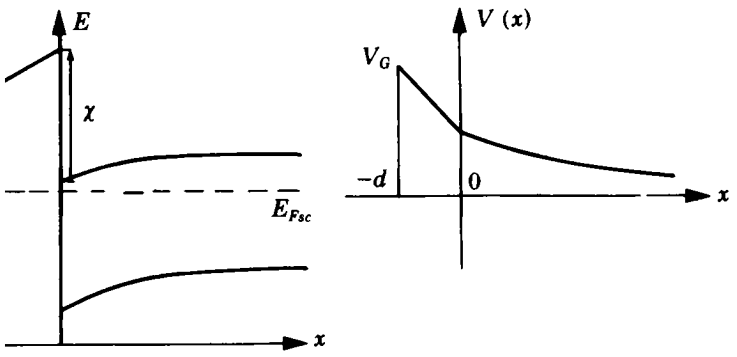
We are in the inversion region as $p(0) > n(0)$: the semiconductor has become p type (hole conduction) near the interface. (See Fig. 10.37.)

(c) $V_G > 0$.

We now have enhancement as $n(0) > N_d$. Electrons accumulate at the interface, as seen on Fig. 10.38.



ınd profile and electrostatic potential in the inversion regime.



ınd profile and electrostatic potential in the enhancement regime.

hold is such that $W(V_G) = 200$ nm. From the expression for get

$_1 = -13.2$ V.

ersion when

$$n\,(\infty) \approx N_d \tag{10.111}$$

$$p\,(0) = \frac{n_i^2}{N_d} \exp\left(\frac{eV_S}{kT}\right) \tag{10.112}$$

so that $\dfrac{V_S}{kT} = -\dfrac{2}{e} \log \dfrac{N_d}{n_i} \sim -0.78$ V.

From the expression for V_S:

$$W(V_G) = \sqrt{-\frac{2\,\epsilon_{sc}\,V_S}{eN_D}} \approx 0.1\ \mu\text{m}.$$

At inversion $W(V_G)$ is less than the channel width. The channel is inverted before it is fully depleted.

From the expression for V_G as a function of V_S: $V_G = -5.65$ V.

10.

$$\frac{1}{R} = \frac{\displaystyle\int_0^\infty n(x)\, e\mu_e\, dx \cdot L}{L} = \int_W^l N_d\, e\mu_e\, dx$$

$$= N_d\, e\mu_e\, l\left(1 - \frac{W}{l}\right), \tag{10.113}$$

$$R\,(V_G) = R_0\left(1 - \frac{W(V_G)}{l}\right)^{-1}, \tag{10.114}$$

where L is the length of a side of the square, l the semiconductor thickness, and μ_e the electron mobility in the semiconductor:

$$G\,(V_G) = \frac{d\,(1/R)}{dV_G} = -\mu_e\, N_D\, e\frac{dW(V_G)}{dV_G}$$

$$= \mu_e \frac{dQ_G}{dV_G} = \mu_e\, C(V_G). \tag{10.115}$$

11. When we increase $|V_G|$ with $V_G < 0$, W grows and can become equal to l before we reach the inversion regime. There are then practically no more mobile carriers in the semiconductor and its resistance is extremely high. In contrast, if l is large enough we reach the inversion regime before completely depleting the semiconductor. The conductance then increases abruptly because it arises from the mobile holes which appear in the inversion region at the oxide–semiconductor interface.

Supplement: Integrated Circuits with MOS Transistors

These are of two types: normally on and normally off, depending on whether the structure at rest ($V_G = 0$) conducts or not. These structures are also sometimes called depletion-mode or enhancement-mode, respectively. This means that when a voltage is applied the structure becomes non-conducting ("normally on") or conducting ("normally off"). Depending on the material of the channel (n or p) this change of state occurs for a positive or a negative voltage.

Figure 10.10 shows an enhancement-mode MOSFET with an n channel: when a voltage is applied between the source and drain electrodes (n type), one of the junctions between source or drain and substrate (p type) is reverse biased and does not allow current to flow.

When a sufficiently large *positive voltage* is applied to the gate, there will be inversion near the oxide, i.e., the appearance of significant numbers of conduction electrons; hence the name "n channel." Analogous reasoning shows that a structure with source-drain electrodes of p type and a substrate of n type (as in the problems) will conduct if a *negative voltage* is applied at the gate. This would be a p-channel enhancement-mode MOSFET.

The depletion structures are entirely analogous. We start from a situation of *inversion at rest* because of the presence of *static charges in the oxide.* These charges induce charges of opposite sign in the semiconductor. As we can easily introduce positive ions into the oxide (Na^+ ions) the usual case involves the introduction of an n channel with p-type source and drain contacts. Depletion-mode MOSFETs are n-channel transistors with n-type source and drain contacts. Applying a negative voltage to the gate then empties the channel of electrons. The channel becomes p type and one of the contacts blocks the current.

In each case it is clear that the channel conductance is a linear function of the applied voltage in a certain range. There is then a linear amplification effect: linear control of the source-drain current I_D by the gate voltage. The parameter $g_m = \partial I_D / \partial V_G$ is called the transconductance of the transistor and is one of the essential parameters for the user.

Microlithography techniques allow one to manufacture n-channel and p-channel MOSFETs simultaneously on the same substrate. In this way one produces complementary structures called CMOS (Complementary MOS) as in Fig. 10.39(a). The equivalent electrical scheme is given in Fig. 10.39(b). These structures have the property of having two stable states in which the current dissipation is infinitesimal. These are used in binary logic applications in which the signal is either 0 or 0.5 V.

One applies voltages $V_{\text{high}} = 5$ V at S_2 and $V_{\text{low}} = 0$ V at S_1 throughout. Assume that the input voltage is 5 V (Fig. 10.39(c)). The gates G_1, G_2 are positively polarized and electrons are attracted. We thus create an n channel

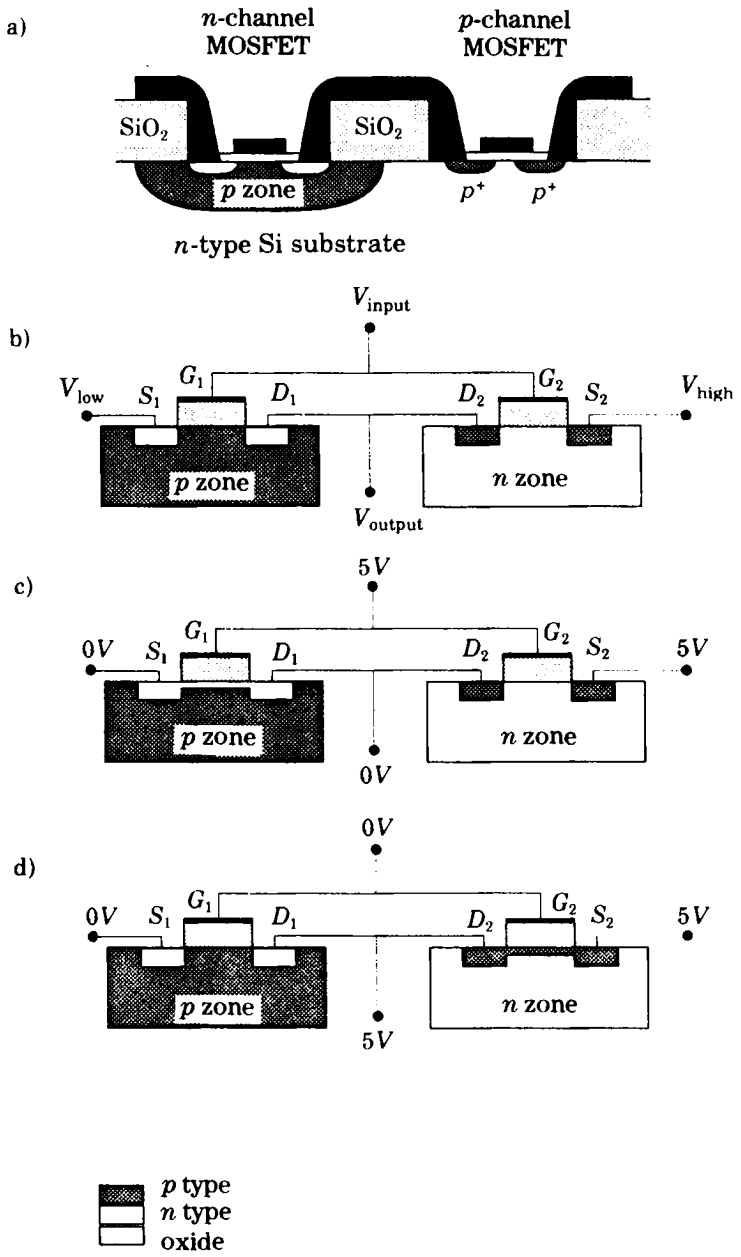


Fig. 10.39. (a) CMOS structure, with two complementary MOSFETs. The p region of the n-channel MOSFET is obtained by diffusion; (b) equivalent electrical scheme; operating as an inverter; (c) $V_{\text{input}} = 5$ V; (d) $V_{\text{input}} = 0$ V.

in the MOS on the left which becomes conducting. By contrast the MOS on the right is blocked since it contains two head–tail junctions in series. In these conditions the MOS behaves as a short circuit and $V_{\text{output}} \sim 0$ V.

On the other hand if $V_{\text{input}} = 0$ the left-hand MOS is blocked if we chose it "normally off," while that on the right conducts if it is "normally on." The output voltage is then 5 V. In both cases one of the two MOS in series is blocked and the current flowing in the circuit is very small. This device is an inverter of very low power dissipation.

MOSFETs constitute the most rapidly developing sector of integrated circuit production (annual growth rates of between 30%–60%). The reason for this success comes from the ease of manufacture (fewer masking levels than in bipolar circuits) and their low dissipation. The degree of integration is also higher because of the reduced size of the transistors, which are the building blocks of each device; indeed, 64-megabit memory chips are currently being developed and will presumably reach the market shortly. On the other hand MOSFETs are slower than bipolar transistors because of the large gate capacitances, caused by the presence of an insulating layer.

Values of the Important Physical Constants

Avogadro's number		N	$= 6.02\times10^{23}$
electron charge		e	$= -1.6\times10^{-19}$ coulomb
electron rest-mass		m	$= 0.91\times10^{-30}$ kg
Planck's constant		h	$= 6.624\times10^{-34}$ J·s
velocity of light		c	$= 2.9979\times10^{8}$ m·s^{-1}
electron specific charge		e/m	$= 1.76\times10^{11}$ C·kg^{-1}
electron radius	$r_0 = e^2/4\pi\epsilon_0 mc^2$		$= 2.82\times10^{-15}$ m
first Bohr radius		a_0	$= 5.29\times10^{-11}$ m
proton–electron mass ratio			$= 1{,}836.1$
Boltzmann's constant		k	$= 1.38\cdot10^{-23}$ J·K^{-1}
Bohr magneton		μ_B	$= -9.27\times10^{-24}$ J·T^{-1}
wavelength associated with 1 eV			$= 1.239$ μm
frequency associated with 1 eV			$= 2.418\times10^{14}$ Hz
energy associated with 1 K			$= 8.616\times10^{-5}$ eV
temperature associated with 1 eV			$= 11{,}605$ K

ome Physical Properties of emiconductors (286 K)

		E_g (eV)	μ_e ($cm^2/V.s$)	μ_h ($cm^2/V.s$)	$\varepsilon=n^2$	a angstr.	d ($g.cm^{-3}$)	$F.T.$ (°C)
IV	C	5.4	1 800	1 200	5.5	3.567	3.51	3 550
	Si	1.15	1 900	480	11.8	5.42	2.42	1 412
	Ge	0.65	3 800	1 800	16.0	5.646	5.36	958
	Sn	0.08	2 500	2 400		6.47	6.0	232
VI	Se	1.6		0.6	8.5	4.35 4.95	4.8	220
	Te	0.33	1 100	560	5.0	4.447 5.915	6.24	452
III V	BP	6		100	11.6	4.537	2.97	3 000
	Al P	2.5	3 500		11.6	5.43	2.85	1 500
	Al As	2.3	1 200	200		5.63	3.81	1 600
	Al Sb	1.52	400	150	10.3	6.13	4.22	1 060
	Ga P	2.25	80	17	8.4	5.44	4.13	1 350
	Ga As	1.42	8 500	400	13.5	5.65	5.31	1 280
	Ga Sb	0.69	4 000	650	15.2	6.095	5.62	728
	In P	1.27	4 600	700	10.6	5.869	4.78	1 055
	In As	0.35	30 000	240	11.5	6.058	5.66	942
	In Sb	0.17	70 000	1 000	16.8	6.48	5.775	525
IV IV	Si C	3.0	60	8	10.2	4.35	3.21	2 700
IV VI	Pb S	0.37	800	1 000	17.9	7.5	7.61	1 114
	Pb Se	0.26	1 500	1 500		6.14	8.15	1 062
	Pb Te	0.25	1 620	750		6.45	8.16	904
V VI	Bi_2Te_3	0.15	1 250	515		10.48	7.7	580
II V	Cd_3As_2	0.13	15 000			8,76	6.21	721
	Cd Sb	0.48	300	300		6,471	6.66	456
II VI	Zn O	3.2	190		8.5	5.18	5.60	1 975
	Zn S	3.65	100		8.3	5.423	4.80	
	Zn Se	2.6	100	16	5.75	5.667	5.42	1 515
	Zn Te	2.15		50	18.6	6.101	5.54	1 239
	Cd S	2.4	200		5.9	5.83	4.82	685
	Cd Se	1.74	500		4.30	6.05	5.81	1 350
	Cd Te	1.50	650	45	11.0	6.48	6.20	1 098
	Hg S	2.5			5.86	5.852	7.67	583
	Hg Se	0.3	18 500		14	6.08	8.5	798
	Hg Te	0.2	22 000	160		6.429	8.42	670

ne table gives the values of the band gap energy, electron and hole mo-lities, high-frequency dielectric constant, crystal lattice constant, density id fusion temperature.

Bibliography

Ashcroft, N. W. and Mermin, M. D. 1976.
Solid State Physics. New York: Holt, Rinehart and Winston. A general solid state physics textbook.

Handbook on Semiconductors. 1981. Amsterdam: North-Holland Publishing Co. A comprehensive presentation of semiconductor physics.

Kittel, C. 1986.
Introduction to Solid State Physics (6th ed.). New York: John Wiley and Sons, Inc. A general solid state physics textbook.

Sze, S. M. ed., 1991.
Semiconductor Devices: Pioneering Papers. Singapore: World Scientific.

Willardson, R. K., Beer, A. C., and Weber, E. R., eds.
Semiconductors and Semimetals (37 volumes.) Boston: Academic Press. Contains reviews on the various aspects of semiconductor research.

Index

Based on courses given at the Ecole Polytechnique in France, this book covers not only the fundamental physics of semiconductors, but also discusses the operation of electronic and optical devices based on semiconductors. It is aimed at students with a good background in mathematics and physics, and is equally suited for graduate-level courses in condensed-matter physics as well as for self-study by engineers interested in a basic understanding of semiconductor devices.

ISBN 0-387-94024-3
EAN
9 780387 940243 >

SBN 0-387-94024-3