2

The Physical Background

2.1 Introduction

SOLID-STATE electronics is based on the properties of semiconductors or, more specifically, on the properties of the junctions between a semiconductor and any another material that can be a metal, a semiconductor or an insulator.
To study solid-state devices it is necessary to understand the properties of the charge carriers whose distribution in space and in energy creates the properties of the junctions and the relationships between the applied voltage, the accumulated charge, and the flowing current.
The behavior of most of the devices can be adequately described by a sort of a combination of the concepts of quantum mechanics and classic physics. This is a practical approach where quantum mechanics provides the explanation of the energy distribution of electrons, but the transport phenomena can be still described by classical concepts.
It is always important to keep in mind that theories in sciences are valid so far they can predict and explain the experimental behaviors. To this regard, the validity of such *semi-classic* models is mainly limited by the dimensions of the devices. So, when the dimensions become smaller than tens of nanometers the classical concepts lose their validity and a full quantum description is required.

2.1.1 The phenomenology of semiconductors

As *nomen omen* the most evident property of semiconductors is their resistivities which lies between those of conductors and insulators. Indeed, the resistivity (ρ) of semiconductors occurs in a range from $\rho = 10^6$ to 10^{-2} Ωm, while in metals it is in the interval from $\rho = 10^{-4}$ and 10^{-8} Ωm and in insulators it goes from $\rho = 10^{10}$ up to 10^{18} Ωm. The inverse of resistivity is the conductivity (σ). It is worth to remind that the resistance of a resistor, defined as the ratio between voltage and current and measured in Ω, is a combination between the resistivity (a property of the material) and the geometrical shape of the resistor. In longitudinal resistors this combination is expressed by the well-known relationship:

$$R = \rho \frac{l}{A}$$

where l is the length of the conductor and A is the section through which the current flows.
The conductivity of semiconductors strongly depends on their chemical composition. Hence, it can be altered by the addition of impurities either in the bulk or at the surface of the material. The

first case is technologically exploited, as it will be discussed later, to modify the conductivity of semiconductors while the second case provides the basis for many chemical sensors.
The limited conductivity made semiconductors not interesting for electric applications. However, theyexhibit other peculiar behaviors whose explanation was made possibile by quantum mechanics. For instance, the relationship between conductivity and temperature is characterized by a *negative temperature coefficient*. Namely, the resistance decreases with the temperature. While, in general, metals show the opposite behavior (*positive temperature coefficient*). This peculiar property of semiconductors is used to sense the temperature in a class of sensors called thermistors. Furthermore, semiconductors are photoconductors. Namely the conductivity increases when the material is shined by a light with a wavelength λ greater than a λ_0 which is specific for each material.

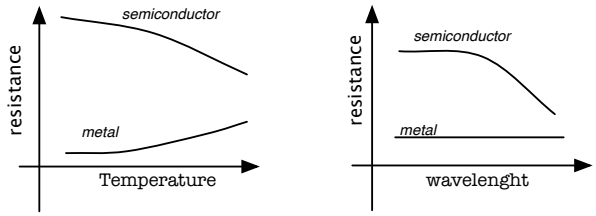


Fig. 2.1. *Behaviour of the resistance of metals and semiconductors as a function of temperature and light.*

These differences are ultimately related to the nature of the bonds that keep the atoms together. Metal atoms are hold together by metal bonds while semiconductors are built by covalent bonds. In case of metal bonds, the electrons participating to the bonds (valence electrons) are not localized but they are equally distributed in the space around the atoms. This leads to the formation of a population of weakly bonded electrons that are almost free to move inside the solid. In case of covalent bonds, each valence electron remains localized in a molecular orbital shared by two adjacent atoms. As a consequence, valence electrons are strongly bonded to their own atoms and a non negligible amount of energy is required to make them free to move. However, in semiconductors the energy provided by the temperature is sufficient to break a limited but not negligible amount of such bonds.

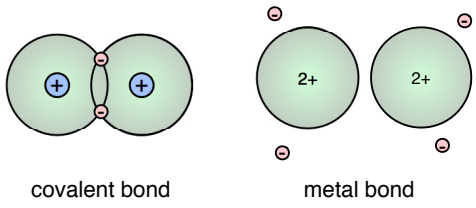


Fig. 2.2. *Simple picture of covalent and metal bonds.*

2.2 Electrons in solids

In quantum mechanics, the energy of particles confined in a closed space is quantized in a finite number of energy levels. The confinement is provided by the forces acting on the particle, and the confinement in space is equivalent to a potential well. Besides the quantization of energy levels, quantum mechanics introduces additional rules to accomodate particles onto the energy levels. Electrons, and in general all the particles with a semi-integer spin, are subjected to the Pauli principle that states that each quantum state can contain only one electron. Since the spin can take two values ($s = \pm\frac{1}{2}$), no more than two electrons are allowed per each energy level. Thus, N levels can accommodate 2N electrons, and 2N electrons require the existence of at least N levels.
Energy levels are calculated solving the Schrodinger equation once the field at which electrons are subjected is known. The solution can be exactly calculated for simple atoms under the hypothesis that the nucleus is still. More complex cases require further simplifications.
Interactions among atoms provides the bonds which enable the atoms to aggregate in structures of growing complexity such as molecules, liquids, and solids. Solids in particular, can aggregate either into ordered structures called crystals or disordered (amorphous) structures.

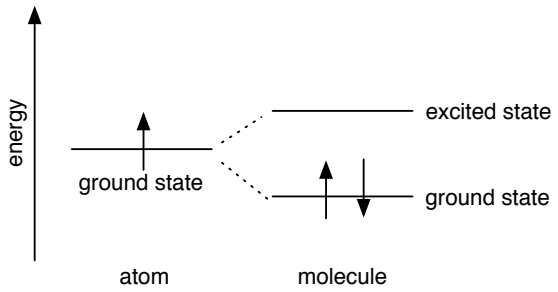


Fig. 2.3. *The molecule of hydrogen offers the simplest example of orbitals multiplication. The ground state of hydrogen atom splits in two molecular orbitals. The lowest is the molecular ground level while the highest is the excited state. The energy of the molecular orbitals are smaller and greater than the atomic orbitals.*

The interactions among atoms involve the electrons of the outer atomic shells. The mutual interaction among these electrons (valence electrons) provides the "glue" that keep the atoms together. Electrons in identical and non interacting atoms should have the same energy levels (orbitals), but when the atoms are interacting, the electrons of one atom "feel" the presence of the electrons in the adjacent atoms, and Pauli principle does not allow that these electrons stay on the same energy level. Thus, to obey to the Pauli principle, the interacting electrons have to slightly change their energy. This leads to a multiplication of energy levels. In practice, the original atomic levels split into a number of orbitals roughly equal to the number of the atoms involved in the interaction. In the case of molecules the atoms are few, and then the orbital multiplication still leads to discontinuous energy levels. But in the case of solids, where the number of atoms is very large (e.g. the density of atoms in silicon is about $5 \cdot 10^{22}\ cm^{-3}$) the multiplication of energy levels lead to a quasi continuum distribution of energy levels called *energy band*.

How energy levels combine together depends on the nature of the atoms. In some cases, the separation between orbitals, occurring in atomic energy levels, is preserved but in other cases it is cancelled and all the orbitals merge into a continuous band of energies.

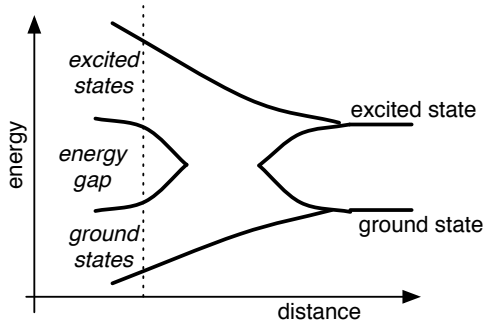


Fig. 2.4. *The figure qualitatively illustrates the process of orbitals multiplication and the merging of distinct energy levels into continuous bands of energies. In figure the orbitals are plotted as a function of distance. The distance is related to the strength of the interaction. According to the interatomic distance, fundamental and ground states may either form a continuous band or split in two bands separated by an energy gap. The first is the case of conductors, the latter is the case of semiconductors.*

2.2.1 Orbitals splitting and coupled oscillators

The split of orbitals is similar to the split of frequencies observed in coupled oscillators. Let us consider for instance two identical LC circuits. The resonant frequencies of the isolated circuits are obviously coincident, but in the coupled circuits the resonant frequencies split in two distinct values: one smaller and one larger than the unperturbed resonant frequency. The analogy between particles and oscillators is supported by the fact that in quantum mechanics electrons are described by waves and stable orbitals are similar to steady oscillators.

2.2.2 Crystals, periodic potentials and energy gaps

Crystals form a particular class of solids where atoms are arranged in a regular periodic pattern. In ideal crystals the pattern is infinitely repeated. In real crystals, the perfect periodicity of the pattern is disturbed by defects such as the dislocation of atoms, and impurities (namely alien atoms included during the crystal growth). A obvious deviation from ideality occurs at the surface where the periodic pattern abruptly stops.
Respect to the valence electrons, the rest of the atom is positively charged, thus the regular pattern of atoms in the crystal gives rise to a periodic potential for the valence electrons. Figure 6 shows a simplified view of the electric potential in a mono dimensional crystal. The potential energy of the electrons of the last atom decays outside. Close to the surface it gives rise to the surface potential. Space periodicity is defined by the so-called Wiener cell. This is the elementary cell defined by the smallest arrangement of atoms that is infinitely repeated. The space periodicity is complemented

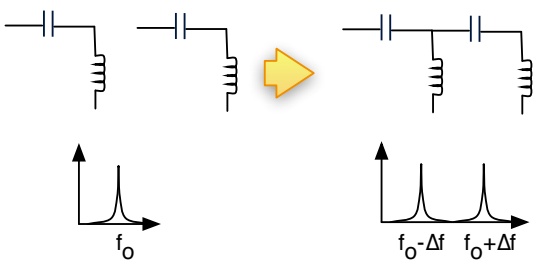


Fig. 2.5. *The resonance frequency of a LC circuit splits in two frequencies when the resonators are coupled. The phenomenon is analog to energy levels splits in interacting atoms*

by the periodic behavior in the reciprocal space. The reciprocal space is the Fourier transform of the real space. The elementary cell in the reciprocal space is called the Brillouin zone.

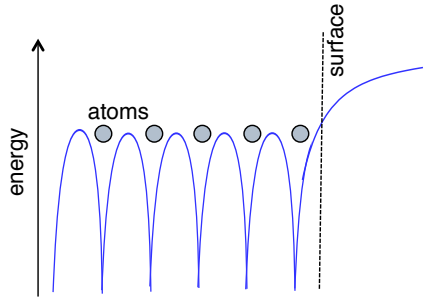


Fig. 2.6. *The periodic electric potential due to a periodic sequence of atoms. The last atom determines the surface potential.*

In quantum mechanics, particles are represented by wave functions which are periodic in time and in space. The frequency ω defines the periodicity in time and the wavelength λ the period in space. The space period is conveniently represented in the reciprocal space by the wavenumber

$$k = \frac{2\pi}{\lambda}$$

.
For this reason, the reciprocal space is also called the k-space. A traveling particle, such as an electron, corresponds to a wave $(exp(i(\omega t - kx))$ where frequency and wavenumber are proportional to the energy $(E = \hbar\omega)$ and the momentum $(p = \hbar k)$ respectively. Where $\hbar = h/2\pi$ and h is the Planck's constant $(h \approx 6.62 \cdot 10^{-34}\ J/s)$.
It is important to observe that only ideal sinusoids are characterized by a unique frequency and a unique wavenumber. Actually, real particles are confined both in space (e.g. solids) and in time. Thus, instead of single values of frequency and wavenumber, real particles are characterized by

distributions of frequencies ($\Delta\omega$) and wavenumbers (Δk). Eventually, real particles correspond to the superposition of many waves called a wave-packet.
Energy and momentum are connected by the so called *dispersion relationship*. This function defines the conditions of propagation of the wave packet and it depends on the forces acting on the particle. The motion equation in quantum mechanics is the Schrodinger equation:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V(x)\Psi(x,t) = i\hbar\frac{\partial\Psi(x,t)}{\partial t} \tag{2.1}$$

For a free particle, namely when the potential is null ($V = 0$), the solution of the Schrodinger equation is a plane wave

$$\Psi(x,t) = A \cdot exp(i(kx - \omega t)) \tag{2.2}$$

A free electron can assume any energy value. The relationship between energy ($\hbar\omega$) and momentum ($\hbar k$) is:

$$E = \frac{\hbar}{2m}k^2 \tag{2.3}$$

A different picture emerges when the particle is exposed to a potential. The simplest case is when the particle is confined into a infinite potential well (particle in a box). In this case the wave vectors depends on the size of the box (L), and the longest possible wavelength is $\lambda = 2L$ and the wavelength of the harmonics is $\lambda = 2L/n$ where n is an integer. Then $p = \hbar k = \hbar 2\pi/\lambda = \hbar n\pi/L$. From the energy momentum relationship ($E = \hbar\omega = \hbar^2k^2/2m$) the energy levels are found

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2}$$

The dispersion relation of equation 2.3 still holds but only a limited set of energy is allowed, and the dispersion relation is not continuous. It is worth to note that when the potential is not zero, the total energy of the particle is $E = T + V$ then the relationship $E = p^2/2m$ is no more valid. The quantity k is not the actual momentum but rather it still describes the number of peaks of the wave inside the box namely the wavenumber.
In a solid the electrons, besides to be confined, undergo the action of the potential generated by the atoms. In case of a crystal this potential is periodic. An important theorem of quantum mechanics (Bloch's theorem) states that a particle in a periodic potential is described by a periodic wave function. Then the properties of the electrons are periodic in the crystal, hence the behavior in the elementary cell is the behavior in the whole material.
Exact solutions require the knowledge of the actual potential, however some simple toy models can elucidate the general properties. To this regard, a simple case is the Kronig-Penney model where the potential is made by an infinite sequence of pulses. The Schrodinger equation applied to such a potential results in a dispersion relationship that is discontinuous in energy. In practice a energy gaps appear. The shape of the dispersion relation approximates the free particle around k=0 and deviates from the free particle as k approaches the border of the elementary cell.

The potentials inside real solids are obviously more complex than those in the Penney-Kronig model. Indeed, atoms are arranged in tridimensional structures and, in general, more than an atomic species is present. As a consequence, the shape of the bands may be rather complex.
Fig. 8 shows the calculated dispersion relationship of silicon. The continuous branches of these plots form the band energies and such a diagram is also known as bands diagram.

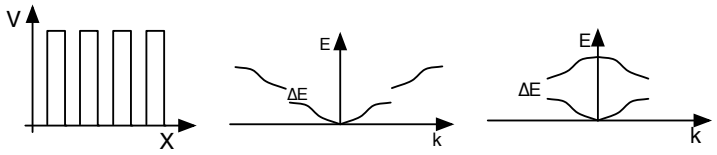


Fig. 2.7. *Kronig-Penney potential periodic in the real space. Dispersion interaction E(k) and dispersion interactions plotted in the first Brillouin zone.*

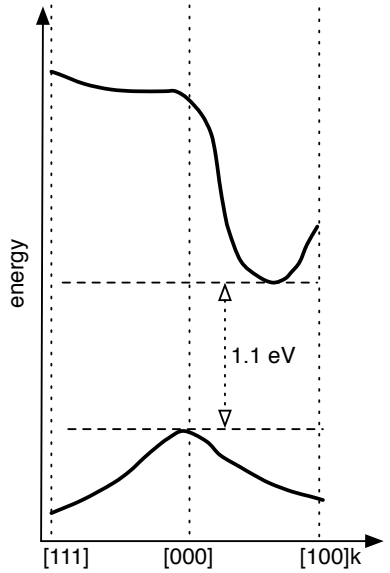


Fig. 2.8. *Calculated band diagram of silicon. The coordinates in k space are given as Miller indexes. This is a method to describe the main directions in crystals .*

The lower band identifies the ground state, the **valence band**, while the upper band is the excited state, the **conduction band**. The electrons in the conduction band are quasi-free particles that can be kept in movement by an applied electric field.
The passage from the valence band to the conduction band is possible if the electrons receive, from an external source, an extra amount of energy and momentum necessary to move from the top of the valence band to the bottom of the conductance band.

2.2.3 The distribution of the electrons in the energy levels

It is straightforward that without any external input of energy the electrons, like any other physical system, occupy the lowest available states. The temperature is a ubiquitous energy contribution that the electrons and the whole material receive. Thermal equilibrium is the condition where all

the elements of the system (electrons, atoms, external world) share the same temperature and there is not any net transfer of energy from one element to the other.
Let us consider a simplified, but useful for our scopes, band diagram where conduction and valence bands are separated by a gap of energy. In practice the diagram is restricted to the bottom of the conduction band and the top of the valence band where electrons and holes are quasi-free particles ($E \propto k^2$).
The distribution of electrons in the available states is ruled by a statistical law. Thus, rather than describing the behavior of each electron, the average behavior of a large population of electrons is considered. It is important to note that statistics is valid for the average but individual electrons can strongly deviate from the collective behavior.
Of course, the statistical approach is justified by the large number of electrons in the material. In the case of silicon, the density of atoms is $5 \cdot 10^{22} \frac{atoms}{cm^3}$. Then, since each atom of silicon has 4 valence electrons, the density of electrons that have to be distributed between valence and conductance bands is $4N = 2 \cdot 10^{23} cm^{-3}$.
On the other hand, all measurable quantities always involve a large number of electrons. As an example, let us consider a tiny current such as 1 pA. This is equivalent to 10^{-12} Coulombs per second across a section of the conductor. Since electron charge is $1.6 \cdot 10^{-19} C$, 1 pA corresponds to a flow of about 10^7 electrons per second.
The concentration of electrons with an energy between E and $E + \Delta E$ is given by the product of the density of available states in the energy intervals times the probability that electrons can actually have energies in that interval.

$$n = \int_E^{E+\Delta E} g(E) \cdot f(E) dE \tag{2.4}$$

The function g(E) is the density of the allowable states. This quantity depends on the nature of the atoms and the characteristics of their interactions. Namely on the overlap and multiplication of the atomic orbitals. The function f(E) is the probability function that depends on the total number of electrons. The probability is a function of the temperature, greatest the temperature largest the probability to find electrons at high energy.
In classical physics, the probability function for non mutually interacting particles is the Boltzmann partition function.
The Boltzmann equation is a direct consequence of the hypothesis of non interacting particles. Indeed, considering a system made of two states, the probability of occupancy of the total system is the product of the probability of occupancy of each state ($p(1,2) = p(1) \cdot p(2)$) and the total energy is the sum of the energies of the two states ($E(1,2) = E(1) + E(2)$). Then the probability to find the system at the energy $E(1,2)$ is: $P(E_1 + E_2) = P(E_1) \cdot P(E_2)$. This condition is fulfilled by the exponential function where the function of a sum is the product of the functions of the individual arguments ($e^{A+B} = e^A \cdot e^B$). Then the probability function can be written as: $P = const \cdot exp(-\beta E)$ where $\beta = 1/kT$ where k is the Boltzmann constant ($k = 1.38 \cdot 10^{-23} JK^{-1}$).
Given $N = n_1 + n_2$ distributed in two energy levels (E_1 and E_2), the ratio between the number of particles in the two states (n_1 and n_2) is:

$$\frac{n_2}{n_1} = exp\left(-\frac{E_2 - E_1}{kT}\right) \tag{2.5}$$

The distribution of particles in the two states is driven by the temperature. At T=0 K all particles lie in the lower states $\frac{n_2}{n_1} = 0$, while at infinite temperature $\frac{n_2}{n_1} = 1$ and the particles are

equally distributed in the two states.
The classical statistical theory fails in case of elementary particles for which quantum concepts holds. In particular, electrons, like any other particle with non integer spin, obey to the Pauli principle of exclusion. Pauli principle states that no more than two electrons can be found in the same state, then, even at T=0 K electrons cannot lie in a single ground state, but rather at least N/2 states are necessary to accomodate N particles, thus at T=0 K, the particles pile up the stack of states until a maximum allowable energy level is reached. The statistical law that incorporate the Pauli principle in the Boltzmann equation is the **Fermi-Dirac function**:

$$f(E) = \frac{1}{1 + exp\left(\frac{E-E_F}{kT}\right)} \tag{2.6}$$

The quantity E_F is called **Fermi level**. It is the highest energy level that can be occupied at T=0 K. The Fermi level depends on the total number of electrons and it is variable with the temperature. It will be shown later that the Fermi level is equivalent to the electrochemical potential of a population of non interacting charged particles.
The shape of the Fermi-Dirac function is qualitatively different in the two cases T=0 K and $T > 0\ K$. At T=0 K, $f(E \leq E_F) = 1$ and $f(E > E_F) = 0$. While, at $T > 0K$, the function assumes the following values: $f(E < E_F) < 1$, $f(E > E_F) > 0$, with the condition: $f(E = E_F) = \frac{1}{2}$. The Fermi-Dirac function is shown in figure 9.

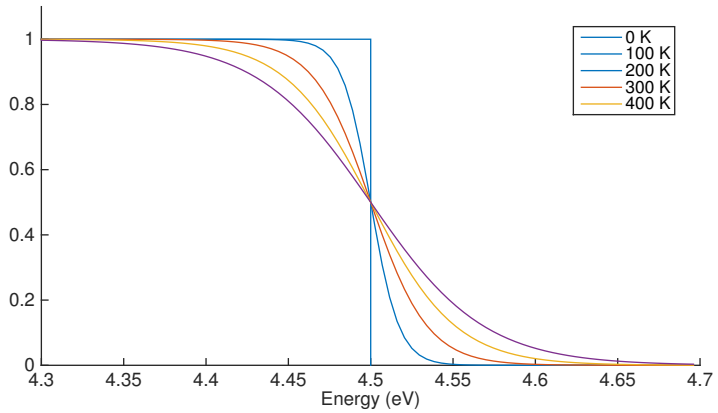


Fig. 2.9. *The Fermi-Dirac function as a function of energy and at different temperatures. As the temperature increases, the probability to find electrons at higher energy becomes large.*

Noteworthy, when $(E - E_F) \gg kT$ the Fermi-Dirac function is approximated by the Boltzmann equation (eq 2.3). At room temperature (T=300K) kT is approximately 26 meV. Then at room temperature, if $E - E_F \gg 26\ meV$ the Fermi- Dirac function can be written as:

$$f(E > E_F) \approx exp\left(-\frac{E - E_F}{kT}\right) \tag{2.7}$$

This exponential function happens to be ubiquitous in all the equations that describe the behavior of electronic devices. Its presence reminds of the statistical nature of the principles on which the devices are based.
Most of the electric characteristics of materials are consequence of the position of the Fermi level respect to the conduction and the valence bands. In semiconductors, the Fermi level occurs inside the gap between the valence and the conduction band. Then, at low temperature, the probability to find electrons in the conductance band is practically zero.
Figure 10 shows the simplified band diagram with superimposed both the Fermi-Dirac function and the density of states. The density of states is obviously zero in the band gap. The states in their respective bands have a parabolic dependence from the energy. The function g(E) will be explicitly calculated later.
In pure silicon, the concentration of electrons in the conductance band at room temperature is of the order or $10^{10} cm^{-3}$. This number determines the small, but non negligible, conductivity of pure silicon.

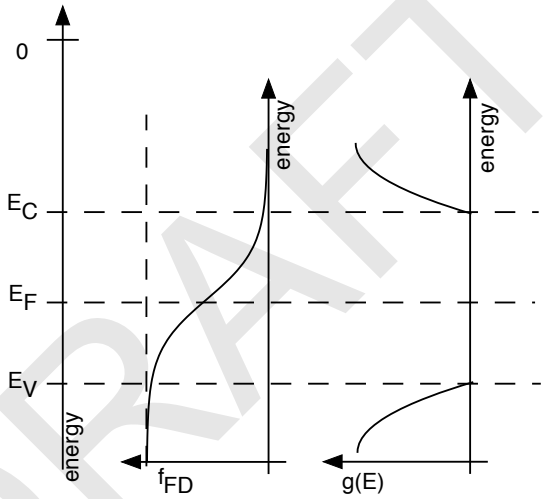


Fig. 2.10. *Comparison of the Fermi-Dirac function and the density of states. The 0 energy level defines the condition of free electron, namely an electron not bounded to the material. As a consequence all the energies of the electrons inside the material are negative (binding energies). The density of states is zero in the band gap.*

Electrons in valence band: the concept of holes

The Fermi-Dirac function describes the probability of finding electrons at a given energy. As the temperature increases, the probability to find electrons in the conductance band increases. This means that electrons engaged in covalent bonds leave their location and can be kept in movement by an applied electric field.
Each electron promoted to the conductance band leaves an empty spot in the valence band, due to the nature of the covalent bond, this vacancy is localized in energy and in space. The total charge

surrounding an unperturbed atom is zero, so an electron leaves the atom, a fixed positive charge is left behind. Under the influence of an electric field, electrons engaged in adjacent bonds can be displaced to occupy the empty position. This movement can be represented either considering the displacement of electrons (negative charges) or the displacement of the empty, positively charged, locations. The empty spots are called **holes**, and they carry a positive charge whose value is the absolute value of the electron charge.
Actually, the charge transport in semiconductors involve electrons at two energy intervals. Those in the conduction band and those in the valence band, to distinguish between them it is convenient to introduce the concept of holes and to treat the transport of electrons in the valence band as the transport of complementary positively charged particles.
Obvioulsy the concept of holes is valid only in the semiconductor. When the semiconductor is contacted by metal electrodes (as always in practice) the holes are either annihilated or created at the contact (the correct terms are "recombined" and "generated") by the electrons of the metal. This ensures that only electrons circulate in the metallic wire while in the semiconductor the current can be due to both electrons and holes.

The effective mass and the free electron approximation

Electrons in crystals are subjected to periodic potentials whose consequence is the dispersion relation between the energy and the momentum. Thus, when we study the behavior of an electron inside a crystal, for instance the motion of an electron under an applied electric field, it is necessary to include in the equation of motion also the internal periodic potentials.
On the other hand, the shape of the dispersion relation $E(k)$ at the bottom of the conduction band and at the top of the valence band is very close to the parabolic behavior that is typical of free electrons. A rescale of the properties of the electron could then allow to treat the electrons in the crystal as free particles.
This approximation is implemented through the concept of effective mass (m^*). This is a very convenient way to embed the potential that keeps the electron in the crystal into the amount of mass and then to apply to the electron the equation of motion of the free particle. This gives rise to an abstract entity (a quasi-particle) with charge q and mass m^*. Such a quasi-particle is the charge carrier in the semiconductor and we will continue to call it electron. The effective mass of the electrons in the crystal is obviously different from the rest mass of actual electrons ($m_0 = 9 \cdot 10^{-31} Kg$).
To calculate the effective mass is necessary to consider that in quantum mechanics particles are described by waves characterized by wavelength (λ), angular frequency ($\omega = 2\pi f$), and wavenumber ($k = 2\pi/\lambda$). Angular frequency and wavenumber are proportional to the energy and the momentum respectively ($E = \hbar\omega \ ; \ p = \hbar k$).
The velocity of propagation of a pure sinusoid is the phase velocity (v_{ph}). Considering the relationship between wavelength and frequency, the phase velocity is:

$$\lambda = \frac{v}{f} \rightarrow \frac{2\pi}{k} = v\frac{2\pi}{\omega} \rightarrow v_{ph} = \frac{\omega}{k} \tag{2.8}$$

Sinusoids are analytical functions existing from $t = -\infty$ to $t = +\infty$ and moving everywhere in space. Of course, a real particle can be observed only for a limited amount of time when it is confined into a limited amount of space (e.g. the solid). According to the Fourier transform theorem waves limited in space and in time correspond to a distribution of frequencies and wavenumbers. Thus, instead of a single pure wave, real particles correspond to a superposition of waves that is

called a wave packet. The velocity of propagation of the wave packet is the group velocity (v_g) that is defined as

$$v_g = \frac{d\omega}{dk} \tag{2.9}$$

The mass of the particle, defined by a dispersion relation $E(k)$, can be calculated from the laws of dynamics. The force acting on the particle is

$$F = \frac{dp}{dt} = \hbar \frac{dk}{dt}$$

and the force is the mass times the acceleration:

$$ma = m\frac{dv}{dt} = m\frac{d}{dt}\frac{d\omega}{dk}$$

then, reminding that $d\omega = \frac{dE}{\hbar}$ we can write:

$$\hbar \frac{dk}{dt} = m\frac{d}{dt}\frac{d\omega}{dk} = \frac{m}{\hbar}\frac{d}{dt}\frac{dE}{dk}$$

multiplying the last expression for $\frac{dk}{dk}$ we get:

$$\hbar \frac{dk}{dt} = \frac{m}{\hbar}\frac{d}{dt}\frac{dE}{dk}\frac{dk}{dk} = \frac{m}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt}$$

from which the definition of the effective mass is obtained:

$$m^* = \frac{\hbar^2}{\frac{d^2E}{dk^2}} \tag{2.10}$$

The effective mass is inversely proportional to the second derivative of the dispersion relation, namely is proportional to the inverse of the curvature of the band. As previously discussed, the bottom of the conduction band and the top of the valence band have a parabolic shape, then $E \propto k^2$ and in this situation the effective mass is constant.
The concept of effective mass is applied both to the electrons and the holes.

Table 2.1. Effective mass of electrons and holes for typical semiconductors

semiconductor	electrons effective mass	holes effective mass
silicon	0.26	0.38
germanium	0.12	0.3
gallium arsenide	0.068	0.5

Usually, the effective mass is smaller than the rest mass, this indicates that the electrons inside the crystal offer less inertia respect to free electrons. The different effective mass of electrons and holes is a consequence of the separated conditions of motion. It is important to remind that holes are actually electrons whose motion looks like a series of leaps from one atom to another. It is interestingly to observe that the concept of holes is a consequence of the definition of the effective mass.

Indeed, since the actual shapes of conductance and valence bands are characterized by opposite curvatures (see fig. 2.8), the effective mass is positive in the conduction band and negative in the valence band. The excitation of electrons in the conduction bands leaves empty states in the valence band, and then charge motion is possible but, since the effective mass is negative, the motion occurs in the opposite direction respect to the electrons in the conduction band. The physical absurdity of a negative mass is removed introducing a positive charge for the mobile particle in the valence band.

2.2.4 The band diagram

The band diagram, even in its simplified form, is the fundamental tool to interpret the electric properties of materials and the behaviour of electronic devices.
The diagram fixes the relative position of the conduction band, the valence band, and the Fermi level. Since the condition E=0 is not accessible, to define the energy values is necessary to introduce a reference value that can be actually observed. A convenient value is the potential energy of a free electron placed immediately outside the material. This level is the *vacuum level* and it corresponds to the surface potential of a given solid material. When more solids are kept in contact, for instance in junctions, the difference of surface potentials corresponds to the built-in potential.
In case of semiconductors, the band diagram can be drawn considering three fundamental experimental quantities: affinity, energy gap, and work function. Instead for a metal, since there is no band gap, the only meaningful quantity is the work function. Figure 11 shows the typical band diagram of semiconductors and metals.

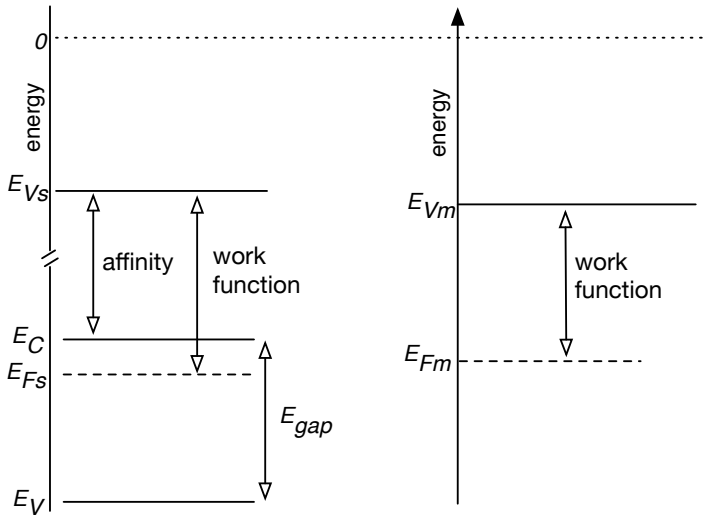


Fig. 2.11. *Simplified band diagram of a semiconductor and a metal. The vacuum level of individual materials are different respect to the absolute energy ladder. Note that in a semiconductor, the affinity and the work function are about 4 times the energy gap, for sake of simplicity the diagrams are usually plotted with the energy gap out of scale. The interrupted energy axis is introduced to mean the differences of scales.*

Electrons affinity

The electrons affinity ($q\chi$) is a material property corresponding to the largest energy necessary to displace an electron from the vacuum level to the inside of the material. In the band diagram $q\chi$ it is the distance between the vacuum level and the conduction band. UNder the assumption that the concentration of electrons in the conductance band is much smaller than the density of available states there are always available states at E_C, thus the affinity does not depend on the density of electrons.
The affinity can be measured with a variety of sophisticated experimental techniques. Inverse photoemission is one of them. In this technique electrons are delivered at very low kinetic energy towards the surface. Due to the low kinetic energy, electrons are absorbed in highest energy levels close to the vacuum level. From this leve they may decade towards the lowest allowable state lying at the bottom of the conduction band. The transition can occur via a number of intermediate steps or with a single step. Each transition may correspond to the emission of a photon. In case of a single transition, the photon with the largest energy is emitted, this energy is approximately equivalent to the distance between the vacuum level and the bottom of the conductance band namely to the affinity.

Table 2.2. Electron affinities of typical semiconductors

Semiconductor	Affinity
silicon	4.05 eV
germanium	4.00 eV
gallium arsenide	4.07 eV

Energy gap

The energy gap is the difference between the bottom of the conduction band and the top of the valence band. The energy gap corresponds to the energy necessary to displace an electron from the valence band to the conduction band. This energy can be provided by an external source such as a photon that releases its energy to an electron of the valence band. When the energy of the photon is larger or equal to the energy gap an increase of conductivity is observed. This phenomenon is called photoconductivity. In case of silicon the energy gap is about 1.1 eV which corresponds ($E_{gap} = hc/\lambda$) to a photon with a wavelength of 1.11 μm. Photons with a wavelength shorter than 1.1 μm can elicit the photoconductivity. In particular, photons in the visible range ($\lambda = 400 - 700\ nm$) excite the photoconductivity in silicon, and this property led to the development of digital cameras.
It is worth to note that the energy gap is slightly dependent on the temperature according to the following equation.

$$E_{gap} = E_{gap_0} - \frac{\alpha T^2}{T + \beta} \tag{2.11}$$

For silicon, $E_{gap_0} = 1.166\ eV$, $\alpha = 0.473\ meV/K$ and $\beta = 636\ K$.

Table 2.3. Energy gaps at room temperature of typical semiconductors

Semiconductor	Energy gap
silicon	1.12 eV
germanium	0.67 eV
gallium arsenide	1.42 eV

Work function

The work function corresponds to the energy necessary to relocate an electron from the inside of a material up to the vacuum level with null kinetic energy.
As shown in figure 11, the work function fixes the position of the Fermi level respect to the vacuum level. The relationship between the Fermi level and the work function can be obtained from the following thermodynamical considerations.

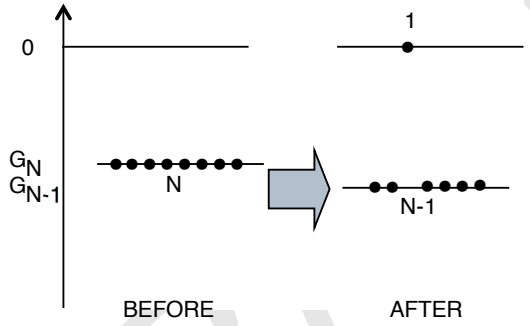


Fig. 2.12. *Before the extraction of one particle, the gas has a free energy G_N after the extraction the number of particles decreases of one unit.*

Let us suppose to remove one particle from a gas of N particles.The total energy of the particles before and after the extractions are: $E_{before} = G_N$ and $E_{after} = G_{N-1} + E_{vac}$
where G_N and G_{N-1} are the Gibbs free energies of a gas of N and N-1 particles calculated at constant pressure, temperature, and volume.
The work function ($q\phi$) is defined as the change of the energy necessary for the extraction process:

$$q\Phi = E_{after} - E_{before} = E_{vac} + G_{N-1} - GN = E_{vac} - \frac{G_N - G_{N-1}}{N - (N-1)} = E_{vac} - \frac{\partial \eta}{\partial N} \tag{2.12}$$

where η is the chemical potential at constant pressure, temperature, and volume. If the particle is charged, the electric potential has to be added and the chemical potential is replaced by the electrochemical potential:

$$\eta = \eta_0 - qV \tag{2.13}$$

Then the work function is the difference between the vacuum level and the electrochemical potential.
The electrochemical potential describes the collective energy of an ensemble of non-interacting charged particles. For an ensemble of electrons, the electrochemical potential is replaced by the Fermi level. This change is made necessary by the Pauli principle and the concept of collective energy is replaced by the energy level whose probability of occurrence is $\frac{1}{2}$. In a later section some additional arguments about the relationship between the Fermi level and the electrochemical potential will be provided.
The work function is a statistical quantity. Hence, even if a single electron could be extracted from any level, the average energy of extraction is the difference between the vacuum level and the Fermi level. More interestingly is the observation that in semiconductors the Fermi level lies in the energy gap, then there are not electrons at the Fermi level. However, since the probability to find electrons at the Fermi level is $\frac{1}{2}$ the average extraction energy is still the work function. Strictly speaking the definition in semiconductors does not hold at 0 K, but any applied energy rises the temperature above the absolute zero and then the definition is always valid.
There are several experimental methods to measure the work function, among them it is worth to mention those base on the thermionic effect, discussed in the introduction, and the photoelectric effect. This latter consists in measuring the current of electrons released from a material shined by a radiation of variable wavelength.
Since the Fermi level depends on the density of quasi-free electrons in the material, the value of the work function is a material constant only in the case of metals. In semiconductors the concentration of electrons can be varied with a technological procedure called *doping* and then even the work function is a variable depending on doping.
When the Fermi level lies in the band gap, the value of the work function is in the range between $q\chi$ and $q\chi + E_{gap}$. In table 3 the work function of some metals used in microelectronics technology is given.

Table 2.4. Typical work function values for some metals. The work function depends on the surface electronic states that can be arranged in a variety of structures, then in some cases an interval rather than a single value is found.

Metal	Work function
Silver	4.26 - 4.74 eV
Gold	4.95 - 5.47 eV
Copper	4.53 - 5.10 eV
Titanium	4.33 eV
Aluminum	4.06 - 4.26 eV

2.3 The statistics of electrons and holes

The concentrations of electrons and holes in conduction and valence bands are ruled by the Fermi-Dirac function. Being the holes a lack of electrons, the probability of finding a hole at a certain

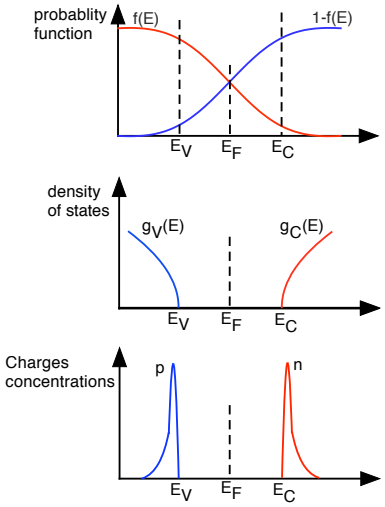


Fig. 2.13. *Probability functions, density of states, and resulting concentrations of electrons and holes. At room temperature electrons and holes energies corresponds to the bottom of the conduction band and the top of the valence band respectively.*

energy is complementary to the probability of finding at the same energy an electron. Then the distribution function for holes is $1 - f_{FD}$. Eventually the concentrations of electrons and holes at the energy E inside the respective bands is:

$$n = \int_{E_C}^{E} g_c(E) \cdot f_{FD}(E) dE$$

$$p = \int_{E}^{E_V} g_v(E) \cdot (1 - f_{FD}(E)) dE$$

The total concentration of electrons in the conduction band is:

$$n = \int_{E_C}^{\infty} g_c(E) \cdot f_{FD}(E) dE \tag{2.14}$$

In normal conditions, the Fermi level lies in the band gap, which is much larger than 26 meV, then $E - E_F \ll kT$ is likely to be valid, and the Fermi-Dirac function can be replaced by its first order approximation:

$$n = \int_{E_C}^{\infty} exp\left(-\frac{E - E_F}{kT}\right) \cdot g_c(E) dE \tag{2.15}$$

the numerator of the argument of the exponential can be split in two parts: $E - E_F = (E_C - E_F) + (E - E_C)$ then:

$$n = exp\left(-\frac{E_C - E_F}{kT}\right) \int_{E_C}^{\infty} exp\left(-\frac{E - E_C}{kT}\right) \cdot g_c(E) dE \tag{2.16}$$

The integral now provides a constant value which is independent on the Fermi level. This is indicated with N_C which corresponds to the total density of states in the conduction band. Note that due to the fast decay of the exponential function, the infinite can be conveniently used as the upper limit of integration.
The density of states is a fundamental quantity of the material, in the next section a description of the density of states calculations is provided.

2.3.1 The density of states

The density of states is the density per unit of volume and per unit of energy of the solutions of the Schrodinger's equation. For our scope, the semiconductor can be modeled as a infinite potential well applied to free particles (conveniently called electrons) with charge q and mass m*. In practice, once the internal potentials are included in the effective mass, the electrons in a solid corresponds to the particle-in-a-box model. The macroscopic shape of the material does not affect the density of the states. so, for sake of simplicity, let us consider a cube whose side is L.
The free electrons conditions means that the potential inside the well is null ($V(x) = 0$). Then the solution of the Schrodinger's equation can be written as a superposition of sine and cosine functions

$$\Psi = A \cdot sin(k_x x) + B \cdot cos(k_x x) \tag{2.17}$$

As discussed above, the boundary conditions are fixed by the potential well, thus $\Psi = 0$ at the borders of the well: $x = 0$ and $x = L$. As a consequence B=0 and the possibile values of k_x are:

$$k_x = \frac{n\pi}{L}, \quad n = 1, 2, 3, ... \tag{2.18}$$

The previous analysis has to be repeated for the other two dimensions (y and z). Thus, each solution corresponds to a cube in the k-space of volume π/L.
The total number of solutions characterized by positive values of k_x, k_y, and k_z and a modulus k of the wavevector is calculated considering one eighth of the volume of a sphere of radius k divided by the volume of the single solution (π/L).

$$N = 2\frac{1}{8}\left(\frac{L}{\pi}\right)^3 \frac{4}{3}\pi k^3 \tag{2.19}$$

The factor 2 gives account of the fact that each solution accommodates two electrons (two opposite spins). The density per energy unit is obtained using the chain rule for the derivative:

$$\frac{dN}{dE} = \frac{dN}{dk}\frac{dk}{dE} = \left(\frac{L}{\pi}\right)^3 \pi k^2 \frac{dk}{dE} \tag{2.20}$$

For the dispersion relation of the free particle ($E = \frac{\hbar^2 k^2}{2m*}$) we obtain:

$$\frac{dk}{dE} = \frac{m^*}{\hbar^2 k} \quad ; \quad k = \frac{\sqrt{2m^* E}}{\hbar} \tag{2.21}$$

then the density of states per unit of energy, for $E > 0$, is:

$$g(E) = \frac{1}{L^3}\frac{dN}{dE} = \frac{8\pi\sqrt{2}}{h^3} m*^{3/2} \sqrt{E} \tag{2.22}$$

The density of states is zero at the bottom of the well and even for negative values.
For the electrons in the conductance band the minimum of energy corresponds to the bottom of the conductance band (E_C). Then the density of the states in the conduction band has to be written scaling the energy to the bottom of the conduction band

$$g(E) = \frac{8\pi\sqrt{2}}{h^3} m*^{3/2} \sqrt{E - E_C} \tag{2.23}$$

An analog calculation leads to the density of states for the holes in the valence band.

2.3.2 The concentration of electrons and holes

Replacing the density of states in eq. 2.16 the concentration of the electrons in the conductance band is obtained:

$$n = N_C \cdot exp\left(-\frac{E_C - E_F}{kT}\right) \tag{2.24}$$

where N_C is:

$$N_C = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} \tag{2.25}$$

The same calculus repeated for the holes provides the concentration of holes in the valence band:

$$p = N_V \cdot exp\left(-\frac{E_F - E_V}{kT}\right) \tag{2.26}$$

where N_V is:

$$N_V = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2} \tag{2.27}$$

N_C and N_V are expected to be different because the effective masses of electrons and holes are different . The effective masses previously introduced have been calculated considering the dynamics properties of electrons and holes. The value of the effective mass to be used to calculate the density of states is different. The values for silicon are $m_n^* = 1.08$; $m_p^* = 0.81$.
N_C and N_V in silicon and at room temperature are:

$$N_C \approx 2.8 \cdot 10^{19}\ cm^{-3}\ ;\ N_V \approx 1.04 \cdot 10^{19}\ cm^{-3} \tag{2.28}$$

These values are almost similar. It is important to remark that all these quantities depends on temperature.
Equations 2.24 and 2.26 are among the fundamental tools to study the behavior of semiconductors and their junctions.

2.3.3 The intrinsic Fermi level

Equations 2.24 and 2.26 connect the density of holes and electrons with the distance of the conduction and valence bands from the Fermi level. Semiconductors are defined by a Fermi level that lies in the band gap, then using equations 2.24 and 2.26 it is possible to calculate the Fermi level in the case of an intrinsic semiconductor.
The intrinsic semiconductor is a pure material where electrons and holes are only generated by the ionization of the atoms of the semiconductor. Thus, for each electron in the conduction band there is a hole in the valence band.
Thus, the definition of an intrinsic semiconductor is $n_i = p_i$ where the subscript i indicates the intrinsic condition:

$$N_C \cdot exp\left(-\frac{E_C - E_{Fi}}{kT}\right) = N_V \cdot exp\left(-\frac{E_{Fi} - E_V}{kT}\right)$$

$$exp\left(-\frac{E_C - E_{Fi}}{kT} + \frac{E_{Fi} - E_V}{kT}\right) = \frac{N_V}{N_C}$$

$$exp\left(-\frac{2E_{Fi} - E_C + E_V}{kT}\right) = \frac{N_V}{N_C}$$

From which the intrinsic Fermi level (E_{fi})is calculated:

$$E_{Fi} = \frac{E_C + E_V}{2} + \frac{KT}{2} ln\left(\frac{N_V}{N_C}\right) \tag{2.29}$$

The first term is the center of the band gap, while the second term depends on the ratio of the effective masses. In silicon and at room temperature (T=300 K) and this quantity is approximately -13 meV.
Thus, with an error of 13 meV we can conclude that the Fermi level in intrinsic silicon lies at the centre of the band gap. This conclusion applies to most of the semiconductors. Consequently, the work function of the intrinsic silicon is

$$q\Phi_i = q\chi + \frac{E_C + E_V}{2} = 4.61\ eV \tag{2.30}$$

2.3.4 Doping

The structure of real crystals is far to be perfect. Rather, real crystals are characterized by a number of defects whose existence is fundamental for the properties of semiconductors and for their use in electronics.
The most important defects are impurities and vacancies. Impurities are involved in the processes of charge transport, while vacancies offer an important technological feature for semiconductors.
A vacancy is a location of the crystal characterized by a missing atom in the lattice. Around a vacancy, the distances between atoms are altered, the lattice is deformed and, as a consequence

of the greater distance between the adjacent atoms, the binding forces are weaker. An important characteristics of vacancies is that they can be filled by an impurity atom added by purpose.
This opportunity is exploited in a process called *doping* that is aimed at altering the balance between electrons and holes and at increasing the conductivity. A pristine semiconductor is doped through the implantation of impurity atoms. this operation can take place through a physical impact implantation of accelerated ions, followed by a thermal diffusion of the impurities from the surface towards the internal region of the material.
Since, defects are usually uniformly distributed, also the dopant impurities will be uniformly distributed at least after a short distance from the surface.
There are two important categories of impurities: those that favor the increase of the concentration of electrons and those that increase the concentration of holes. In silicon these conditions are fulfilled by pentavalent and trivalent atoms respectively.

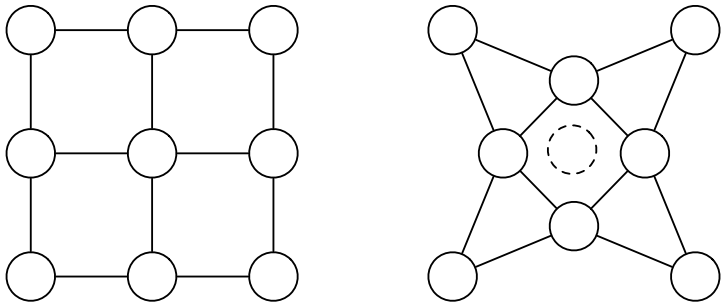


Fig. 2.14. *In a perfect crystal (left) the atoms are arranged in a regular lattice and all the interactions are of the same magnitude. In case of a vacancy (right) the atoms around the empty position are displaced and the interatomic interactions are less intense.*

In silicon, pentavalent atoms (e.g. phosphorous, arsenic, and other elements of the V group of the periodic table) are called N-type dopants or **donors**. When a pentavalent atom replaces the position of a silicon atom in the crystal only four of the five available valence electrons are engaged in a covalent bond with an adjacent silicon atom. The fifth electron remains bonded to its own atom but at an energy level rather close to the conduction band of the crystal. In practice, the distance between the bottom of the conduction band and the energy level of the idle electron is about 10% the energy gap. Due to the thermal energy, a conspicuous portion of these electrons leaves the phosphorous and populates the conduction band. However, even if the energy difference is small, according to the Boltzmann probability function, only a fraction of the phosphorous levels are actually transferred into the conduction band.
The loss of one electron changes the total charge around the phosphorous atom that instead of being neutral becomes positively charged. This positive charge is fixed and it is not displaced by an applied electric field.
Given a density N_D of donors the statistics allows to calculate the percentage of effectively ionized donors. Let E_D the energy level of donor states and let us consider that both $E_C - E_F$ and $E_D - E_F$ are larger than kT, thus the Fermi-Dirac equation can be replaced by the Boltzmann equation. The concentration of electrons in the conductance band is dominated by the ionized

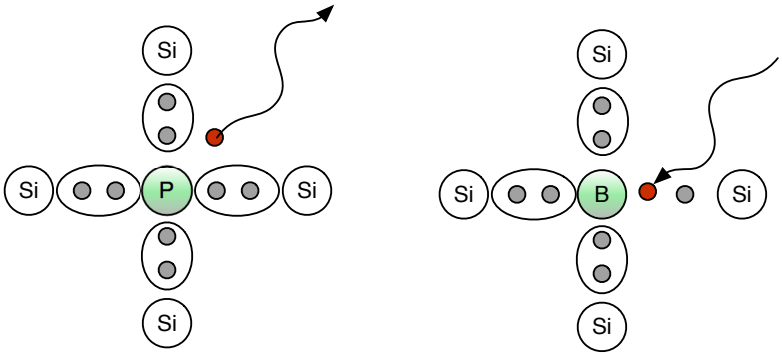


Fig. 2.15. *Pictorial representation of the doping effect of pentavalent (left) and trivalent (right) atoms. In case of pentavalent one of the electrons of the dopant atom is not engaged in a covalent bond and can be promoted to the conduction band. In case of a trivalent doping, one of the adjacent silicon remains with an unpaired electron that could be fulfilled by an electron from a nearby silicon atom giving rise to a hole.*

donors $n = N_C exp(-\frac{E_C - E_F}{kT})$. In the same way, the concentration of electrons occupying the N_D donor levels (n_D) is $n_D = N_D exp(-\frac{E_P - E_F}{kT})$. Of course $N_D = n + n_D$, then the percentage of ionized donors is

$$\frac{n}{n + n_D} = \frac{N_C exp(-\frac{E_C - E_F}{kT})}{N_C exp(-\frac{E_C - E_F}{kT}) + N_D exp(-\frac{E_P - E_F}{kT})} \tag{2.31}$$

Dividing by $exp(\frac{-E_F}{kT})$ we get the following expression:

$$\frac{n}{n + n_D} = \frac{1}{1 + \frac{N_D}{N_C} exp(\frac{E_C - E_D}{kT})} \tag{2.32}$$

The above equation is qualitatively correct, and actually some small correction term due to the degeneracy of donor levels should be introduced. However, the energy level E_D depends on the nature of the dopant atom. In any case, the fraction of ionized donors depends on the doping concentration; it decreases as the doping concentration increases. The relationship between percentage of ionized donors and concentration of donors is shown in fig. 16 in the case of phosphorous in silicon where the energy level of the donor state occurs at 0.045 eV below the bottom of the conduction band.

At room temperature, for a doping of 10^{18} cm^{-3} about 71 % of donors are actually ionized. However, since the exact number of donors is unknown and the order of magnitude of ionized donors is equal to the order of magnitude of the total number of donors, it is customary to assume that N_D donors give rise to N_D electrons in the conduction band. These are equilibrium average values; in practice, donors continuously lose and acquire an electron, and the average quantity of ionized atoms is given by equation 2.32.

A opposite behavior is obtained using trivalent atoms (e.g. boron, aluminum, and other elements of the group III of the periodic table). Trivalent atoms in silicon are P-type dopants or **acceptors**. A boron atom that replace a silicon in the crystal leaves one of the adjacent silicon atoms with an

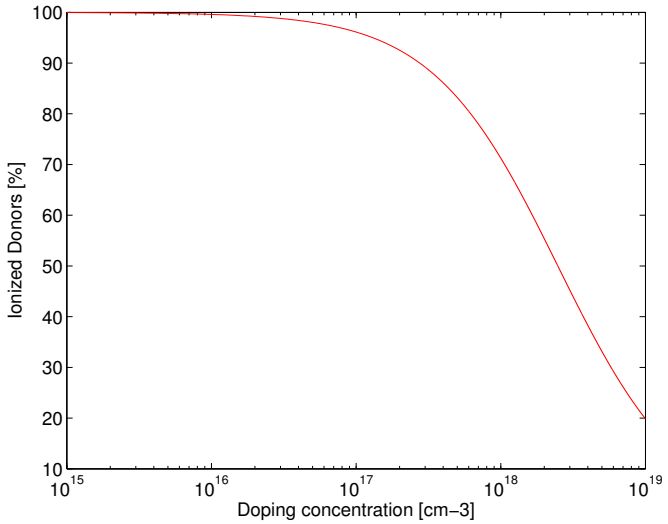


Fig. 2.16. *Percentage of ionized donors as a function of the doping concentration. The calculations are related to phosphorous in silicon at room temperature and a factor 2 is introduced in eq. 2.30 to take into account the degeneracy factor of the donor level. Note that as the doping concentration increases the distance between the conduction band and the Fermi level decreases and the Boltzmann approximation tends to be less valid.*

unpaired electron. Since the unpaired electron is not engaged in a covalent bond, its energy level is slightly higher than the top of the valence band. Namely the electron is less bound to its atom. The magnitude of the distance is comparable with that observed between pentavalent electrons and conduction band. It is useful to remind that the energy of paired electrons is lower (more negative) than unpaired electrons, and the minimization of this energy leads to the stability of the chemical bonds.

The statistics of donors applies also to acceptor states, then then plot in figure 16 is valid also for P-type doping. As a consequence of the distribution of electrons in the acceptor states, the total charge around the boron atom instead of being neutral becomes negative. This negative charge is fixed and cannot be moved by an applied electric field.

The electron that moves from an adjacent silicon atom to fulfill the octet of a silicon adjacent to the boron leaves a hole that can be occupied by another electron. The hole moves through the crystal, while the boron remains negatively charged.

The statistics of donors is valid also for acceptors, then we consider that N_A acceptors give rise to N_A holes in the valence band.

It is important, to note that in both cases the creation of a free mobile charge is not compensated by the creation of a mobile charge of opposite sign (as it happen in intrinsic semiconductors), but the countercharge is a fixed non mobile charge. Eventually, doping results in an increase of one of the two charge carriers and, to respect the neutrality of the material, in a concentration of fixed charges.

Table 2.5. Charges produced by doping

type	mobile charges	fixed charges
N type	electrons (negative)	donors (positive)
P type	holes (positive)	acceptors (negative)

The laws ruling the statistics of electrons and semiconductors are still valid for a doped material. Then, in order to accommodate the increase of electrons or holes concentrations with equations 2.24 and 2.26 it is necessary that the Fermi level changes its value.
Then if n increases, the distance between the conduction band and the Fermi level has to decrease and on the other hand, if p increases, the distance between the Fermi level and the valence band decreases. Eventually, the position of the Fermi level defines the doping. It has been shown in the previous section that the Fermi level of a intrinsic semiconductor is approximately at the center of the band gap, for N type materials the Fermi level lies close to the conduction band, and on the contrary if the Fermi level is close to the valence band the material is P type.
Typical concentrations of dopants in silicon are in the range $10^{15} - 10^{18} cm^{-3}$. This value has to be compared with the concentration of silicon atoms that is about $10^{23} cm^{-3}$. Then the doping is of the order of one atom of impurity each 10 millions of silicon atoms. Strikingly, this tiny quantity is sufficient to change the electric characteristics, but on the other hand, it leaves untouched the other parameters of the material, such as the energy gap, the electron affinity, the density and so on.
At such level of concentration the dopant atoms are sparse in the material. Thus, there is not interactions among the dopant atoms, and their atomic orbitals do not degenerate in bands. For this reason, the dopant energy level mentioned before can still be considered as a single energy level rather than a band.

Mass action law

According to equations 1.24 and 1.26, the Fermi level defines the concentrations of electrons and holes. In a doped material the concentration of only one of the two species increases, and since the material is in equilibrium this cannot leave unaffected the other.
The product of the concentrations of the two species is ruled by the action mass law. For an intrinsic semiconductor this product is $n_i p_i = n_i^2$. In general, using equations 17 and 18 this product can be calculated under any condition.

$$np = N_C exp\left(-\frac{E_C - E_F}{kT}\right) \cdot N_V exp\left(-\frac{E_F - E_V}{kT}\right) = N_C N_V exp\left(\frac{-E_C + E_F - E_F + E_V}{kT}\right) \tag{2.33}$$

Thus, the product does not depend on the Fermi level, namely it does not depend on doping.

$$np = n_i^2 = N_C N_V exp\left(-\frac{E_{gap}}{kT}\right) \tag{2.34}$$

The product np is maintained constant under any kind of doping and it depends, besides than on the temperature, on the energy gap.
In silicon, at T=300 K, $n_i^2 \approx 1.45 \cdot 10^{20} cm^{-6}$. Then the intrinsic concentration of charge carriers is

$n = p = \sqrt{n_i^2} \approx 10^{10} cm^{-3}$.
In case of doping, let us suppose N-type, according to the mass action law the concentrations are:

$$n = N_D \ ; \ p = \frac{N_D}{n_i^2} \tag{2.35}$$

Then, if $N_D = 10^{17} cm^{-3}$ then $n = 10^{17} cm^{-3}$ and $p = 10^{20}/10^{10} = 10^3 cm^{-3}$. This great inequality of concentrations is a striking consequence of doping, whose major effect is to discriminate the charge carriers into majority and minority charges.

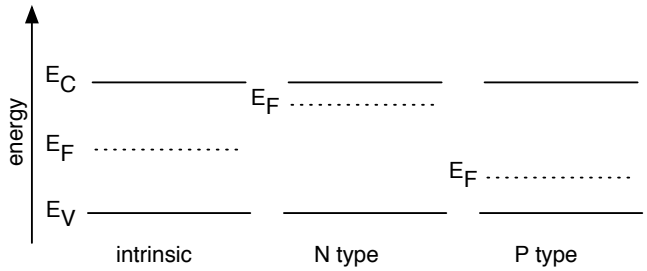


Fig. 2.17. *The relative position of the Fermi level with respect to the conduction and valence bands energies signals the kind of doping of the semiconductor.*

The concentration of doping determines the position of the Fermi level, and then the work function. In a N-type semiconductor we can write the concentration of electrons with respect to the intrinsic concentration :

$$n = N_D = N_C \ \exp\left(-\frac{E_C - E_F}{kT}\right) = N_C \ \exp\left(-\frac{E_C - E_i - E_F + E_i}{kT}\right) \tag{2.36}$$

$$n = N_D = N_C \ \exp\left(-\frac{E_C - E_i}{kT}\right) \exp\left(\frac{E_f - E_i}{kT}\right) \tag{2.37}$$

Where E_i is the Fermi level of the intrinsic semiconductor.

$$N_D = n_i exp\left(\frac{E_f - E_i}{kT}\right) \tag{2.38}$$

Thus the concentration of electrons depends on the distance between the Fermi level and the intrinsic Fermi level:

$$E_F = E_i + kT \ln(\frac{N_D}{n_i}) \tag{2.39}$$

if $N_D = 10^{17} cm^{-3}$ the distance between the Fermi level and its intrinsic value is 0.40 meV.
Then, the work function of the doped semiconductor is:

$$q\Phi = E_{vac} - E_F = E_{vac} - E_i - kT \ln\left(\frac{N_D}{n_i}\right) = q\Phi_i - kTln\left(\frac{N_D}{n_i}\right) \tag{2.40}$$

where $q\Phi_i$ is the work function of the intrinsic semiconductor. For silicon: $q\Phi_i = 4.61eV$. Similar equations hold in case of P type doping replacing $E_C - E_F$ with $E_F - E_V$.
Finally, it is important to remind the hypothesis on which these calculations have been performed. In particular they are based on the approximation of the Fermi-Dirac function with the Boltzmann probability function. This holds if the distance between the conduction band and the Fermi level is of the order of 2 - 3 times kT. When the doping is large, the Fermi level lies too close to the conduction band and the approximation is no more valid. In silicon, this happens when N_D and N_A are greater than $10^{19}cm^{-3}$. Beyond this value, the Fermi level invades the bands and the semiconductor assumes a metal-like character. Such a semiconductor is said degenerate.

2.4 Charge transport: the drift-diffusion model

In the first part of this chapter the properties of electrons and holes at the equilibrium have been illustrated. The equilibrium condition is important but, of course, we are interested to describe the relations between the currents and the voltages that are manifested when the electronic system is out of equilibrium. The material at the equilibrium lies the condition to describe the non equilibrium behaviour, which is actually a perturbation, sometimes small, of the equilibrium state.
In this section the charge transport phenomena in the framework of a classic approach to electric current are described. In this context the current flowing in a semiconductor is the sum of two components: the drift current and the diffusion current. This approach is when the dimensions of the material are larger than few nanometers, a more precise requirement about the validity of the classical model is provided later.
To study the motion of the charges is necessary to consider the forces acting on the charges, namely the electric force $F = q\mathcal{E}$, where $\mathcal{E}$ is the electric field, and their effects on the acceleration of the charge $F = ma$.
The relationship between the electric field and the current depends on two properties of the particle: the charge and the mass. Inside of a semiconductor we find two mobile particles: electrons and holes. The charge of electrons and holes is the elementary charge ($q = 1.6 \cdot 10^{-19}\ C$) that is positive for holes and negative for electrons, and the masses are the effective masses previously introduced.
The electric current is the amount of charge flowing across a section of the conductor per time unit ($I = Q/T$). It is more convenient to consider the density of current that is defined independently from the section. In this way, a mono dimensional description of the devices is possible.

$$j = \frac{Q}{TA} \quad \left[\frac{C}{s \cdot m^2}\right] \tag{2.41}$$

The current (measured in Amperes) is obtained from the density of current simply multiplying the density of current for the area of the section.
The density of current is the macroscopic manifestation of the movement of individual charges. The connection between the current and the velocity of the charges can be obtained considering the instantaneous work done by the applied electric field on the charge: $dL = Fdx = q\mathcal{E}dx$. Replacing the field with the voltage ($\mathcal{E} = \frac{V}{w}$) the work is $dL = q\frac{V}{w}dx$. Where w is the length of the conductor. At the equilibrium, this work is equal to energy dissipated by the current itself Pdt where P is the electric power ($P = VI$). From this equality the definition of current is obtained:

$$q\frac{V}{w}dx = iVdt \rightarrow i = \frac{q}{w}\frac{dx}{dt} = \frac{qv}{w} \tag{2.42}$$

The above calculation is derived from the Ramo's theorem which establish the relationship between the microscopic motion of a charge and the current observed in an external circuit.
The above definition describes the current produced by a single particle; real currents are due to the motion of a density of particles. Then the moving charge is replaced by $qnAw$ namely the product of the density of the charges (nq) times the volume of the conductor.
In a semiconductor, where two charge carriers exist, the total current is the sum of the currents of electrons and holes:

$$j_n = qnv_n \ ; \ j_p = qpv_p \tag{2.43}$$

where q is the elementary charge, n and p are the densities of electrons and holes, and v_n and v_p are the velocities of electrons and holes respectively.

2.4.1 Thermal velocity

The temperature of a gas of non interacting particles is proportional to the average kinetic energy of the particles. This implies that particles are kept in motion even in absence of an external force. The equipartition theorem assigns a kinetic energy equal to $kT/2$ to each degree of freedom. The quantity proportional h the temperature is the kinetic energy, namely the square of the velocity. Thus, the thermal motion is isotropic, and the average position of the particles does not change with time. To this regard it is important to remind that the velocity is a vector whose average value can be zero even if the speed (the magnitude of velocity) is different from zero.

$$\frac{1}{2}m_n^* v_{th}^2 = \frac{3}{2}kT \tag{2.44}$$

In silicon, the effective mass of electrons is $m_n^* = 0.26 \cdot m_0$, and then the thermal velocity at room temperature (T=300 K) is of the order of $10^7 \ \frac{cm}{s}$.

2.4.2 Drift current

A net displacement of charges is achieved by the application of a voltage drop across the material. The applied voltage changes the potential energy of the electrons inside the material. The potential energy corresponds to the energy of the bottom of the conduction band (for electrons) and the top of the valence band (for holes). Then in presence of an applied voltage, the band diagram is altered.

According to quantum mechanics, an electron kept in movement in a perfect periodic potential should not undergo any scattering process. In this condition, the electron does not lose the acquired kinetic energy and the velocity instead of reaching a stable value grows to a maximum value that ultimately depends on the length of the conductor. Such a motion, without scatters and energy dissipation, is said ballistic.
Actually, in real materials the atoms are not arranged in perfect regular lattices, alterations in positions and nature of the atoms occur and they are collectively called lattice defects. The major sources of defects are the impurities (doping among them) and the vacancies.

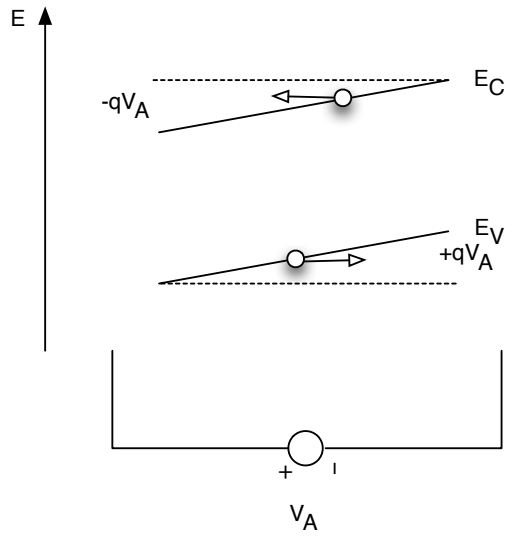


Fig. 2.18. *Applied voltage alters the potential energy of electrons ($\Delta E_C = -qV_A$) and holes ($\Delta E_V = +qV_A$). Electrons and holes moves due to the electric field. The energy above (or below for holes) the bands are the acquired kinetic energies.*

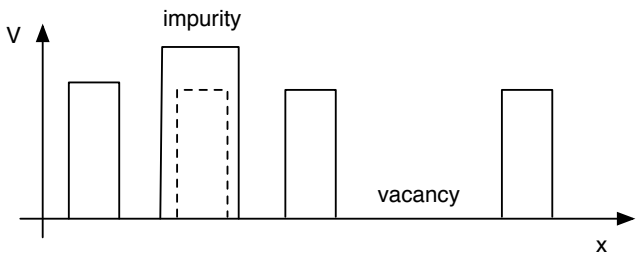


Fig. 2.19. *Impurities are different atoms, then their potential is different from that of the natural atom of the lattice. Vacancies are missing atoms that can be considered as a lack of potential pulse.*

Furthermore, due to thermal motion, even the atoms fluctuate around their equilibrium positions. The vibration of atoms around their equilibrium positions gives rise to collective modes of oscillation that are treated as quasi-particles called *phonons* endowed with proper energy, momentum, and dispersion relations.
Defects in the atoms arrangement alters the profile of the potential breaking the perfect periodicity. Eventually, defects of any nature play the role of scattering centers. Electrons can scatter with these centers losing whole or part of the kinetic energy acquired during the motion. Scatter events and the consequent transfer of energy from the electrons to the atoms are at the origin of the Joule effect.

In order to derive a simple, but effective, model of voltage-current relationship, let us assume that in each scatter event, the electron loses all the kinetic energy and the momentum acquired from the acceleration produced by the electric field. In this way after each scatter, the electron starts again to acquire energy.
Let us introduce τ_c : the average time between to consecutive scatters. This is also known as relaxation time. Then, the average momentum acquired between two consecutive scatters is:

$$p = F \cdot \tau_c \rightarrow m_n^* v_d = -q\mathcal{E}\tau_c \tag{2.45}$$

v_d is the *drift velocity*, which is the average speed of displacement of electrons subject to a electric field $\mathcal{E}$.

$$v_d = -\frac{q\tau_c}{m_n^*}\mathcal{E} \tag{2.46}$$

The previous relation establishes a proportion between the electric field and the average velocity. This is the microscopic version of the Ohm's law. Noteworthy, the drift velocity is independent on the size of the material. Note that as the scatter probability approaches zero, τ_c becomes infinite and the drift velocity diverges to infinite.
An important, complementary quantity of the average time between scatter is the free mean path (l_c) that is defined as the average distance travelled between two consecutive scatters.

$$l_c = \tau_c v_d \tag{2.47}$$

Note that the above model is valid only if the electrons reach the electrodes after a large number of scatters. When the path is too short, $l \leq l_c$ the scatters do not occur and the charge transport is ballistic. All the above considerations are still valid in case of holes.
The quantity reassuming the characteristics of the motion of electrons and holes is the **mobility** that is the proportion between the drift velocity and the electric field.

$$\mu_n = \frac{q\tau_c}{m_n^*} \; ; \; \mu_p = \frac{q\tau_c}{m_p^*} \tag{2.48}$$

The drift velocities have to be written considering the sign of the charge of the particles as:

$$v_d = -\mu_n\mathcal{E} \; ; \; v_d = +\mu_p\mathcal{E} \tag{2.49}$$

The mobility depends on the nature of the material and on its purity. This is particularly important for semiconductors where mobility depends on the density of defects. Among the defects, the doping atoms are particularly important because they are added with a purpose.
The mobilities of electrons and holes at room temperature for intrinsic silicon are of the order of:

$$\mu_n = 1400 \; \frac{cm^2}{Vs} \; ; \; \mu_p = 500 \; \frac{cm^2}{Vs} \tag{2.50}$$

From the mobility, which can be experimentally determined, it is possible to estimate the relaxation time. For instance in the case of electrons in silicon:

$$\tau = \frac{m_n^* \mu_n}{q} = \frac{1400 \cdot 10^{-4} \cdot 0.26 \cdot 9.1 \cdot 10^{-31}}{1.6 \cdot 10^{-19}} \approx 2 \cdot 10^{-13} \; s \tag{2.51}$$

Eventually, the drift currents due to electrons and holes are

$$J_n = -qn(-v_d) = qn\mu_n\mathcal{E} \; ; \; J_p = +qp(+v_d) = qp\mu_p\mathcal{E} \tag{2.52}$$

and the total drift current is

$$J = J_n + J_p = -q(n\mu_n + p\mu_p)\mathcal{E} \tag{2.53}$$

The quantity $-q(n\mu_n + p\mu_p)$ is the total conductivity (σ) of the semiconductor, so that the Ohm's law is synthetically written as $j = \sigma\mathcal{E}$. Note that $\sigma = 1/\rho$. The measured current in a piece of a material of length w and section A is:

$$I = J \cdot A = \sigma \cdot A\frac{V}{w} \rightarrow V = \frac{1}{\sigma}\frac{w}{A}I \tag{2.54}$$

that is the usual definition of the electric resistance assumed at the beginning of this chapter.

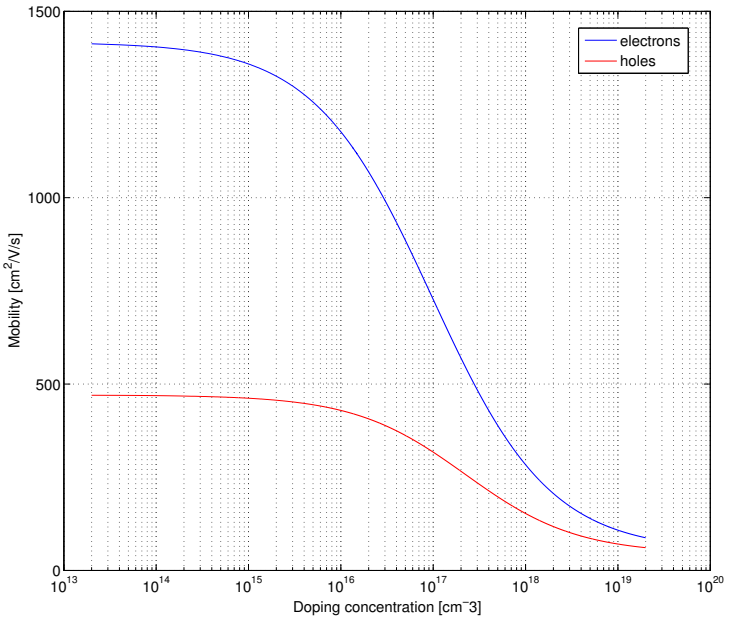


Fig. 2.20. *Mobilities of electrons and holes in silicon as a function of dopant atoms concentrations. Electrons mobility is calculated for arsenic and phosphorous impurities and the mobility of holes in case of boron dopant.*

Velocity saturation

At large electric fields the mobility tends to deviate from its constant value. This because at large electric field some additional mechanisms intervene to limit the velocity. In silicon, the most important of these mechanisms is the increase of the scattering probability with the atoms of the lattice.

These phenomena become important when the drift velocity becomes larger than the thermal velocity. These electrons are called hot because their temperature (according to the equipartition theorem) is larger than the background temperature. Hot electrons, among the other properties, can activate the scatter with phonons described by a different dispersion relation (optical phonons). The details of these interactions are outside the scope of this textbook, here it is important to keep in mind that as the electric field increases the mobility decreases and the velocity reaches a saturation value. The observable consequence is the limitation of the current.
The saturation velocity in silicon is about $10^7 \frac{cm}{s}$ for electrons, and it is reached for a saturation electric field $\mathcal{E}_{sat} = 10^4 \frac{V}{cm}$. Saturation may become important when voltage is applied across short distances, for instance, at the distance of $1\mu m$ the saturation field is obtained with voltage drop of only 10 mV.
In semiconductors characterized by a different relative position of the conductance and valence band the relationship between drift velocity and electric field is more complex and it gives rise to peculiar behaviors. These will be discussed in a later chapter.

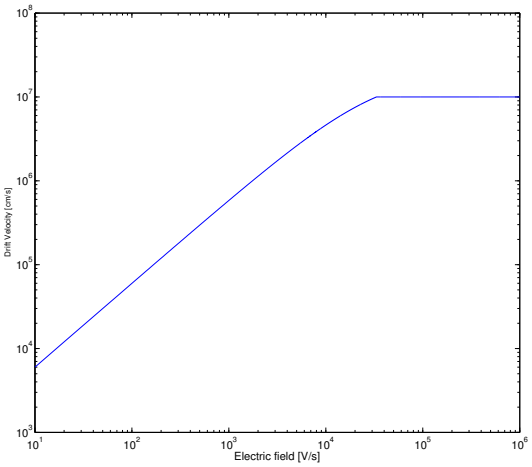


Fig. 2.21. *Approximated behavior of electrons drift velocity vs. electric field in silicon.*

2.4.3 Diffusion current

Contrarily to metals, semiconductors can maintain an internal non homogeneous distributions of charge carriers. Due to the thermal motion, particles that are non homogeneously distributed tend to equate their density. This process is called diffusion and it is observed for any mobile set of particles such as the gas molecules in atmosphere.
In case of charged particles this process gives rise to a current that is called diffusion current. Thus, in a semiconductor it is possible to observe a current even without an applied voltage. The energy necessary to the motion is thermal but the magnitude of the current is proportional to the gradient of the charges density.

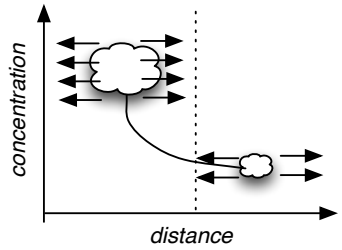


Fig. 2.22. *Thermal flow is proportional to the concentration of particles. In case of a gradient, the flow impinging onto a surface from the most populated region is larger than the flow coming from the less populated side. Note that the thermal flow is absolutely isotropic and the diffusion current is observed in the line of separation between regions of different concentrations.*

A gradient of concentration of particles n gives rise to a flux of particles. The relationship between the flux and the gradient is the first law of Fick:

$$F = -D \cdot \frac{dn}{dx} \tag{2.55}$$

Where D is the diffusion coefficient. The negative signs indicates that the direction of the flow is opposite to the direction of growth of the concentration as shown in figure 22.
The diffusion coefficient has the dimension of cm^2/s. In case of charged particles, such as electrons and holes, the electric current associated to the flow is $J = q \cdot F$, Electrons and holes are characterized by different diffusion coefficients: D_n and D_p. Thus, the diffusion current of electrons and holes is:

$$J_n = (-q) \cdot \left(-D_n \cdot \frac{dn}{dx}\right) = q \cdot D_n \cdot \frac{dn}{dx}; \quad J_p = (+q) \cdot \left(-D_p \cdot \frac{dn}{dx}\right) = -q \cdot D_p \cdot \frac{dn}{dx}; \tag{2.56}$$

Eventually, the total current of electrons and holes is given by the sum of the drift and the diffusion current

$$J_n = qn\mu_n\mathcal{E} + qD_n\frac{dn}{dx} \; ; \; J_p = qp\mu_p\mathcal{E} - qD_p\frac{dp}{dx} \tag{2.57}$$

The diffusion coefficient describes the motion of the charges under the influence of the gradient concentration. This quantity is similar to the mobility that describes the motion of the charges under the influence of the electric field. In order to calculate the relationship between D and μ let us consider the case of equilibrium of the current of electrons.
Due to the drift-diffusion model the equilibrium is not merely the absence of current, but rather a situation where the diffusion current is compensated by a drift current and vice versa. Hence, in a semiconductor the equilibrium can be achieved also when electric field and a gradient of concentration of charges are simultaneously present.
The equilibrium condition is:

$$J_{diff} = J_{drift} \rightarrow qn\mu_n\mathcal{E} = -qD_n\frac{dn}{dx} \tag{2.58}$$

Let us replace $\mathcal{E} = -dV/dx$ and let us calculate the mobility considering that being the voltage a function of x, also the concentration is function of the position.

$$-n(x)\mu_n \frac{dV}{dx} = -D_n \frac{dn}{dx} \rightarrow \mu = \frac{D_n}{n(x)} \frac{dn}{dV} \tag{2.59}$$

In order to calculate dn/dV, let us consider that any potential difference is added to the conduction energy band. Thus the concentration n(x) can be written as:

$$n(x) = N_C exp\left(-\frac{E_c - qV(x) - E_f}{kT}\right) = N_C exp\left(-\frac{E_c - E_f}{kT}\right) exp\left(\frac{qV}{kT}\right) = n_0 \cdot exp\left(\frac{qV}{kT}\right) \tag{2.60}$$

Where n_0 is the concentration of electrons calculated where the voltage is null. Then, the mobility is

$$\mu = \frac{D_n}{n(x)} \frac{dn}{dV} = \frac{D_n}{n(x)} \frac{q}{kT} n(x) = \frac{q}{kT} D_n \rightarrow D_n = \frac{kT}{q} \mu_n \tag{2.61}$$

A similar expression can be found for the holes diffusion coefficient. The above relationship is called *Einstein-Smoluchowski equation.*
The diffusion coefficient is the product of the mobility and a quantity that corresponds to the voltage equivalent temperature (V_T). At room temperature ($T = 330\ K$) the thermal voltage is about 26 mV. in practice, the temperature provides the energy for the diffusion current whose magnitude is determined by the gradient of concentration and the mobility.

Total current, electrochemical potential and Fermi level

Drift and diffusion currents seems to be generated by two different phenomena. Drift current originates from an electric field, namely the gradient of the applied voltage, while the diffusion current is due to a gradient of charges density.
The Einstein-Smoluchowski equation suggests the existence of a strong relationship between the two currents, and actually, they can be derived from the gradient of a unique quantity that is called electrochemical potential.
Indeed, since $\mathcal{E} = -\frac{dV}{dx}$ and $D_n = \frac{kT}{q}\mu_n$ the total current of electrons is proportional to the gradient of a unique potential function:

$$J_n = -qn\mu_n \frac{dV}{dx} + q\frac{kT}{q}\mu_n \frac{dn}{dx} = n\mu_n \left[-q\frac{dV}{dx} + \frac{kT}{n}\frac{dn}{dx}\right] = n\mu_n \frac{d\eta}{dx} \tag{2.62}$$

where η is called the electrochemical potential, defined as:

$$\eta = \eta_0 + kT\ln(n) - qV \tag{2.63}$$

It is easy to show that the above relationship also applies to the Fermi level. Indeed, from eq. 2.24

$$E_F = E_c + kT\ln(n) - kT\ln(N_c) \tag{2.64}$$

The first term is the potential energy equivalent to $-qV$, and the last term is a constant equivalent to η_0.

Then in the Boltzmann approximation ($E_C - E_F \ll kT$) the Fermi level and the electrochemical potential of a gas of charged particles are exactly coincident. Then as the gradient of the electrochemical potential determines the current of particles, the gradient of the Fermi level determines the current of charges.
As a consequence, the equilibrium condition where the sum of all the currents is zero is achieved when the gradient of the Fermi level is zero, namely when the Fermi level is constant throughout the whole material.

2.5 The non uniform distribution of dopant atoms and the built-in potential

Contrarily to metals the electric field inside a semiconductor can be steadily different from zero. Thus, it is possible to maintain a non uniform distribution of fixed and mobile charges. The uneven distribution of charges is maintained by an internal potential that is called *built-in potential*.
To study this situation, let us consider a semiconductor with a continuos distribution of dopant atoms extended from a region where donors dominate to a region where acceptors are the majority. Let us consider an ideal experiment where the distribution of dopant atoms is instantaneously created. At the time t_0 the concentrations of electrons and holes is $n = N_d$ and $p = N_a$ and the electric field is zero everywhere.

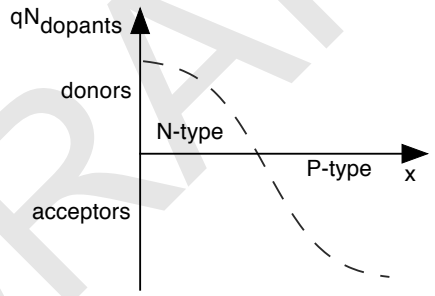


Fig. 2.23. *Non uniform distribution of dopant atoms. Two kinds of dopants are used in this example, so the material is partially N-type and partially P-type*

In terms of band diagram, the above described situation is shown in figure 23. Since the concentrations of electrons and holes is not constant a diffusion current emerges. However, as the mobile charges move, the total charge locally changes giving rise to an electric field that prompts a drift current that compensate the diffusion current.
The system evolves towards the equilibrium that corresponds to a constant Fermi level. This condition is obtained imposing a curvature to the other energy levels: the vacuum level, the conduction and the valence band. Since the conduction band and the valence bands are the potential energies of electrons and holes, the curvature of the bands corresponds to the internal potential energy.
To evaluate this energy, let us consider that at each coordinate x, the concentration of electrons and holes is given by:

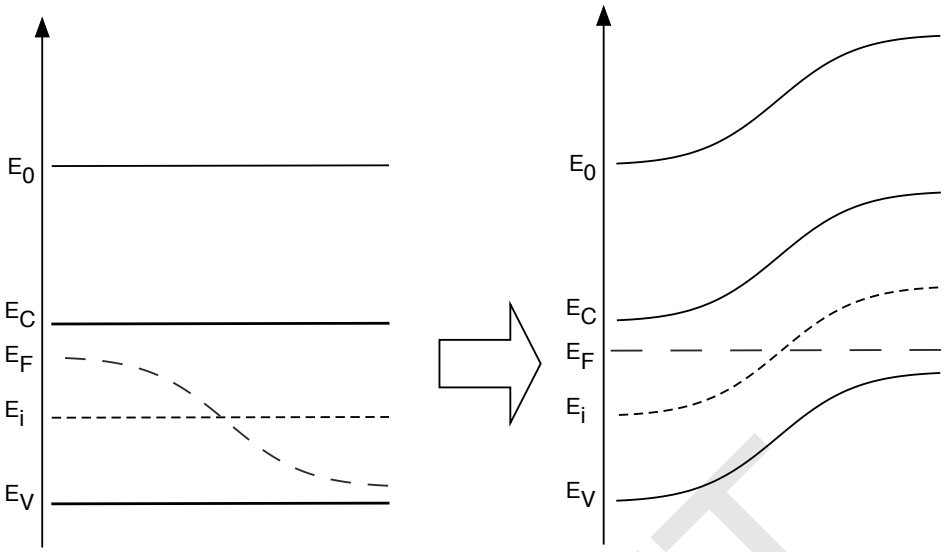


Fig. 2.24. *The dopant distribution of fig. 21 is depicted in the band diagram. Left: the band diagram immediately after the doping. The Fermi level is not constant and its distance from the bands depends on the doping concentration. Right: at the equilibrium the Fermi level is constant but all the bands are curved. The intrinsic Fermi level is plotted to allow for a rapid identification of the kind of total doping.*

$$n(x) = N_C exp\left(-\frac{E_C(x) - E_F}{kT}\right) \; ; \; p(x) = N_V exp\left(-\frac{E_F - E_V(x)}{kT}\right) \tag{2.65}$$

Let us measure all the energies with respect to the intrinsic Fermi level whose distance from the bands is constant.

$$E_C - E_F = (E_C - E_i) - (E_F - E_i) \tag{2.66}$$

Then the potential of the electrons can be written as:

$$\phi = -\frac{1}{q}(E_C - E_F) = -\frac{1}{q}(E_C - E_i) + \frac{1}{q}(E_F - E_i) = \phi_0 + \phi_i \tag{2.67}$$

Since the potential can be defined with respect to any constant, we can consider ϕ_i, the distance between the Fermi level and the intrinsic level, as the potential of the electrons.
ϕ_i counteracts the concentration gradient. Then the drift current is equal and opposite to the diffusion current and the system is in equilibrium. The sign of the potential ϕ_i defines the kind of doping, if positive the material is N-type and if negative the material is P-type.

At the equilibrium, the total current is zero, this condition determines the relationship between the built-in potential and the concentration of electrons.

$$J_n = qn\mu_n\mathcal{E} + qD_n\frac{dn}{dx} = 0 \tag{2.68}$$

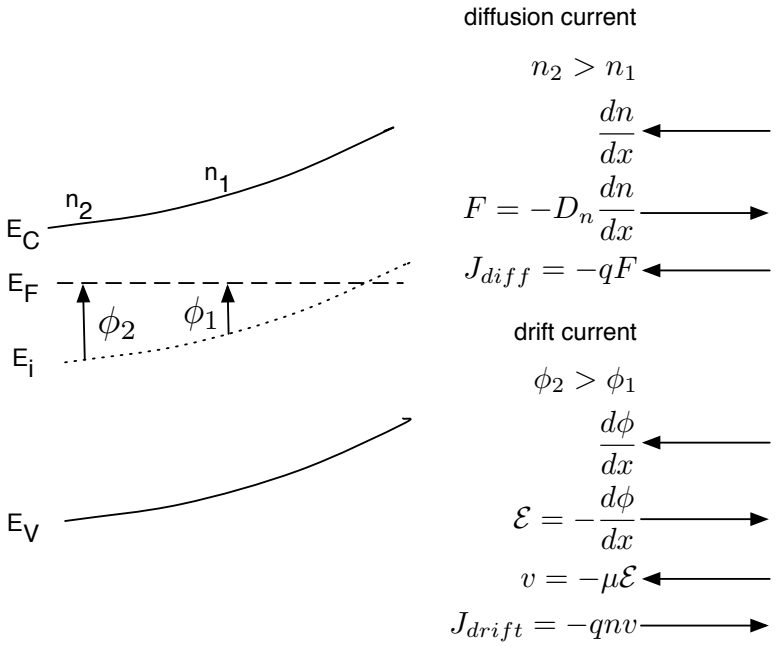


Fig. 2.25. *Figure shows an example of potential ϕ_i and the related diffusion and drift currents. The direction of the vector of all the quantities composing the two currents are shown. Eventually, drift and diffusion currents are opposite and the defined potential can keep in equilibrium the system. All the considerations holds for holes.*

from which:

$$\mathcal{E} = -\frac{d\phi}{dx} = -\frac{D_n}{\mu_n}\frac{1}{n}\frac{dn}{dx} = -\frac{kT}{q}\frac{1}{n}\frac{dn}{dx} \tag{2.69}$$

where the Einstein relation was used.
The potential can be calculated integrating the previous equation from the point 1 to the point 2.

$$\int_{\phi_1}^{\phi_2} d\phi = \int_{n_1}^{n_2} \frac{kT}{q}\frac{1}{n}\frac{dn}{dx}dx \tag{2.70}$$

from which the built-in potential is calculated:

$$\phi_2 - \phi_1 = \frac{kT}{q} ln\left(\frac{n_2}{n_1}\right) \tag{2.71}$$

The previous equation allows to calculate the potential across a perturbed region. The analytical behavior of the potential ($\phi(x)$) is calculated from the **Poisson equation** that relates the charges distribution with the electric potential.

$$\frac{d\phi^2}{dx^2} = -\frac{\rho(x)}{\epsilon_s} \tag{2.72}$$

The density of charge ρ is contributed by the four kinds of charges that are found in a semiconductor:

$$\rho = p - n + N_d - N_a \tag{2.73}$$

where p and n are the mobile charges and N_d and N_a are the densities of donors and acceptors. The concentration of electrons is:

$$n = N_C exp\left(-\frac{E_C - E_F}{kT}\right) = N_C exp\left(-\frac{E_C - E_i}{kT}\right) exp\left(\frac{E_F - E_i}{kT}\right) = n_i exp\left(\frac{q\phi_i}{kT}\right) \tag{2.74}$$

a similar expression is found for the holes

$$p = n_i exp\left(-\frac{q\phi_i}{kT}\right) \tag{2.75}$$

Then the Poisson equation can be rewritten as:

$$\frac{d\phi^2}{dx^2} = \frac{q}{\epsilon_s}\left[n_i exp\left(\frac{q\phi_i}{kT}\right) - n_i exp\left(-\frac{q\phi_i}{kT}\right) - N_d + N_a\right] \tag{2.76}$$

introducing the hyperbolic sine ($sinh(x) = (e^x - e^{-x})/2$):

$$\frac{d\phi^2}{dx^2} = \frac{q}{\epsilon_s}[2n_i sinh(\frac{q\phi_i}{kT}) - N_d + N_a] \tag{2.77}$$

In order to solve the above equation it is necessary to know the distribution of donors and acceptors. In this textbook, the Poisson equation is solved only under simple assumptions.
Charge distribution, electric field, and built-in potential are fundamental quantities to characterize the properties of junctions between materials. They will be thoroughly calculated, even if in ideal conditions, for all the junctions studied in this textbook.

Quasi-neutrality condition

Given the initial situation in fig. 23, the equilibrium is reached with a negligible displacement of charges. In practice, at the equilibrium the amount of electrons and holes is still given by the concentration of donors and acceptors.
This assumption may be justified through a numerical example.
Let us consider a N-type silicon ($\epsilon_s = 11.7$) where the concentration of donors changes from 10^{16} to $10^{18} cm^{-3}$ at a distance of $0.5\mu m$. The built-in potential generated by this gradient of concentration is:

$$\Delta\phi = \frac{kT}{q} ln(\frac{n_2}{n_1}) = 0.026\ ln\frac{10^{18}}{10^{16}} \approx 0.12\ V \tag{2.78}$$

The electric field is

$$\Delta\mathcal{E} = \frac{\Delta\phi}{\Delta x} = \frac{0.12}{0.5\ 10^{-4}} = 0.2\ 10^4\ \frac{V}{cm} \tag{2.79}$$

The derivative of the electric field can be calculated from the Poisson equation considering the charges in a N-type semiconductor where $N_a = 0$ and $p = \frac{n_i^2}{N_D}$ is negligible.

$$\frac{d\phi^2}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{q}{\epsilon_s}[n - N_d] \tag{2.80}$$

from which

$$n - N_d = \frac{d\mathcal{E}}{dx}\frac{\epsilon_s}{q} \tag{2.81}$$

Considering the finite differences, the difference between mobile electrons and fixed donors is:

$$n - N_d = \frac{\Delta\mathcal{E}}{\Delta x}\frac{\epsilon_s}{q} = \frac{0.2\ 10^4}{0.5\ 10^{-4}}\frac{11.7\ 8.8\ 10^{-14}}{1.6\ 10^{-19}} \approx 10^{14} cm^{-3} \tag{2.82}$$

Then, $n = N_d - 10^{14}$ that is negligible with respect to N_d. Therefore, the local concentration of electrons is always equal to the local concentration of donors, and the equilibrium condition is reached moving just a negligible amount of mobile charges.

2.6 The Chua formalism of electric network elements

Electronics is interested to study the relationship between $v(t)$ and $i(t)$ as they occur in pure materials and in their combinations. The relationship between voltage and current is complex, and it is usual in electric network theory to decompose such a relation in three elements that put into evidence three different phenomena occurring in the matter when voltage and current are considered. Voltage and current are actually the observable macroscopic quantities of two other quantities that, in some sense are more fundamental than them: the electric charge and the flux of magnetic field. The relationship between observable and internal quantities is mediated by the operators of integral and derivative.
The relations between the four electric quantities (internal and observable) are conveniently represented in a diagram originally introduced by Chua.

The ideal elements connecting the four variables are:
Resistance: $v = f_R(i)$. In case of linearity $f_R = R$. This is the Ohm's law, but in general the relationship is non linear.
Capacitance: $v = f_C(q)$. Even in this case the linearity defines the standard capacitor element, but in devices non linear behaviors appears.
inductance: $\varphi = f_L(i)$. In the linear case $f_L = L$.
Eventually, Chua introduced, for symmetry reasons, a fourth element connecting charge and magnetic flow that he called memristor: $\varphi = f_M(q)$. It is easy to observe that in case of linearity a memristor is simply a resistor, but in case of non linearity it gives rise to a component whose conductivity depends on the amount of charge that have flown through the device, such as in electrochemical cell. In some sense this element preserves a memory of past events and, as its name suggests, it is a sort of a resistor with memory.

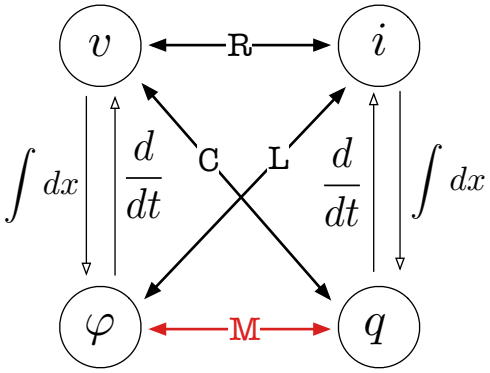


Fig. 2.26. *Relation between electric quantities and the properties of the matter.*

$$\varphi = f_M(q) \rightarrow \frac{d\varphi}{dt} = v = \frac{\partial f}{\partial q}\frac{\partial q}{\partial t} = M(q)i \tag{2.83}$$

In the following of the textbook the electronic devices will be described in terms of non linear resistors (I/V characteristics) and non linear capacitors(C/V characteristics) . It is clear that these quantities are abstraction of a more complex behavior similar to the case of mechanics where any system is described as a collection of combination of masses, springs, and damping elements.

3

The Metal-Semiconductor junction

3.1 Introduction

The junctions between materials are the building blocks of electronic devices. The properties of electrons and holes in their native materials strongly affect the properties of the junction.
The first junction that is considered involves a semiconductor and a metal. This is the basic structure necessary to connect any piece of semiconductor to an electric circuit. Under particular conditions the metal-semiconductor junction can give rise to a non linear relationship between the current and the voltage. Eventually, the metal-semiconductor junction may behave as either a rectifier or a ohmic contact.
The formation of a junction elicits a change of the properties of the charge carriers across the interface, these changes involve a region across the physical boundary between the materials. This region is the junction itself and its properties determines the behaviour of the whole device.
We will be interested to study the properties of junctions in two conditions: in thermal equilibrium and under an applied voltage.
Metal- semiconductor junctions serve also as a tool to define the general methods to study the junctions. Strictly speaking, a junction is the region of contact between materials. Ideally, this is a plane surface where one material abruptly ends and the other begins.
Another ideal approximation is to consider the semiconductor uniform and homogeneous until the surface. Actually, this is quite far from reality because the surface is a region of very large non homogeneities. Indeed, the regular crystal lattice inside the semiconductor (the bulk) is faded to be interrupted at the surface. Thus, the interactions among atoms close to the surface are different respect to those in the bulk. This leads to the formation of additional energy states for the electrons (surface states) that may also lie in the energy gap. In this chapter, we begin to treat the ideal junction neglecting the surface states. The effects of the surface states will be discussed at the end of the chapter.
The main assumption about a junction is that the mobile charges can move from one material to the other. It is worth to point out that this condition depends on the chemical bonds that tie together the two materials. The energy and the density of the states of the electrons in the two sides of the junction are obviously different; thus if electrons can cross the interface they tend to leave the states at largest energies to occupy any free allowable state at lower energy. Thus, currents from one material to the other and vice versa appear as the junction is formed. At the equilibrium the total current has to be zero.

Equilibrium does not mean that all currents are null but rather that their algebraic sum is zero. In practice, at the equilibrium, the junction is still crossed by charges in both directions, but the average of the total current at each coordinate of the system is null. We have seen in the previous chapter that, if the Fermi-Dirac function is approximated by the Boltzmann equation, the null current condition is provided by the constant Fermi level. In the the next section a more general extension of this property is given.

3.1.1 The general rule of junctions at the equilibrium

Let us consider a junction of two materials labeled as 1 and 2. The electrons in one materials may occupy free states in the other material and vice versa. The charge transfer from one material to the other occurs without the input of any external energy source but the temperature, that is uniform everywhere. It is important to remark that the uniformity of temperature is a general assumption valid, except where noted, throughout this textbook.
Let us call G the rate of transfer (electrons per second) of electrons from one material to the other. As a general rule, the rate of any transition between states can be written as the product of the density of filled states, from where the electrons move, times the density of the empty available states times the probability that the transfer can occur.
In other words, the displacement of electrons from one material to the other requires that the electrons move from a filled state to land into an empty state and this transition has to be physically possible. The first two terms are provided by the statistical laws but the third implies that the junction is permeable to the electrons.
To calculate the rate of transition of electrons across the junction let us consider $n(E)$: the density of filled states, $v(E)$: the density of empty states and k the transition probability.
The rate of transitions from material 1 to 2 is $G_{1\to2} = n_1(E) \cdot v_2(E) \cdot k$ and in the opposite direction it is $G_{2\to1} = n_2(E) \cdot v_1(E) \cdot k$. At the equilibrium:

$$G_{1\to2} = G_{2\to1} \to n_1(E)v_2(E) = n_2(E)v_1(E) \tag{3.1}$$

The density of filled and empty states are calculated from the statistics as $n(E) = g(E)f_D(E)$ and $v(E) = g(E)(1 - f_D(E))$ where f_D is the Fermi-Dirac function.
Since the electrons do not change energy, the energy can be omitted in the equations, and the equilibrium condition provides:

$$n_1v_2 = n_2v_1 \to g_1f_1g_2(1 - f_2) = g_2f_2g_1(1 - f_1) \to f_1 = f_2 \tag{3.2}$$

The equilibrium condition does not depend on the density of states, namely it does not depend on the nature of the materials but only on the Fermi-Dirac function. The equilibrium is achieved when the Fermi Dirac function is the same everywhere. In practice this means that the electrons are in equilibrium if the probability to find electrons at a given energy is the same everywhere. Since we assume uniform the temperature, the Fermi Dirac functions is constant only if the Fermi level is uniform.
Eventually, a junction at the equilibrium, and any material in general, requires that the Fermi level is constant everywhere.
Since the Fermi levels in the pristine materials, are generally different, a uniform Fermi level is achieved after a net transfer of charges from one material to the other. This rule holds for any kind of material (metal or semiconductor) kept in contact, namely where electrons can freely flow from one material to another.

3.2 The metal-semiconductor junction at the equilibrium

The behaviour of the metal-semiconductor junction depends on the relative magnitude of the work function of the metal ($q\Phi_m$) and the semiconductor ($q\Phi_s$) and the kind of doping (either N-type or P-type), This gives rise to four combinations: N-type and $q\Phi_m > q\Phi_s$, N-type and $q\Phi_m < q\Phi_s$, P-type and $q\Phi_m > q\Phi_s$, and P-type and $q\Phi_m < q\Phi_s$.
Let us firstly consider the case of a junction made of a N-type semiconductor and a metal such that $q\Phi_m > q\Phi_s$. As an example, chromium and N-type silicon.
The ideal junction is a useful model where the semiconductor is uniform until the surface (no surface states) and it is characterized by an isotropic distribution of dopant atoms (N_D). The surface of the semiconductor is perfectly planar and the metal grows in the direction parallel to the surface.
Most of the properties of the junction can be derived from the band diagrams. The only relevant quantity for metals is the work function that in case of chromium is about $q\Phi_{Au} = 4.60\ eV$. On the other hand, the band diagram of the semiconductor is characterized by three quantities: affinity, energy gap, and work function. For silicon we have: $q\chi_{Si} = 4.05\ eV$; $E_{gap_{Si}} = 1.12\ e$, however the work function, namely the position of the Fermi level, depends on the concentration of the doping according to eq. 2.40.

$$q\Phi = q\Phi_i - kTln\left(\frac{N_D}{n_i}\right) \tag{3.3}$$

where $q\Phi_i = q\chi + \frac{E_{gap}}{2} = 4.61\ eV$. In case of $N_D = 10^{16}\ cm^{-3}$ the work function is $q\Phi_{Si} = 4.25\ eV$. All these quantities allow to design the band diagrams of the two materials before the junction is formed.

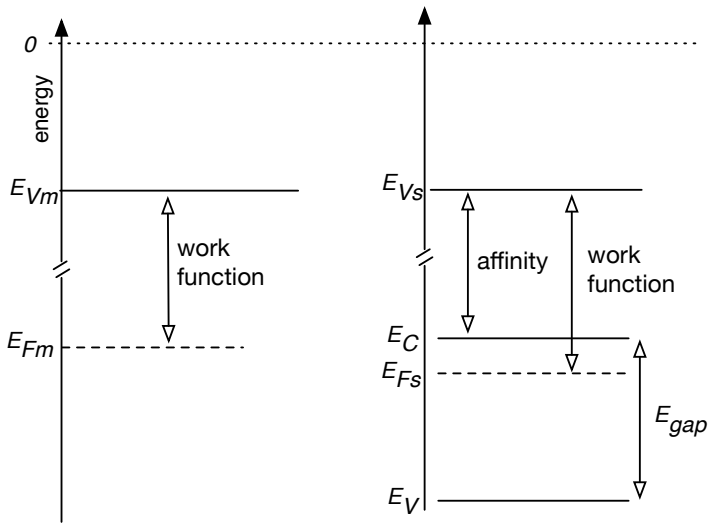


Fig. 3.1. *Band diagrams of chromium and N-type silicon before the formation of the junction. Note that the drawing is not in scale being the affinity almost 4 times larger than the energy gap.*

The band diagram at the equilibrium is achieved as a consequence of a displacement of electrons according to the mutual position of the Fermi levels in the two materials. In this example, the Fermi level of the semiconductor is higher, in energy, than the Fermi level of the metal, namely in terms of work function $q\Phi_m > q\Phi_s$. It is worth to remark that the displacement of electrons requires also the availability of states in the material of destination.

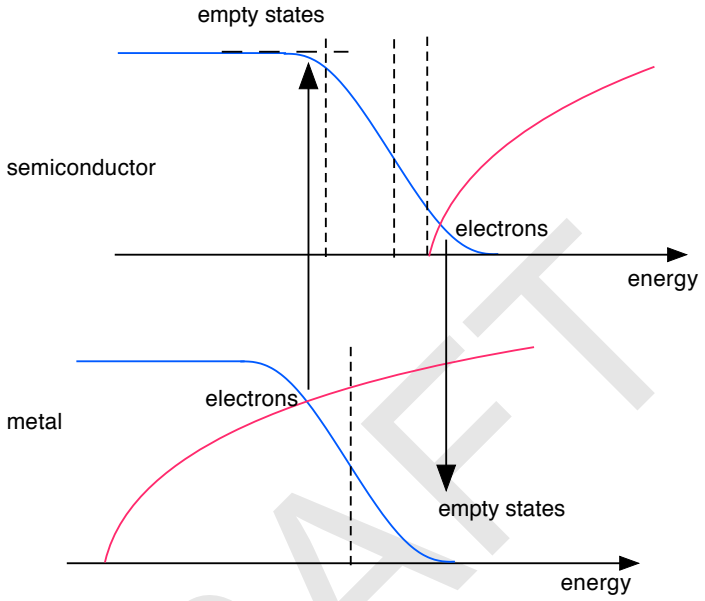


Fig. 3.2. *Fermi-Dirac function and density of states of semiconductor and metal.*

As shown in figure 3.2, electrons in the conductance band are less numerous with respect to those in the metal but they have access to a larger density of empty states. On the other hand, the transfer of the many electrons in the metal is hindered by the scarcity of available free states in the valence band of the semiconductor. The empty states in the valence band are the holes whose number is given by the action mass law $p = n_i^2/N_d$. In this numerical example the density of states available in the valence band is $p = 10^4\ cm^{-3}$.
Immediately after the formation of the junction, the current of electrons from the semiconductor to the metal is larger than the current flowing in the opposite direction. As the electrons leave the semiconductor, the density of holes increases and the current from the metal to the semiconductor increases. Eventually, the equilibrium (zero total current) is reached after that a net amount of charges are transferred from the semiconductor to the metal.
Since the electric field inside metals is null, the excess electrons coming from the semiconductor accumulate on the metal surface at the interface. On the other hand, the electrons that left the semiconductor leave behind a region where the total charge is negative since it is dominated by the fixed donor charges.
In this region close to the junction before the formation of the junction we find $n = N_d$ and when the junction reaches the equilibrium $n < N_D$. A volume of the material where the total charge is

different than zero is called a *space charge region*.
The alteration of the charge is limited to a region immediately close to the interface with the metal while the rest of the material (the bulk) is left unchanged. Note that the size of the region depends on the density of the dopant atoms.
The concentration of electrons and holes is still abiding the law of the statistics: in particular, the concentration depends on the difference between the Fermi level and the conduction band.

$$n = N_C \cdot exp\left(-\frac{E_C - E_F}{kT}\right) \tag{3.4}$$

In the space charge region the difference between the conduction band and the Fermi level changes to take into account the decrease of electrons density. The affinity and the energy gap being related to the intimate nature of the semiconductor do not change. Then due to the loss of electrons, in the region close to the interface the difference between the conduction band and the Fermi level increases. This band bending indicates that the region is depleted by electrons. For this reason, such a region is also called a *depletion layer*.
All these elements contribute to draw the equilibrium band diagram of the whole metal-semiconductor system at the thermal equilibrium. Note that the band diagram of insulated materials are floating because of the impossibility to measure the energy distance from E=0. In the junction, the constant Fermi level provides an anchor to draw the band diagram and then to establish the energy variation from one material to the other.
The drawing of equilibrium band diagram is an important tool to understand the properties of junctions. It may be easily accomplished, in any situation, following the steps listed in table 1

Table 3.1. Steps to draw the equilibrium band diagram

1	identify the interface and the space charge region in the semiconductor
2	draw a unique Fermi level
3	draw the unaltered band diagram of the metal and the bulk of the semiconductor
4	draw a continuous curve connecting the vacuum level of the metal and the vacuum level of the bulk
5	draw the conduction band parallel to the vacuum level (constant affinity)
6	draw the valence band parallel to the conduction band (constant energy gap)

Following these steps, the equilibrium band diagram shown in figure 3.3 is obtained.

The equilibrium between the currents is maintained by an energy barrier applied to the electrons of both the materials. Due to the original differences in work function and the different concentrations of electrons, the equilibrium requires two barriers with different heights: one is applied to the electrons of the metal ($q\phi_B$) and the other is applied to the electrons of the semiconductor (ϕ_i) where $q\phi_i$, is smaller than $q\phi_B$.
The barrier applied to the electrons of the semiconductor ($q\phi_i$) is the *built-in potential*, and it is equal to the difference of the work functions of the two materials.. It is also the energy difference between the vacuum energy level at the surfaces of the two materials. This quantity is a potential difference measured between the surfaces of the two materials. It is called *contact potential difference* or *Volta potential*, it is an observable manifestation of the internal potentials.
The amount and the behaviour of potential, electric field, and depletion layer size can be calculated solving the Poisson equation (eq. 2.72)

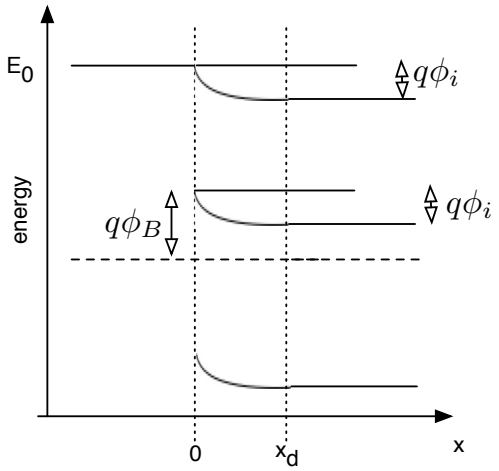


Fig. 3.3. *Equilibrium band diagram of metal-semiconductor system. The coordinate x=0 indicates the interface, and the coordinate x_d indicates the end of the perturbed region, this x_d is the length of the depletion layer.*

The solution of the Poisson equation requires the knowledge of the distribution of mobile and fixed charges. It has been mentioned above that one of the hypothesis of the ideal junction is the uniform doping of the semiconductor, as a consequence, in a non perturbed N-type semiconductor also the electrons are uniformly distributed and the total charge in any volume of the semiconductor is null. However, at the equilibrium in the semiconductor, in the region from $x = 0$ to $x = x_d$ the concentration of electrons is less than N_d and it is variable as showed by the bending of the conduction band. The concentration of electrons depends on the potential according to eq. 2.73.
The Poisson equation for a N-type semiconductor can be written as:

$$\frac{d\phi^2}{dx^2} = -\frac{q}{\epsilon_s}\left[N_d - N_d exp\left(\frac{q\phi}{kT}\right)\right] \tag{3.5}$$

The previous equation has to be solved in the depletion layer where the concentration of electrons is variable and then the derivative of the potential is different from zero.
The Poisson equation can be solved numerically, but a practical solution can be obtained making some assumption about the distribution of mobile charges. In particular we can assume that in the perturbed region the fixed charge of donors is much greater than the mobile charge of electrons ($N_D \gg n$). This condition is called *deep depletion* hypothesis.
The deep depletion hypothesis is justified by the fact that the concentration of electrons is an exponential function of $(E_C - E_F)$, then small changes in energy gives great changes in concentration. Of course in the neighbour of x_d the deep depletion approximation fails, a description of the behaviour around x_d more close to the reality is discussed in chapter 4.

The total distribution of charge is shown in figure 3.4. The negative charge is accumulated in a thin layer at the surface of the metal, and the positive charge, formed by donor charges, extends in the semiconductor. Obviously, the total charge in the whole device is always zero.

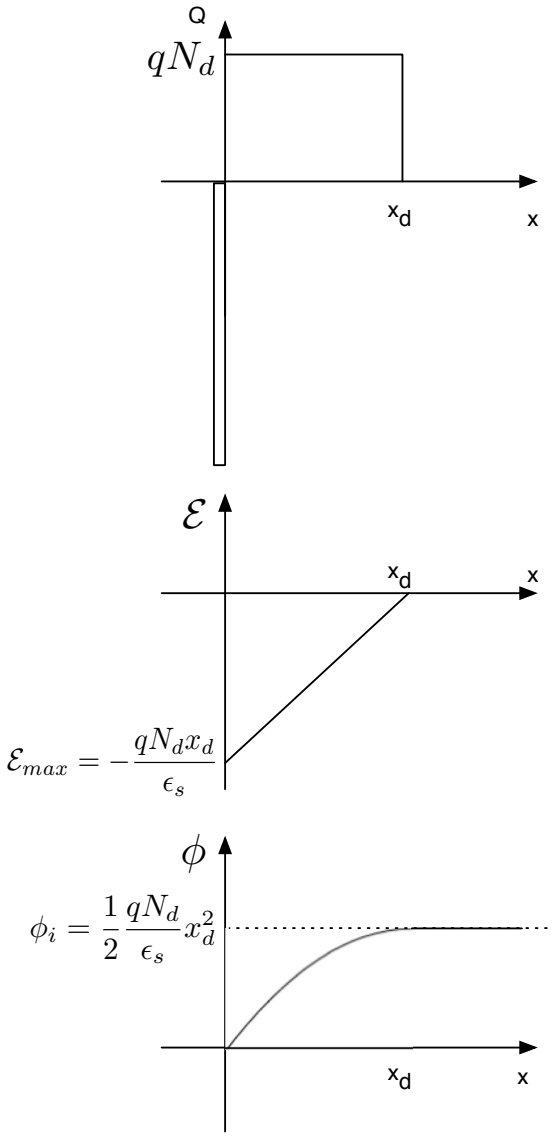


Fig. 3.4. *Charge distribution, electric field, and potential of a ideal metal-semiconductor junction at the equilibrium under the hypothesis of deep depletion and uniform doping.*

$$+Q = qN_d x_d;\ -Q = -qN_d x_d \tag{3.6}$$

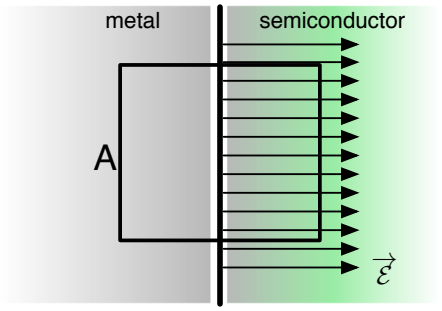
Note that these are charge densities in unit of $C \cdot cm^{-2}$.
The relationship between the charge density and the electric field is given by the Gauss's theorem:

$$\varphi = \oint_S \overrightarrow{\mathcal{E}} \cdot d\overrightarrow{A} \tag{3.7}$$

where φ is the electric flux, $\overrightarrow{\mathcal{E}}$ is the electric field and $d\overrightarrow{A}$ is a vector pointing orthogonal to the infinitesimal area element. The surface integral is extended to a closed surface S. Using the divergence theorem, the Gauss theorem can be written in differential form: $\nabla \cdot \overrightarrow{\mathcal{E}} = \frac{\rho}{\epsilon}$ that in one dimension is:

$$\frac{d\mathcal{E}}{dx} = \frac{\rho(x)}{\epsilon} \rightarrow \mathcal{E}(x) = \mathcal{E}_0 + \int_0^x \frac{\rho(x)}{\epsilon} dx \tag{3.8}$$

At $x < 0$, inside the metal the electric field is zero by definition.
The electric field at the interface ($x = 0$) is generated by a surface density of charge. This can be calculated in figure 3.5 by a simple application of the Gauss's theorem considering that the electric field generated by a uniformly distributed sheet of charges is orthogonal to the surface and the electric field towards the metal is null.



$$\oint_S \overrightarrow{\mathcal{E}} \cdot d\overrightarrow{A} = \mathcal{E}A = \frac{\sigma A}{\epsilon}$$
$$\mathcal{E} = \frac{\sigma}{\epsilon}$$

Fig. 3.5. *Electric field at $x = 0$ is calculated considering the flux through a cylinder around the interface and immediately above the metal. The only surface crossed by the electric field is the area A of the cylinder towards the semiconductor.*

The field generated by the charges at the surface of the metal is calculated at the surface of the semiconductor. Then, the involved dielectric constant is the dielectric constant of the semiconductor (e.g. $\epsilon_{Si} = 11.7\epsilon_0$).

$$\mathcal{E}(0) = -\frac{qN_d x_d}{\epsilon_s} \tag{3.9}$$

fro eq. 3.8 the electric field in the depletion layer is calculated:

$$\mathcal{E}(x) = \mathcal{E}_0 + \int_0^x \frac{qN_d x_d}{\epsilon} dx \tag{3.10}$$

where $\mathcal{E}_0$ is the electric field at $(x = 0)$. The solution is:

$$\mathcal{E}(x) = -\frac{qN_d x_d}{\epsilon_s} + -\frac{qN_d x}{\epsilon_s} \rightarrow \mathcal{E}(x) = \frac{qN_d}{\epsilon_s}(x - x_d) \tag{3.11}$$

Finally, in the bulk of the semiconductor $(x > x_d)$ the total charge is zero, and the electric field is null. Given the electric field, the electric potential can be calculated:

$$\phi = \phi_0 - \int_0^x \mathcal{E}(x) dx \tag{3.12}$$

Inside the metal $(x < 0)$ the potential is null and in the depletion layer is:

$$\phi(x) = -\int_0^x \frac{qN_d}{\epsilon_s}(x - x_d) dx = -\frac{qN_d}{\epsilon_s}\left(\frac{1}{2}x^2 - x x_d\right) \tag{3.13}$$

at $x = x_d$ the potential reaches the maximum value:

$$\phi_{max} = \frac{1}{2}\frac{qN_d}{\epsilon_s} x_d^2 = \phi_i \tag{3.14}$$

Since the electric field at $x > x_d$ is null, the maximum value of the potential is maintained in the bulk of the semiconductor. This quantity is the built-in potential of the metal-semiconductor junction.
The built-in potential is equal to the difference of the work functions, then it is known since the beginning when the materials for the junctions are chased.
From the built-in potential we can calculate the size of the depletion layer.

$$x_d = \sqrt{\frac{2\phi_i \epsilon_s}{qN_D}} \tag{3.15}$$

Following the band diagram in figure 3.3, $N_d = 10^{16}\ cm^{-3}$; $q\phi_s = 4.25\ eV$; $q\phi_m = 4.60\ eV$ from which we calculate $\phi_i = 0.25\ V$ and $x_d = 179\ nm$.

Finally, we can observe from eq. 15 that the extension of the depletion layer depends on the doping, in particular larger is the doping more narrow is the depletion layer. To this regard, is worth to note that also the built-in potential depends on the doping but as a logarithm. Then the above stated behaviour remains valid.

3.3 Biased metal-semiconductor junction

The equilibrium condition illustrated in the previous section is altered by the application of an external voltage. The voltage is applied through two contacts made by metal wires and the two sides of the metal-semiconductor system. The voltage is applied by means ot additional metal-metal and a metal-semiconductor junctions. This gives rise to a tautology: in order to study the behaviour of a junction we need another junction of the same kind. For the moment, let us consider that the junctions between the system under study and the circuit are negligible, namely they are ohmic

contacts. We will study later the properties of the ohmic contacts and the conditions under which they occurs.
The system under study is formed by three distinct regions: the metal, the junction, and the semiconductor. Thus the voltage is applied to a series of three different materials each with a distinct electric property.
In particular, the metal is characterized by a very large population of mobile electrons ($n \approx 10^{22} cm^{-3}$), the doped semiconductors contains a smaller quantity of electrons ($n = N_d$), and finally the junction is practically depleted of mobile electrons ($n \ll N_d$). Then, the three regions are characterized by very different conductivities and the depletion layer is the less conductive element of the series. As a consequence, the applied voltage drops almost completely across the depletion layer.
This is an important assumption considered valid throughout this textbook. The applied voltage falls across the depletion layer, and the electric field in the bulk of the semiconductor is null. for this reason the bulk of the semiconductor is also called neutral zone.

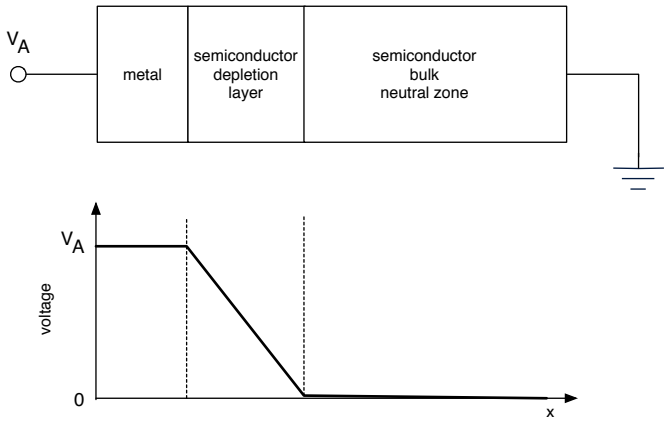


Fig. 3.6. *Distribution of the applied voltage along the metal-semiconductor system. The voltage in the bulk of the semiconductor is practically considered negligible.*

The external voltage violates the conditions of thermal equilibrium and then the statistical law derived in chapter 1 are no more valid. However, for modest values of the applied voltage it is still possible to continue to calculate the concentrations of electrons and holes using the statistical laws. This condition is the so-called *quasi-equilibrium hypothesis*.
The applied voltage shown in figure 5 locally modifies the energy of the electrons. In particular electrons acquire an extra potential energy equal to qV, since the charge of the electron is negative, if V_A is positive the energy is shifted towards negative values. The electrons in the metal, are subjected to the same voltage, then the energy of all the electrons in the metals is translated downward of the quantity qV_A. On the other hand, the energy of the electrons in the bulk of the semiconductor remains unchanged.
Eventually, the application of an external voltage, does not change the barrier applied to the electrons of the metal ($q\phi_b$) but it changes the barrier applied to the electrons of the semiconductors

$(q\phi_i - qV_A)$: when V_A is positive the barrier $(q\phi_i)$ decreases and when V_A is negative the barrier increases. These two conditions are called *forward bias* and *reverse bias*.

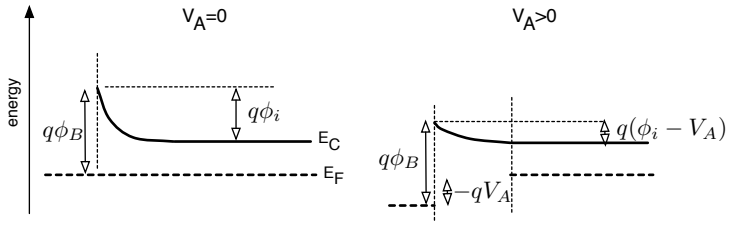


Fig. 3.7. *Electrons potential energy (conduction band) before and after the application of an external voltage* V_A.

The uneven change of the barriers breaks the equilibrium between the currents. The current from the metal is unchanged and it maintains the thermal equilibrium value. On the other hand, the current from the semiconductor changes. In case of forward bias the barrier decreases and the current increases while under reverse bias the barrier increases and the current decreases. This phenomenon depicts the behavior of a diode and it will be quantitatively studied in the next section. The bias also affects the depletion layer size.

$$x_d = \sqrt{\frac{2\epsilon_s}{qN_D}(\phi_i - V_A)} \tag{3.16}$$

Then in case of a reverse bias the depletion layer becomes larger and under forward bias it becomes thinner.
The behavior of the metal-semiconductor system under an applied voltage can be conveniently studied separating the capacitive and the resistive effects according to the classical approach adopted in network theory. h

3.3.1 The capacitance of the junction

The capacitance measures the modulation of the concentration of charges due to the applied voltage.

$$C = \left|\frac{dQ}{dV}\right| \tag{3.17}$$

The amount of fixed charges in the depletion region is proportional to x_n, then since the length of the depletion layer is modulated by the applied voltage, also the charge therein contained varies according to the applied voltage

$$Q = qN_D x_d = qN_D\sqrt{\frac{2\epsilon_s}{qN_D}(\phi_i - V_A)} = \sqrt{2qN_D\epsilon_s(\phi_i - V_A)} \tag{3.18}$$

The capacitance is simply calculated applying the definition. Note that as many other quantities in this textbook, the following equation actually describes the density of capacitance (F/m^2).

$$C = \left| \frac{dQ}{dV_A} \right| = \sqrt{2qN_D\epsilon_s} \frac{1}{2(\phi_i - V_A)} = \epsilon_s \sqrt{\frac{qN_D}{2\epsilon_s} \frac{1}{\phi_i - V_A}} \tag{3.19}$$

At the end

$$C = \frac{\epsilon_s}{x_d} \tag{3.20}$$

The depletion layer behaves as the dielectric of a capacitor with parallel plates.
It is interesting to note that the capacitance diverges when $V_A \geqslant \phi_i$, anyway in these conditions the quasi-equilibrium hypothesis my be no more valid and the calculated quantities do not represent the real device.
The capacitance is better manifested when the resistive behavior is negligible and this happens under reverse bias.
Eq. 3.19 shows that the capacitance is not constant, but it rather depends on the applied voltage. The behavior of the capacitance as a function of the applied voltage (the C/V curve) is a fundamental tool to measure some important parameters of electronic devices. In the case of metal-semiconductor junction, it provides a measure of the doping concentration and the built-in potential. To this scope, let us consider the inverse of the square of equation 3.19

$$C = \frac{1}{C^2} = \frac{2(\phi_i - V_A)}{q\epsilon_s N_D} \tag{3.21}$$

Then, measuring C at different values of V_A and plotting $\frac{1}{C^2}$ as a function of V_A we obtain that the experimental points are aligned along a straight line whose slope contains the concentration of dopant atoms and the intercept with the horizontal axis is the built-in potential.

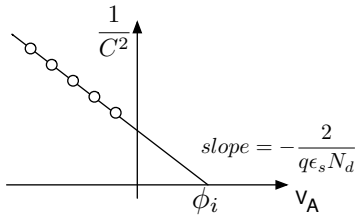
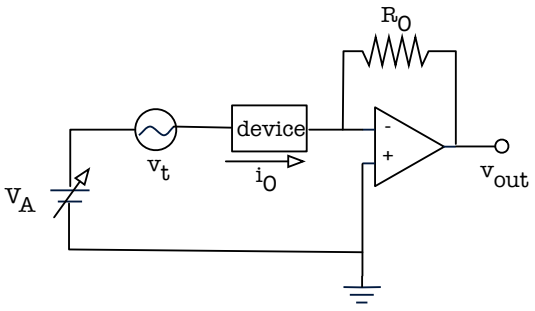


Fig. 3.8. *$\frac{1}{C^2}$ as a function of V_A and the relations of the curve parameters with the built-in potential and the donors concentration. Experimental points are collected under reverse bias where the current is negligible and the device behaves as a almost pure capacitor.*

Experimental set-up for the C/V curve measurement

In order to measure the C/V curve it is necessary to bias the device with a d.c. voltage (V_A) and to superimpose an a.c. signal. The time variable part of the total applied signal allows for the extraction of a current proportional to the capacitance value. Figure 3.9 shows a simple example



$$V_0 = V_A + v_t e^{j\omega t}$$
$$i_0 = C(V_A) \cdot \frac{dV_0}{dt} = C(V_A) j\omega e^{j\omega t}$$
$$v_{out} = -R_0 C(V_A) j\omega e^{j\omega t}$$
$$\rightarrow C(V_A) = \frac{|v_{out}|}{R_0 \omega v_t}$$

Fig. 3.9. *Electronic circuit for the measurement of the C/V curve of a device.*

of a circuit to measure the C/V curve. The circuit is based on a current to voltage converter made with a operational amplifier.

It is important to consider that V_A fixes the value of the capacitance as reported in equation 3.19 while the a.c. signal is the probe necessary to measure the capacitance. The requirement is $\|v_t\| \ll V_A$ so that the probe signal does not alter the value of the capacitance.
The frequency ω of the probe signal is not particularly important for the metal-semiconductor junction, but it is very important in other system such as the metal-oxide-semiconductors (see chapter 9). From an electronic point of view, it is interesting to note that in order to correctly work, the circuit in figure 3.9 has to behave as a differentiator amplifier. Due to the frequency response of the operational amplifier, this is ensured only if ω is smaller than the frequency at which the transfer function of the feedback network meets the transfer function of the amplifier.

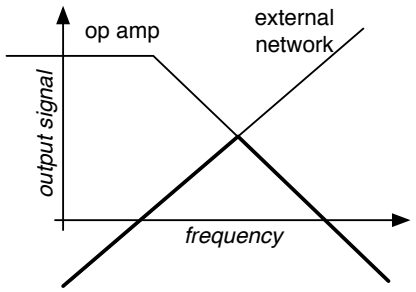


Fig. 3.10. *Transfer function of the circuit of figure 8.*

3.3.2 The I/V characteristics

The rectifying I/V curve of metal-semiconductor junctions was known since the beginning of the twentieth century when the use of a diode made of lead sulfide (PbS), a mineral known as galena,

and phosphor bronze was actually introduced as the first solid state electronic device. PbS is a small band gap semiconductor and phosphor bronze is an alloy of tin with a small amount of copper and phosphor. The contact was simply made leaning the wire on the mineral surface. Such components where quite popular before the advent of modern devices.The diode made of metal-semiconductor junction is called **Schottky diode** after Walter Schottky, the german physicist that developed in the 30's this device.
The relationship between the current and the applied voltage can be calculated following two approaches. The first considers the current as the result of the emission of electrons from one material towards the other, in practice at the interface two unchanged the total current is different from zero. In this approach the actual shape of the barrier is not relevant but only its height.
In the second approach the current is directly calculated from the drift/diffusion model of current. Calculations are restricted to the depletion layer where both the electric field and the concentration gradient may exist. The two approaches qualitatively reaches the same result with a small difference about the reverse current.
The quest for a more accurate model may be useless because of the presence of the surface states which strongly affect the behaviour of real metal-semiconductor junction. Then the thermionic and the drift-diffusion models are sufficient to explain the experimental characteristics of the Schottky diode.
It is important to note that in both the models the concentration of electrons in the conduction band in the non equilibrium condition is still calculated using the statistical law of eq. 2.24. The use of the equilibrium quantities is guaranteed by the assumption of quasi-equilibrium; namely the applied voltage introduces a small perturbation of the equilibrium quantities. Let us postpone the discussion about the validity of the assumption after the study of the PN junction.

The thermionic current model

At thermal equilibrium the total current across the device is null; namely, the absolute value of the current that flow from the metal to the semiconductor and from the semiconductor towards the metal are equal: $\|J_{MS}\| = \|J_{SM}\|$.
In this model we consider that the current at the interface ($x = 0$) is given by the thermal motion of the charges, actually this is particularly true for the current directed towards the metal where the electric field is zero while towards the semiconductor the charges are subjected to the electric field ($\mathcal{E}(0)$) given by eq. 3.9. Since at the equilibrium the two currents are equal we can calculate the current only considering the thermal motion.
The thermal flow of a gas of particle impinging onto a surface is studied by the kinetic theory of gases and it is given by:

$$F = \frac{n v_{ave}}{4} \tag{3.22}$$

The above equation is called law of Knudsen. The average velocity of the particles (v_{ave}) results from the Maxwell distribution of velocities:

$$v_{ave} = \sqrt{\frac{8kT}{\pi m}} \tag{3.23}$$

For the electrons in the semiconductor the mass is replaced by the effective mass.

At $x = 0$ the current due to the flow of electrons is:

$$J_{th} = \frac{1}{4} q n_0 v_{ave} \tag{3.24}$$

The thermal current is proportional to the concentration of the electrons at the interface (n_0). This quantity can be calculated from eq. 2.24 considering that $\phi_B = E_C(x = 0) - E_F$:

$$n_0 = N_C \cdot exp\left(-\frac{q\phi_B}{kT}\right) \tag{3.25}$$

Then the thermal current is:

$$J_{th} = \frac{1}{2} q v_{ave} N_C \cdot exp\left(-\frac{q\phi_B}{kT}\right) = K N_C \cdot exp\left(-\frac{q\phi_B}{kT}\right) \tag{3.26}$$

Where $K = \frac{1}{2} q v_{ave}$. The barrier ($q\phi_B$) can be written as: $q\phi_B = q\phi_i + (E_C - E_F)_{bulk}$ and then the thermal current is:

$$J_{th} = -K N_C \cdot exp\left(-\frac{q\phi_i + (E_C - E_F)_{bulk}}{kT}\right) = K N_C \cdot exp\left(-\frac{(E_C - E_F)_{bulk}}{kT}\right) \cdot exp\left(-\frac{q\phi_i}{kT}\right) = K N_D \cdot exp\left(-\frac{q\phi_i}{kT}\right) \tag{3.27}$$

The current across the interface depends on the doping concentration and on the height of the barrier in the semiconductor.
When $V_A \neq 0$ the equilibrium is no more valid, but the current J_{MS} remains unchanged because the V_A does not affect ϕ_B, but the current J_{SM} changes because the barrier for the electrons of the semiconductor is reduced of a quantity qV_A. Namely:

$$J_{MS} = K N_D \cdot exp\left(-\frac{q\phi_i}{kT}\right);\ J_{SM} = K N_D \cdot exp\left(-\frac{q(\phi_i - V_A)}{kT}\right); \tag{3.28}$$

Noteworthy, the applied voltage modifies only the current originated from the semiconductor that is the material with less electrons.
The total current is:

$$J = J_{SM} - J_{MS} = K N_D \cdot exp\left(-\frac{q\phi_i}{kT}\right) \cdot \left(exp\left(\frac{qV_A}{kT}\right) - 1\right); \tag{3.29}$$

This is the typical I/V characteristic of the diode $J = J_0(exp(\frac{qV_A}{kT}) - 1)$.
The quantity J_0 is the reverse current and it corresponds to the thermal current from the metal to the semiconductor.
The reverse current can be written as:

$$J_0 = \frac{1}{2} q v_{ave} N_C exp\left(-\frac{q\phi_B}{kT}\right) \tag{3.30}$$

Replacing in eq. 29 the definition of average thermal velocity and N_C (eq. 21 of chapter 1) we get:

$$J_0 = \frac{1}{4} q \sqrt{\frac{8kT}{\pi m_n^*}} 2 \left(\frac{2\pi m_n^* kT}{h^2} \right)^{3/2} exp\left(-\frac{q\phi_B}{kT} \right) \tag{3.31}$$

That can be written as

$$J_0 = RT^2 exp\left(-\frac{q\phi_B}{kT} \right) \tag{3.32}$$

where $R = \frac{4\pi m_n^* k^2 q}{h^3}$ is the Richardson constant. This is the thermionic current equation ruling the working principles of the thermionic devices discussed in the preface. For free electrons ($m_n^* = m_0$) $R = 120 \frac{A}{cm^2 K^2}$.
The simple thermionic model has a remarkable agreement with the data fitting the experimentally measured current in a range wider than 5 orders of magnitudes in forward bias when the device is forward bias.

The drift/diffusion model

The current can also be directly calculated from the drift/diffusion model. In particular, the current can be calculated integrating the drift/diffusion equation in the depletion layer where both the potential and the gradient of concentration are different from zero. The straightforward hypothesis for this approach is that the depletion layer is sufficiently large to support the definition of the mobility and the diffusion constant. This means that the depletion has to be at least longer than few electron mean free paths.
Since electric field and electrons concentrations are a function of x. The drift and diffusion currents evolve along the depletion layer in order to maintain a constant current in the device:

$$J = q \left[n\mu_n \mathcal{E}(x) + D_n \frac{dn}{dx} \right] \tag{3.33}$$

using the definition of electric potential and the Einstein relationship it becomes:

$$J = qD_n \left[-\frac{q}{kT} n \frac{d\phi}{dx} + \frac{dn}{dx} \right] \tag{3.34}$$

It is worth to remark that the relationship between the diffusion constant and the mobility was derived under thermal equilibrium conditions, then its use is allowed by the quasi-equilibrium assumption.
The solution of eq. 3.36 can be obtained integrating both sides in the depletion layer from $x = 0$ to $x = x_d$ and from $\phi = 0$ to $\phi = \phi_i - V_A$.
The integral can be solved multiplying the equation to $exp(-\frac{q\phi}{kT})$.

$$J \int_0^{x_d} exp(-\frac{q\phi}{kT}) dx = qD_n \left[\int_0^{x_d} -\frac{q}{kT} n \; exp\left(-\frac{q\phi}{kT} \right) \frac{d\phi}{dx} dx + \int_0^{x_d} exp\left(-\frac{q\phi}{kT} \right) \frac{dn}{dx} dx \right] \tag{3.35}$$

The first of the two integrals in the right hand side can be solved using the rule of integration by parts: $\int fg' = fg - \int f'g$) where $f = n$ and $g = exp(-\frac{q\phi}{kT})$

$$qD_n\left(\left[n\ exp\left(-\frac{q\phi}{kT}\right)\right]_0^{x_d}-\int_0^{x_d}exp\left(-\frac{q\phi}{kT}\right)\frac{dn}{dx}dx+\int_0^{x_d}exp\left(-\frac{q\phi}{kT}\right)\frac{dn}{dx}dx\right)=qD_n\left[n\ exp\left(-\frac{q\phi}{kT}\right)\right]_0^{x_d} \tag{3.36}$$

with the following boundary conditions: $q\phi(0)=0,\ q\phi(x_d)=q\phi_B-(E_C-E_F)_{bulk}-qV_A$; $n(0)=N_Cexp(-\frac{q\phi_B}{kT});\ n(x_d)=N_Cexp(-\frac{(E_C-E_F)_{bulk}}{kT})$. Replacing the boundary conditions we get:

$$qD_n\left[N_Cexp\left(-\frac{(E_C-E_F)_{bulk}}{kT}\right)exp\left(-\frac{q\phi_B-(E_C-E_F)_{bulk}-qV_A}{kT}\right)-N_Cexp(-\frac{q\phi_B}{kT})\cdot 1\right] \tag{3.37}$$

$$qD_n\left[N_Cexp\left(-\frac{q\phi_B}{kT}\right)exp\left(\frac{qV_A}{kT}\right)-N_Cexp\left(-\frac{q\phi_B}{kT}\right)\right]=qD_nN_Cexp\left(-\frac{q\phi_B}{kT}\right)\left[exp\left(\frac{qV_A}{kT}\right)-1\right] \tag{3.38}$$

Then the total current is given by:

$$J=\frac{qD_nN_Cexp(-\frac{q\phi_B}{kT})}{\int_0^{x_d}exp(-\frac{q\phi}{kT})dx}\left[exp\left(\frac{qV_A}{kT}\right)-1\right] \tag{3.39}$$

That can be immediately recognized as the diode equation $J=J_0(exp(-\frac{qV_A}{kT})-1)$.
To complete the calculation it is necessary to solve the integral at the denominator, whose solution depends on the applied voltage. Then, differently from the thermionic model, the drift/diffusion model predicts that the reverse current depends on the applied voltage.
In order to calculate the integral the analytical form of the potential derived in eq. 3.13 is used. A simplified solution is obtained replacing the potential with its first-order approximation. In this way, the integral is under estimated, and since the integral is at the denominator of the reverse current, the reverse current will be over estimated.

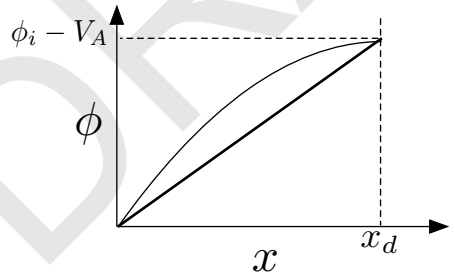


Fig. 3.11. *Potential in the depletion layer and its linear approximation. Dashed area is the difference in the integral between approximated and "exact" solutions.*

The integral at the denominator of eq. 3.39 is approximated as:

$$\int_0^{x_d}exp\left(-\frac{q\phi}{kT}\right)dx\approx\int_0^{x_d}exp\left(-\frac{q}{kT}\frac{\phi_i-V_A}{x_d}x\right)dx=\frac{kT}{q}\frac{x_d}{\phi_i-V_A}\left[1-exp\left(-\frac{q}{kT}(\phi_i-V_A)\right)\right] \tag{3.40}$$

The solution can be further simplified considering that $\phi_i - V_A \gg \frac{kT}{q} \approx 26\ meV$. This condition is surely fulfilled under reverse bias when $V_A < 0$. With this approximation, the exponential in bracket is negligible with respect to 1, then replacing x_d with the expression of eq. 3.15 we get:

$$\int_0^{x_d} exp\left(-\frac{q\phi}{kT}\right) dx \approx \frac{kT}{q}\sqrt{\frac{2\epsilon_s}{qN_D(\phi_i - V_A)}} \tag{3.41}$$

Eventually, the reverse current is:

$$J_0 = qD_n N_C exp\left(-\frac{q\phi_B}{kT}\right)\frac{q}{kT}\sqrt{\frac{qN_D(\phi_i - V_A)}{2\epsilon_s}} \tag{3.42}$$

It can be observed that the electrons in $x = 0$ are subjected to the electric field $\mathcal{E}_{max}$, so the electrons injected from the metal to the semiconductor are accelerated by the electric field at the interface and produce a drift current $J_0 = qn_0\mu_n\mathcal{E}_{max}$. Considering the expression of the electric field at the interface (eq. 3.9), the depletion layer width under bias (eq. 3.16) and the Einstein relation we obtain

$$J_0 = q \cdot \frac{q}{kT} \cdot N_C exp\left(-\frac{q\phi_B}{kT}\right)\sqrt{\frac{2qN_D(\phi_i - V_A)}{\epsilon_s}} \tag{3.43}$$

Apart a term $\frac{1}{2}$, likely due to the approximations necessary to solve the drift-diffusion model, we can say that the reverse current is the current due to the electrons that from the metal cross the barrier Φ_B. Note that the reverse current depends on the doping of the semiconductor because of the electric field at the interface and not because of the concentration of electrons.
With respect to the thermionic model, the drift diffusion model introduces a slight dependence of the reverse current from the applied voltage. We see in the next section that actually even Φ_B has a small dependence from the applied voltage and then the differences between the two models are attenuated. On the the hand, we have always to consider the limit of the ideal metal-semiconductor junction with respect to the real device.
In conclusion, the junction between a N-type semiconductor and a metal whose work function is greater than the work function of the semiconductor behaves as a rectifier.
The reason of the non linear I/V curve lies in the existence of the depletion layer: a region at the surface of the semiconductor depleted of mobile majority charges. Since the interface is depleted of electrons, the applied voltage drops across the depletion layer and then the barriers keeping in equilibrium the current are differently affected by the applied voltage. The barrier applied to the electrons of the metal (ϕ_B) is unchanged and the barrier applied to the electrons of the semiconductor (ϕ_i) is altered. The applied bias modulates only the current from the semiconductor, it is increased in forward bias and it is depressed in reverse bias.
The current-voltage relationship is non linear and it depends exponentially on the applied voltage. This result can be obtained following either the thermionic current model or the drift/diffusion current model and both the models give the same result.
With respect to the PN junction diode (that is discussed in chapter 4) the Schottky diode is usually more fast (only majority charges are involved) and the voltage drop in forward bias is smaller. On the other hand, the reverse current is larger since it is a current from a metal (a large reservoir of electrons) and the behavior of the device is affected by the surface states whose effects are discussed below.

The same behavior can be obtained with the symmetric situation of a junction between a P-type semiconductor and a metal of smaller work function: $\phi_S > \phi_m$. In this case the device is still a diode and the charge carriers are holes. Due to the smaller mobility of holes this configuration is not convenient for real devices. The band diagram and the equilibrium electrostatic quantities are shown in figure 12.

3.3.3 Barrier height lowering

The electric potential has been calculated above considering only the double layer of charges formed by the electrons at the surface of the metal and the fixed charges distributed in the semiconductor. From the point of view of an electron in the depletion layer an additional electric field has to be considered. This supplementary potential comes from the fact that a charge in the semiconductor is actually found in a dielectric material close to a charged conductor plane. From electrostatics it is known that a charge (q) close to the surface of a metal experiences a Coulomb force equivalent to that produced by a virtual charge (image charge) of the same magnitude but opposite sign and placed at the symmetrical point behind the charged plane. The situation is depicted in figure 13.

The image charge gives rise to an electric field $\mathcal{E}_i$ and a potential:

$$\phi_i(x) = -\int_x^\infty \mathcal{E}_i dx = \frac{q}{16\pi\epsilon_s x} \tag{3.44}$$

This potential is additive to the potential generated by the double layer. The potential of the image charge is confined at few nanometers from the surface. In this space the potential of the double layer can be linearly approximated as: $\phi_{DL} = -\mathcal{E}_{max}x$. Note that one potential is decreasing and the other is increasing, thus there is a coordinate where the potential reaches a maximum value (see figure 13).

$$\frac{q}{16\pi\epsilon_s x_{max}} = -\mathcal{E}_{max}x_{max} \rightarrow x_{max} = \sqrt{\frac{q}{16\pi\epsilon_s \mathcal{E}_{max}}} \tag{3.45}$$

The barrier lowering is equal to the maximum electric field times x_{max} that depends also on the maximum electric field. Eventually, the barrier lowering is proportional to the fourth root of the built-in potential and then, under applied voltage, it is inversely proportional to the fourth power of the semiconductor potential barrier:

$$\Delta\phi_B = K(\phi_i - V_A)^{\frac{1}{4}} \tag{3.46}$$

Then then height of the barrier applied to the electrons of metal and semiconductor slightly depends on the applied voltage.

3.4 Ohmic contact

A junction made by a metal and a N-type semiconductor is rectifying when the work function of the metal is greater than the work function of the semiconductor. We see in this section that if the relative magnitude of the work functions is inverted the I/V curve becomes linear. Such a junction is said non-rectifying. This condition occurs when an N-type semiconductor is joint to a metal whose work function is smaller than that of the semiconductor. In this situation, the equilibrium

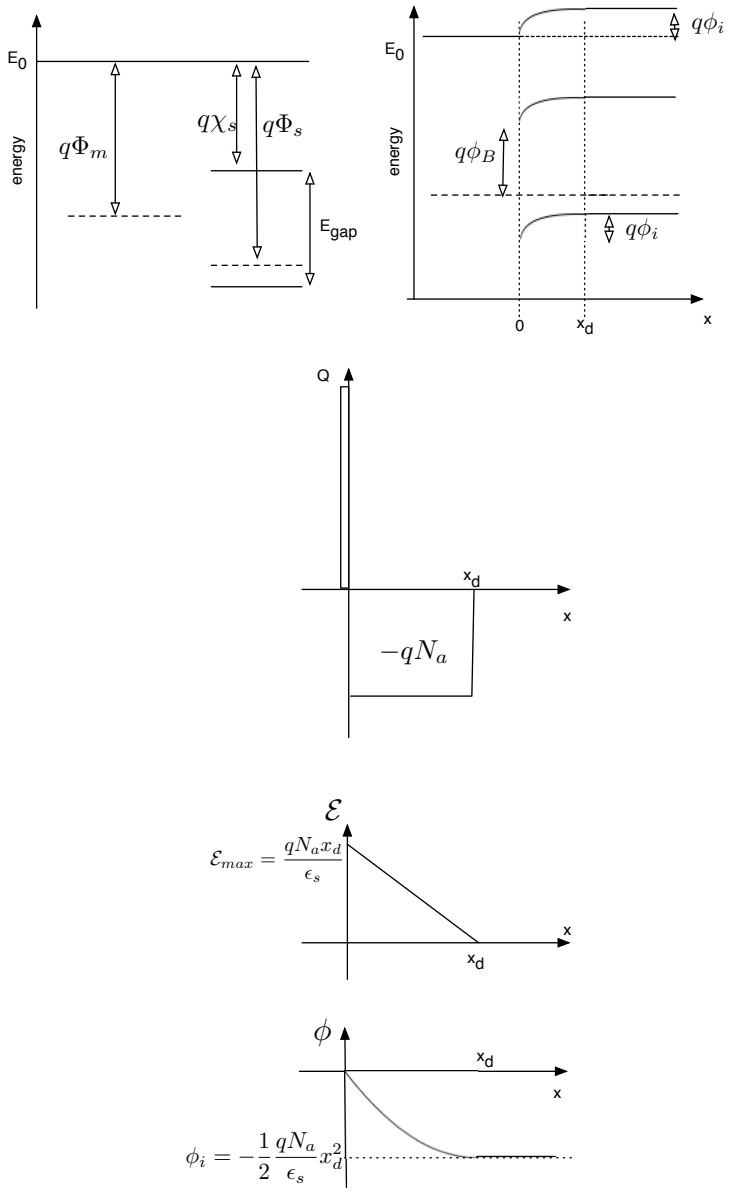


Fig. 3.12. *Metal-semiconductor junction where the work function of the semiconductor is greater than that of the metal, and the semiconductor is P-type. The figure shows the band diagrams before and after the junction and the electrostatic quantities (charge, electric field, and potential) at the equilibrium*

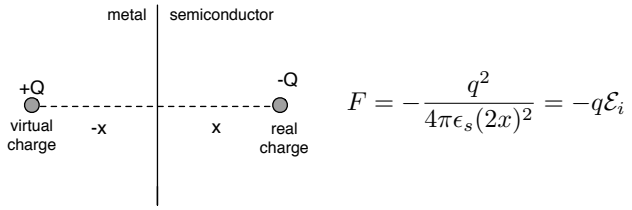


Fig. 3.13. *Coulomb force between a conductor plane and a charge -q, the image charge is placed at distance 2x from the real charge. The field $\mathcal{E}_i$ is the electric field generated by the virtual charge.*

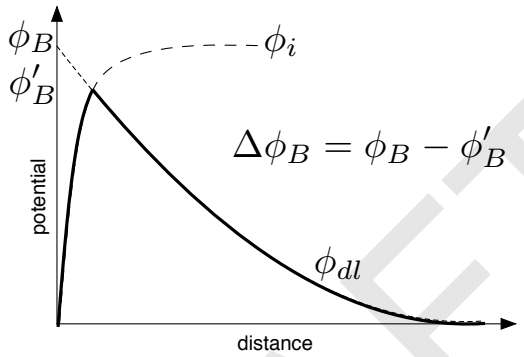


Fig. 3.14. *Total potential is given by the sum of the potential due to the double layer and the potential of the virtual image. As a consequence the barrier at the interface $\Delta\phi_b$ is reduced to the corrected value $\Delta\phi'_b$.*

is reached by a displacement of electrons from the metal to the semiconductor. This leads, in the semiconductor, to a region at the interface with the metals where the concentration of electrons is larger than the bulk. Such a region is called an *accumulation layer*. The electrons that left the metal leave behind a distribution of positive charges, that as usual for a metal, forms a thin layer at the interface with the semiconductor. The situation is depicted in the equilibrium band diagram shown in figure 3.15.
The double layer of charges is now formed by a thin layer of positive charges in the metal and a distribution of mobile electrons in the semiconductor. This gives rise to a built-in potential whose sign is opposite to the previous case and whose magnitude is still given by the difference between the work functions.

Differently than metals semiconductors can allow for a distribution of charges in their volume. Then, the excess of mobile charges are not accumulated at the surface of the semiconductor but they are distributed through the semiconductor (see figure 3.15). The mobile charges are not bounded to the fix donor atoms and then the distribution of the excess electrons cannot be easily predicted. The total charge in the accumulation layer is $Q = (-n + N_D) - n'$ where $n = N_D$ is the charge in the non perturbed semiconductor and n' are the charges transferred from the metal. The profile of the accumulation charge depends on the profile of the potential.
The charge density at the interface is:

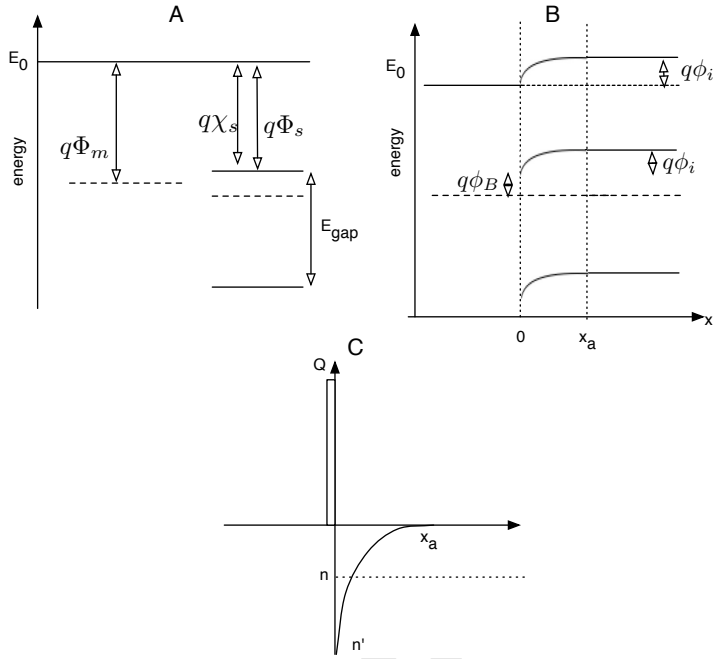


Fig. 3.15. *Band diagram and charge distribution of a non-rectifying junction. A: Band diagrams of pristine materials. B: equilibrium band diagram. The downward bending of the conduction band signals the increase of electrons concentration that is extended for a lenght x_a. C: Distribution of charges, the concentration of electrons before the junction is marked by a dotted line.*

$$n_s = N_C exp\left(-\frac{q\phi_B}{kT}\right) = N_C exp\left(-\frac{(E_C - E_F)_{bulk}}{kT}\right) exp\left(-\frac{q\phi_i}{kT}\right) = N_D exp\left(-\frac{q\phi_i}{kT}\right) \quad (3.47)$$

The behavior of the concentration of the accumulated electrons is an exponential function of the potential:

$$n(x) = n_s exp\left(\frac{q\phi(x)}{kT}\right) \quad (3.48)$$

The analytical behavior of the accumulation charge distribution, and the related potential, can be found solving the Poisson equation.

$$\frac{d^2\phi}{dx^2} = -\frac{\rho(x)}{\epsilon_s} = -\frac{qn(x)}{\epsilon_s} = \frac{qn_s}{\epsilon_s} exp\left(\frac{q\phi(x)}{kT}\right) \quad (3.49)$$

The problem consists in determining the behavior inside a dialectric of a charge distribution whose value at surface is considered known (n_s).
To solve the equation let us consider that $\frac{d\phi}{dx} = -\mathcal{E}$, then

$$\frac{d^2\phi}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{d\mathcal{E}}{d\phi}\frac{d\phi}{dx} = \mathcal{E}\frac{d\mathcal{E}}{d\phi} \tag{3.50}$$

Poisson equation can then be written as

$$\mathcal{E}\frac{d\mathcal{E}}{d\phi} = \frac{qn_s}{\epsilon_s} exp\left(\frac{q\phi(x)}{kT}\right) \tag{3.51}$$

This can be integrated in $d\mathcal{E}$ from 0 to $\mathcal{E}$ and in $d\phi$ from $-\phi_i$ to ϕ

$$\int_0^{\mathcal{E}} \mathcal{E} d\mathcal{E} = \frac{qn_s}{\epsilon_s}\int_{-\phi_i}^{\phi} exp\left(\frac{q\phi}{kT}\right) d\phi \tag{3.52}$$

which gives:

$$\frac{1}{2}\mathcal{E}^2 = \frac{n_s kT}{\epsilon_s}\left[exp\left(\frac{q\phi}{kT}\right) - exp\left(-\frac{q\phi_i}{kT}\right)\right] \tag{3.53}$$

since $q\phi_i \gg kT$ the electric field produced by the accumulation charges is:

$$\mathcal{E} = \sqrt{\frac{2n_s kT}{\epsilon_s}} exp\left(\frac{q\phi}{2kT}\right) \tag{3.54}$$

And the potential is:

$$-\frac{d\phi}{dx} = \sqrt{\frac{2n_s kT}{\epsilon_s}} exp\left(\frac{q\phi}{2kT}\right) \tag{3.55}$$

The integration of the equation is easily achieved by a separation of the variables

$$\int_0^x \sqrt{\frac{2n_s kT}{\epsilon_s}} dx = -\int_0^{\phi} exp\left(-\frac{q\phi}{2kT}\right) d\phi \tag{3.56}$$

and the solution is:

$$exp\left(-\frac{q\phi}{2kT}\right) = \sqrt{\frac{q^2 n_s}{2\epsilon_s kT}} x + 1 \tag{3.57}$$

The above equation depends on an important parameter of the material called the **Debye length**

$$L_D = \sqrt{\frac{\epsilon_s kT}{q^2 n_s}} \tag{3.58}$$

The Debye length is a fundamental quantity ruling the separability of charges in a material. In practice, it is the length scale at which a dipolar charge distribution can be created by an electric field. As the dielectric constant increases the Debye length becomes larger and the charges, as well the electric field, are more widely distributed in the material.
Using the Debye length, the solution of the Poisson equation can be written as:

$$exp(-\frac{q\phi}{2kT}) = 1 + \frac{x}{\sqrt{2}L_D} \tag{3.59}$$

This relation allows to calculate the distribution of the accumulated charges

$$\rho = qn' = qn_s exp\left(\frac{q\phi}{kT}\right) = qn_s \frac{1}{(1+\frac{x}{\sqrt{2}L_D})^2} \tag{3.60}$$

the excess of charges decay in the semiconductor as x^{-2} .
The potential, that can be directly calculated from equation 3.61, has a logarithmic behavior

$$\phi = -\frac{2kT}{q} ln\left(1+\frac{x}{\sqrt{2}L_D}\right) \tag{3.61}$$

The depth of the accumulation region (x_a can also be calculated from the boundary condition: at $\phi(x_a) = -\phi_i$:

$$x_a = \sqrt{2}L_D \left[exp\left(\frac{q\phi_i}{kT}\right) - 1\right] \tag{3.62}$$

The Debye length is a reference distance for the size of the accumulation layer, approximately half of the accumulated charge lies at a distance of $\sqrt{2}L_D$ from the interface with the metal.
As an example, let us consider a N-Type silicon doped with a concentration $N_D = 10^{16}\ cm^{-3}$ of donors. The work function is then $q\Phi_s = 4.25\ eV$. The junction is formed with a metal with a work function $q\Phi_m = 4.10\ eV$. Such a work function can be found in aluminum.
At the equilibrium, the built-in potential is $\phi_i = (1/q)(q\Phi_m - q\Phi_s) = -0.15\ V$. The concentration of accumulated charges at the interface is $n_s = 3.3 \cdot 10^{18}\ cm^{-3}$, the Debye length is $L_D = 2.3\ nm$ and the depth of the accumulation layer is $x_a = 55\ nm$. The total accumulated charge is calculated integrating eq. 3.60 from 0 to x_a whose result is about $1.17 \cdot 10^{19}\ cm^{-3}$.

3.4.1 Ohmic contact

The presence of an accumulation layer, instead of a depletion layer, changes drastically the distribution of the voltage applied to a metal-semiconductor system. Indeed, while the depletion layer corresponds to a region of negligible conductivity, in the accumulation layer the concentration of charges is larger than in the rest of the semiconductor. Thus the conductivity is larger in the accumulation layer with respect to the bulk. As a consequence, the applied voltage tends to drop completely in the bulk of the semiconductor and no power is dissipated in the contact region. The absence of power dissipation is a practical definition of a ohmic contact, namely the contact simply vehicles the current and the voltage to the material of interest.

It is important to note that as we have discussed in the case of the Schottky diode, the applied voltage modifies the band bending in the contact region. This is still valid also for the non-rectifying contact. Here for negative applied voltages the band bending tends to decrease, namely the concentration of accumulated charges becomes smaller until the condition of flat-band is reached, at more negative values the band may bend upward and the initial accumulation layer becomes turned into a depletion layer. This change of character induced by the applied voltage is common to all the junctions among semiconductors. it will be fully manifested in the metal-oxide-semiconductor junction, here it introduces a caveat about the fact that behaviors predicted from the equilibrium condition strictly hold for perturbation around the equilibrium condition. On the other hands, this is the condition for which the quasi-equilibrium hypothesis holds. However, the effective ohmic character of junction is valid only in a limited interval of voltage.

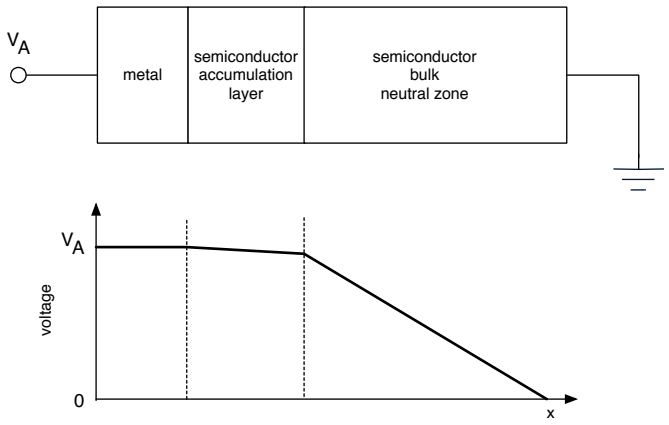


Fig. 3.16. *Distribution of the applied voltage in case of an accumulation layer at the metal-semiconductor interface. Practice. The drop of voltage in the contact region is practically negligible.*

The ohmic contact is also found in the symmetric case of P-type semiconductor where the work function of the metal exceeds that of the semiconductor. Eventually, figure 3.17 reassumes the four metal-semiconductor junctions and their behaviours around the equilibrium condition.

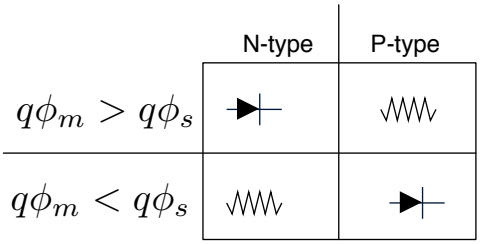


Fig. 3.17. *The combinations of dopings and work functions give rise to four possible junction.*

It is important to remark that the phenomena outlined in this chapter and in particular the double layer of charges and the built-in potential occur between any couple of materials both metals and semiconductors. The main difference in case of metals is the dimensions of the junction. The fact that no electric field can exist inside a metal limits the layer of charges displaced by the equilibrium to lie in bi-dimensional sheets located at the interface between the materials. The size of the junction is then extremely short, and electrons can cross it via tunnel effect.
The built-in potential among metals is of the same order of magnitude of the built-in potentials observed with semiconductors. However these potentials are canceled in any closed network so they do not affect the currents and the voltages in the circuit. Actually, the built-in potential depends on temperature so they are cancelled only if all the junctions of the network stay at the same temperature. In case of temperature gradients inside the circuit the built-in potentials are

no more canceled and they become observable. The thermoelectric phenomena are based on non homogenous temperature distribution in circuits. They are at the basis of important devices such as the thermocouple (a sensor to measure temperature differences) and the Peltier cell (an actuator to cool and to heat small masses).

3.4.2 Tunnel ohmic contacts

The previous section shown that to obtain a ohmic contact with a semiconductor is necessary a metal whose work function is in a particular relationship with the work function of the semiconductor. This condition is not commonly met. The work function of silicon with 10^{16} cm^{-3} donors is about 4.25 eV, and this quantity is smaller than the work function of most of the metals of technological interest (see table 4 of chapter 1).
A more convenient approach to the fabrication of ohmic contacts is offered by the tunnel effect.
The tunnel effect is a typical quantum phenomenon contradicting the classical physics beliefs. It states that given a energy barrier, a particle whose energy is smaller than the height of the barrier has a non zero probability to pass through the barrier.
The probability is proportional to the product of the height of the barrier and its width. In practice, if the barrier is confined in a very short space, of the order of nanometers, the probability to pass through becomes significant.
A very narrow Schottky barrier can be obtained with a highly doped semiconductor. Equation 16 shows that the size of the depletion layer is inversely proportional to the concentration of mobile charges. Then, regardless the difference between the work function, the junction between a metal and a heavily doped semiconductor always results in a ohmic contact because the depletion layer is so narrow that the electrons can be transferred from one material to the other even if their energy is smaller than the energy at the top of the barrier.
A ohmic tunnel contact in silicon becomes possible when the doping concentration is larger than 10^{19} cm^{-3}. This is about the limit of degeneracy beyond which the Fermi level is almost inside the conduction band (in case of N-type) of the valence band (in case of P-type). Note that with such a doping the Fermi-Dirac function is no more approximated by the Boltzmann equation and then the equation used to calculate the device properties are not valid.
The ohmic contact requires a thin layer of heavy doped semiconductor in contact with the metal. In this way the same metal giving rise to a rectifying junction can be used for a ohmic contact.
The junction between the normal semiconductor and the heavily doped layer (indicated with a superscript + or - according to the kind of doping) gives rise to an accumulation layer in the normal semiconductor and then it behaves as a ohmic contact. the situation in terms of the shape of the conductance band at the equilibrium is shown in figure 3.18.

Figure 3.19 shows a realistic planar configuration of a Schottky diode where the same metal is used for the diode and the ohmic contact.

3.4.3 Space charge limited current

In the previous discussion of charge transport, it has been assumed that the current in the semiconductor is made by the mobile charges in the conductiion and in the valence band. In this condition, the flow of current does not alter the charge equilibrium in the semiconductor. Then the total charge density remains zero and, according to the Poisson equation, the derivative of the electric field is null and the electric field is constant in the semiconductor. This assumption lead to the Ohm's law

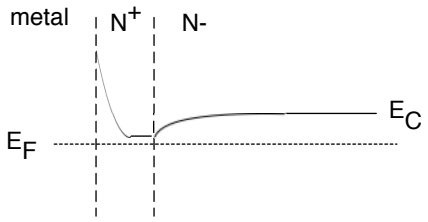


Fig. 3.18. *Conductance band at the equilibrium of the system: metal-N^+-N semiconductor. The depletion layer at the metal-semiconductor interface is narrow enough to be crossed by tunnel effect, while at the junction between semiconductors an accumulation layer is formed. Due to large concentration of electrons in the N^+ material the lost of electrons towards the N semiconductor is negligible.*

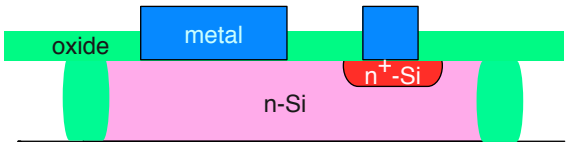


Fig. 3.19. *Realistics scheme of a Schottky diode in planar technology. the deep oxide layers insulate the device from the rest of the wafer.*

where the current and the voltage is proportional.
However, in a structure such as the $N^+ - N$ junction used for ohmic contacts, the N^+ layer being more heavily doped can, inject at sufficiently high voltage, a concentration of charges larger than the donor density in the N-type material. In this condition, the density of total charge is the semiconductor is no more null, and a space charge region takes place in the semiconductor. When the injected charges is much larger that N_D the density of charge is: $\rho = q(N_D - n) \approx -qn$. Then the Poisson equation can be written as:

$$\frac{d^2\Phi}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{qn}{\epsilon_s} \tag{3.63}$$

Replacing, n in the current density definition and in the small electric field regime (so that $v = \mu\mathcal{E}$) we get:

$$j = qnv = \epsilon_s \mu \mathcal{E} \frac{d\mathcal{E}}{dx} \tag{3.64}$$

which is integrated as:

$$\frac{j}{\epsilon_s \mu} \int_0^L dx = \int_0^{\mathcal{E}} \mathcal{E}_{max} \frac{d\mathcal{E}}{dx} \tag{3.65}$$

where L is the lenght of the N-type semiconductor and $\mathcal{E}_{max}$ is the largest electric field at the edge of the semiconductor: $\mathcal{E}_{max} = -\frac{V}{L}$. Solving the integrals and replacing $\mathcal{E}_{max}$ with the applied voltage we get:

$$j = \frac{1}{2}\frac{\epsilon_s \mu}{L}\mathcal{E}_{max}^2 = \frac{1}{2}\frac{\epsilon_s \mu}{L^3}V^2 \tag{3.66}$$

A more accurate calculation, results in

$$j = \frac{9}{8}\frac{\epsilon_s \mu}{L^3}V^2 \tag{3.67}$$

This is called the Mott-Guerney law. It expresses the deviation from the linear Ohm's law in semiconductors where a mobile charge larger than the equilibrium charge is injected. This behaviour is not limited to $N^+ - N$ junctions, but it may also be manifested in ohmic Schottky contacts where the metal injects charges in the semiconductor or under a localized injection of photogenerated charges.

3.5 Surface states

The metal-semiconductor junction takes place at the surface of the semiconductor where the metallic layer is deposited. In this portion of the material we cannot neglect the fact that the surface of semiconductors is very different from the bulk. Indeed disregarding the impurities that can be accumulated on the surface, the regular pattern of atoms arranged according to the crystalline structure is broken at the surface where part of the bonds of the last layer of atoms are used to bind the lateral atoms.The nature of these bonds is obviously different from the bonds of the periodic crystal, then the energy of these states is different with respect to the conduction and valence bands. Thus, a number of additional electronic states appears in proximity of the surface. Actually, the modifications start few atomic layer below the surface.
The density of these states is roughly equal to the density of the surface atoms. Then if n_0 is the concentration of atoms per volume, the density at atoms at the surface is $n_0^{\frac{2}{3}}$. In the case of silicon, there are about $5 \cdot 10^{22} \frac{atoms}{cm^3}$ and then the density of atoms at the surface and the surface states, is about $10^{15} \frac{atoms}{cm^2}$.
The most important of these states are the so called Tamm-Shockley states whose energy falls in the energy gap of the semiconductor. The maximum of density of surface states occurs at about one third of the energy gap. Noteworthy, it the semiconductor is N-type, the Tamm-Shockley states lies surely below the fermi level. Then at the surface, the electrons provided by the doping instead of populating the conduction band are segregated in the surface states. This means that in proximity of the material the semiconductor is depleted of mobile charges. This gives rise to a bend-bending and a built-in potential that naturally occur at the surface of the semiconductor.

Furthermore, the density of the surface states is much greater than the density of surface dopant atoms. Indeed, if $N_D = 10^{17} \frac{atoms}{cm^3}$ the surface density of dopant atoms is about $10^{11} \frac{atoms}{cm^2}$. About four orders of magnitude smaller that the density of states. Thus most of the surface states are empty.
When such a semiconductor is used for a Schottky diode, the electrons necessary to equilibrate the Fermi level are actually provided by the surface states and the electrons from the semiconductor are still subjected to the built-in potential due to the surface states. In practice, the metal does not alter the built-in potential. In this condition the Fermi level is said to be pinned by the surface states. This condition makes the Schottky diode independent from the work function of the metal and the concentration of doping. Besides, since the actual distribution of surface states is unpredictable, the device cannot be properly designed.

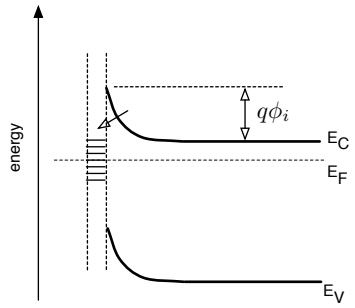


Fig. 3.20. *Surface band-bending due to a surface distribution of Tamm-Shockley states.*

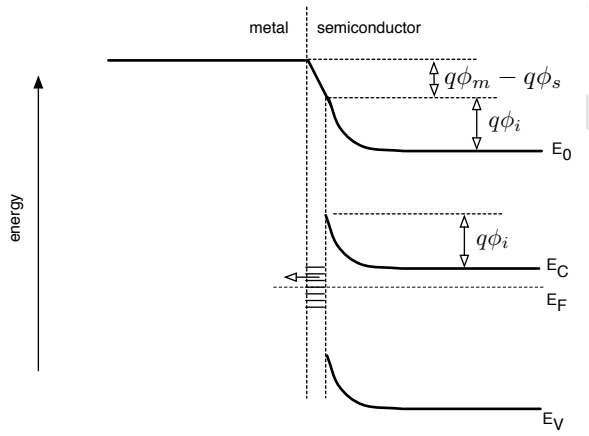


Fig. 3.21. *In case of a metal-semiconductor junction, the built-in potential due to the difference of work functions of metal and semiconductor vanishes in the surface region and the potential barrier applied to the electrons in the conductance band remains unchanged.*

3.6 Numerical example

In this section some of the equations describing the ideal Schottky diode are explicitly calculated in order to provide a quantitative evaluation of the developed model.
The example considers a junction formed by chromium and N-type silicon. The work function of chromium is about $q\Phi = 4.95\ eV$ and the silicon is uniformly doped with a density of donors equal to $N_D = 10^{-17}\ cm^{-3}$.
Figure 3.22 shows the equilibrium band diagram. In figure 3.23 the electrostatic quantities are plotted. The results of the Poisson equation have obviously been used to draw the band diagram. Finally, figure 3.24 shows the dependence on the applied voltage of the depletion layer width, the current and the junction capacitance. For sake of a more clear representation, the current is plotted in a logarithmic scale, then the reverse current appears positive and the origin is not plotted.

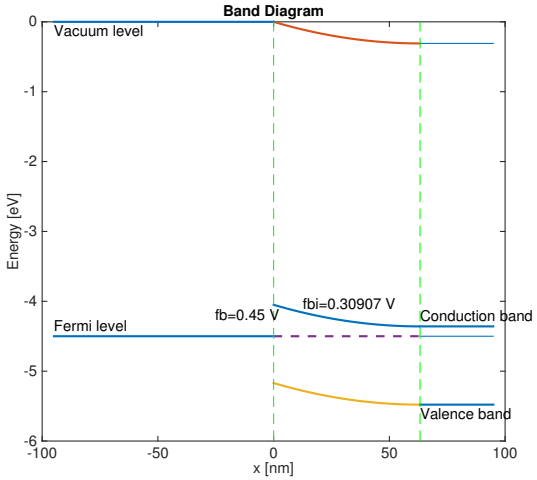


Fig. 3.22. *Equilibrium band diagram of a chromium-N-type silicon junction with* $N_D = 10^{-17}\ cm^{-3}$.

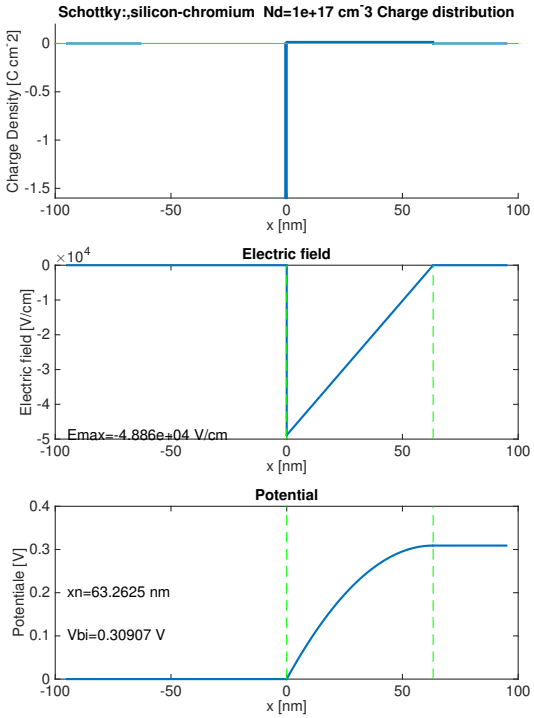


Fig. 3.23. *Behavior at the equilibrium of the charge density, the electric field and the potential for a chromium-N-type silicon junction with* $N_D = 10^{-17}\ cm^{-3}$.

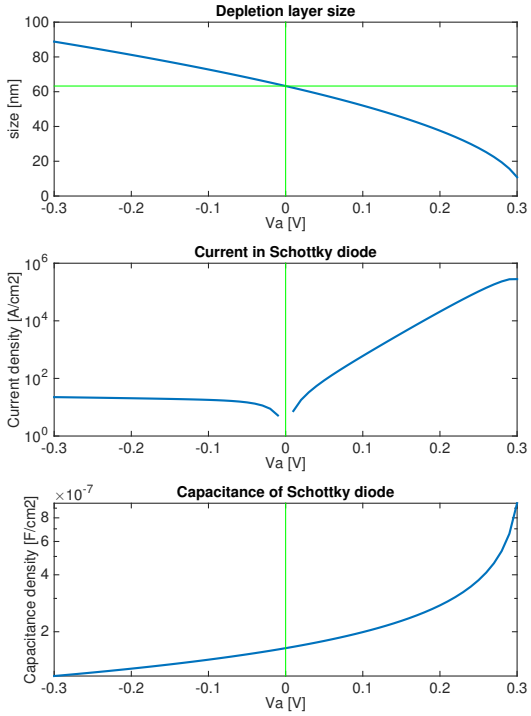


Fig. 3.24. *Space charge region width vs. applied voltage, current vs. applied voltage, and junction capacitance vs. applied voltage in a gold-N-type silicon junction with* $N_D = 10^{-16}\ cm^{-3}$. *The current is calculated with the thermionic model.*