Principles of Semiconductor Devices

Chapter 1: Review of Modern Physics

1.1 Introduction

The fundamentals of semiconductors are typically found in textbooks discussing quantum mechanics, electro-magnetics, solid-state physics and statistical thermodynamics. The purpose of this chapter is to review the physical concepts, which are needed to understand the semiconductor fundamentals of semiconductor devices. While an attempt was made to make this section comprehensible even to readers with a minimal background in the different areas of physics, readers are still referred to the bibliography for a more thorough treatment of this material. Readers with sufficient background in modern physics can skip this chapter without loss of continuity

Chapter 1: Review of Modern Physics

1.2 Quantum Mechanics

Quantum mechanics emerged in the beginning of the twentieth century as a new discipline because of the need to describe phenomena, which could not be explained using Newtonian mechanics or classical electromagnetic theory. These phenomena include the photoelectric effect, blackbody radiation and the rather complex radiation from an excited hydrogen gas. It is these and other experimental observations which lead to the concepts of quantization of light into photons, the particle-wave duality, the de Broglie wavelength and the fundamental equation describing quantum mechanics, namely the Schrödinger equation. This section provides an introductory description of these concepts and a discussion of the energy levels of an infinite one-dimensional quantum well and those of the hydrogen atom.

1.2.1 Particle-wave duality

Quantum mechanics acknowledges the fact that particles exhibit wave properties. For instance, particles can produce interference patterns and can penetrate or "tunnel" through potential barriers. Neither of these effects can be explained using Newtonian mechanics. Photons on the other hand can behave as particles with well-defined energy. These observations blur the classical distinction between waves and particles. Two specific experiments demonstrate the particle-like behavior of light, namely the photoelectric effect and blackbody radiation. Both can only be explained by treating photons as discrete particles whose energy is proportional to the frequency of the light. The emission spectrum of an excited hydrogen gas demonstrates that electrons confined to an atom can only have discrete energies. Niels Bohr explained the emission spectrum by assuming that the wavelength of an electron wave is inversely proportional to the electron momentum.

The particle and the wave picture are both simplified forms of the wave packet description, a localized wave consisting of a combination of plane waves with different wavelength. As the range of wavelength is compressed to a single value, the wave becomes a plane wave at a single frequency and yields the wave picture. As the range of wavelength is increased, the size of the wave packet is reduced, yielding a localized particle.

1.2.2 The photo-electric effect

The photoelectric effect is by now the "classic" experiment, which demonstrates the quantized nature of light: when applying monochromatic light to a metal in vacuum one finds that electrons are released from the metal. This experiment confirms the notion that electrons are confined to the metal, but can escape when provided sufficient energy, for instance in the form of light. However, the surprising fact is that when illuminating with long wavelengths (typically larger than 400 nm) no electrons are emitted from the metal even if the light intensity is increased. On the other hand, one easily observes electron emission at ultra-violet wavelengths for which the number of electrons emitted does vary with the light intensity. A more detailed analysis reveals that the maximum kinetic energy of the emitted electrons varies linearly with the inverse of the wavelength, for wavelengths shorter than the maximum wavelength.

The experiment is illustrated with Figure 1.2.1:
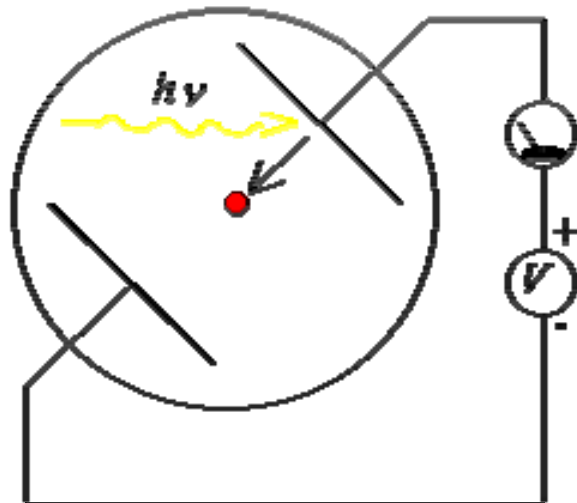
Figure 1.2.1.:
Experimental set-up to measure the photoelectric effect.

The experimental apparatus consists of two metal electrodes within a vacuum chamber. Light is incident on one of two electrodes to which an external voltage is applied. The external voltage is adjusted so that the current due to the photo-emitted electrons becomes zero. This voltage corresponds to the maximum kinetic energy, *K.E.*, of the electrons in units of electron volt. That voltage is measured for different wavelengths and is plotted as a function of the inverse of the wavelength as shown in Figure 1.2.2. The resulting graph is a straight line.
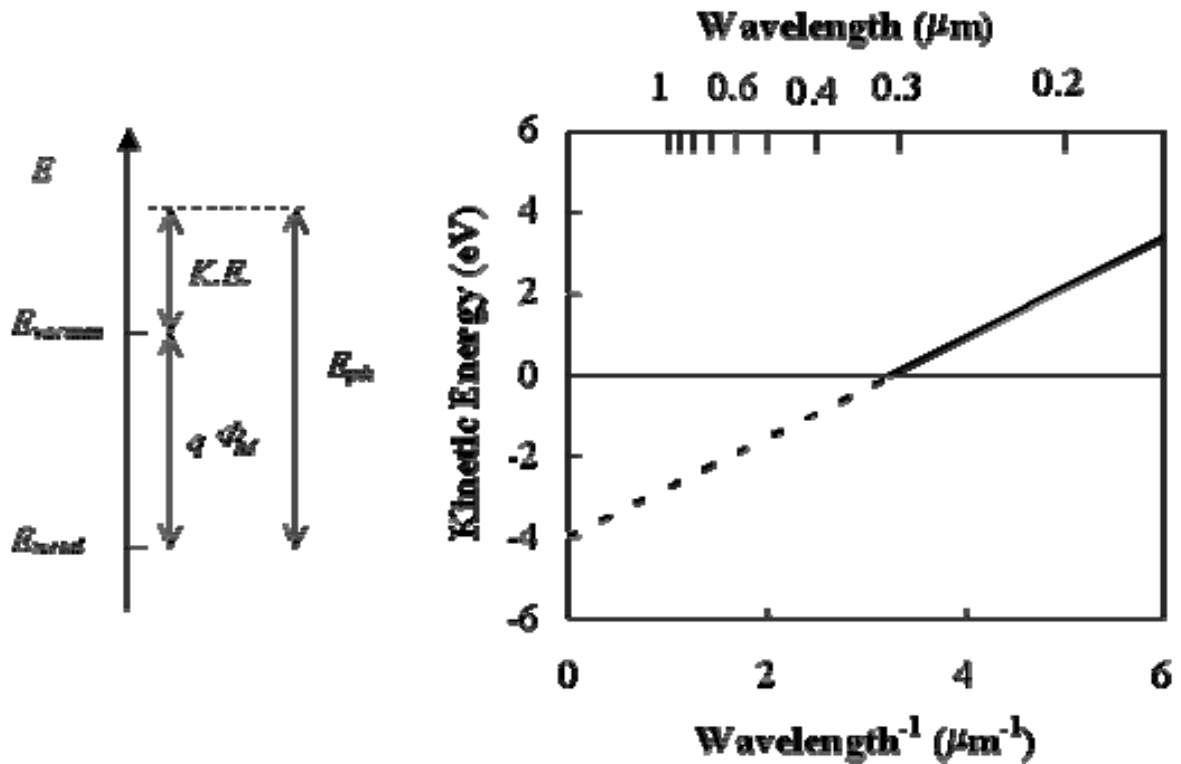
Figure 1.2.2 :
Maximum kinetic energy, *K.E.*, of electrons emitted from a metal upon illumination with photon energy, E_{ph}. The energy is plotted versus the inverse of the wavelength of the light.

Albert Einstein explained this experiment by postulating that the energy of light is quantized. He assumed that light consists of individual particles called photons, so that the kinetic energy of the electrons, *K.E.*, equals the energy of the photons, E_{ph}, minus the energy, $q\Phi_M$, required to extract the electrons from the metal. The work function, Φ_M, therefore quantifies the potential, which the electrons have to overcome to leave the metal. The slope of the curve was measured to be 1.24 eV/micron, which yielded the following relation for the photon energy, E_{ph}:

$$E_{ph} = h\nu = \frac{hc}{\lambda} \qquad\qquad (1.2.1)$$

where h is Planck's constant, ν is the frequency of the light, c is the speed of light in vacuum and λ is the wavelength of the light.

While other light-related phenomena such as the interference of two coherent light beams demonstrate the wave characteristics of light, it is the photoelectric effect, which demonstrates the particle-like behavior of light. These experiments lead to the particle-wave duality concept, namely that particles observed in an appropriate environment behave as waves, while waves can also behave as particles. This concept

applies to all waves and particles. For instance, coherent electron beams also yield interference patterns similar to those of light beams.

It is the wave-like behavior of particles, which led to the de Broglie wavelength: since particles have wave-like properties, there is an associated wavelength, which is

called the de Broglie wavelength and is given by:
$$\lambda = \frac{h}{p}$$

(1.2.2)

where λ is the wavelength, h is Planck's constant and p is the particle momentum. This expression enables a correct calculation of the ground energy of an electron in a hydrogen atom using the Bohr model described in Section 1.2.4. One can also show that the same expression applies to photons by combining equation (1.2.1) with $E_{ph} = p\ c$.

Example   1.1

A metal has a workfunction of 4.3 V. What is the minimum photon energy in Joule to emit an electron from this metal through the photo-electric effect? What are the photon frequency in Terahertz and the photon wavelength in micrometer? What is the corresponding photon momentum? What is the velocity of a free electron with the same momentum?
Solution

The minumum photon energy, E_{ph}, equals the workfunction, Φ_M, in units of electron volt or 4.3 eV. This also equals:

$$E_{ph} = q\Phi_M = 1.6\times10^{-19} \times 4.3 = 6.89\times10^{-19} \text{ Joule}$$

The corresponding photon frequency is:

$$\nu = \frac{6.89\times10^{-19}}{6.626\times10^{-34}} = 1040 \text{ THz}$$

The corresponding wavelength equals:

$$\lambda = \frac{hc}{E_{ph}} = \frac{6.626\times10^{-34} \times 3\times10^8}{6.89\times10^{-19}} = \frac{1.24 \ \mu m}{E_{ph} \ (eV)} = 0.288 \ \mu m$$

The photon momentum, p, is:

$$p = \frac{h}{\lambda} = \frac{6.626 \times 10^{-34}}{0.288 \times 10^{-6}} = 2.297 \times 10^{-27} \frac{kg \, m}{s}$$

And the velocity, v, of a free electron with the same momentum equals:

$$v = \frac{p}{m_0} = \frac{2.297 \times 10^{-27}}{9.11 \times 10^{-31}} = 2522 \, m/s$$

Where m_0 is the free electron mass.

$$E_{ph} = q\Phi_M = 1.6 \times 10^{-19} \times 4.3 = 6.89 \times 10^{-19} \, Joule$$

1.2.3 Blackbody radiation

Another experiment which could not be explained without quantum mechanics is the blackbody radiation experiment: By heating an object to high temperatures one finds that it radiates energy in the form of infra-red, visible and ultra-violet light. The appearance is that of a red glow at temperatures around 800° C which becomes brighter at higher temperatures and eventually looks like white light. The spectrum of the radiation is continuous, which led scientists to initially believe that classical electro-magnetic theory should apply. However, all attempts to describe this phenomenon failed until Max Planck developed the blackbody radiation theory based on the assumption that the energy associated with light is quantized and the energy quantum or photon energy equals:
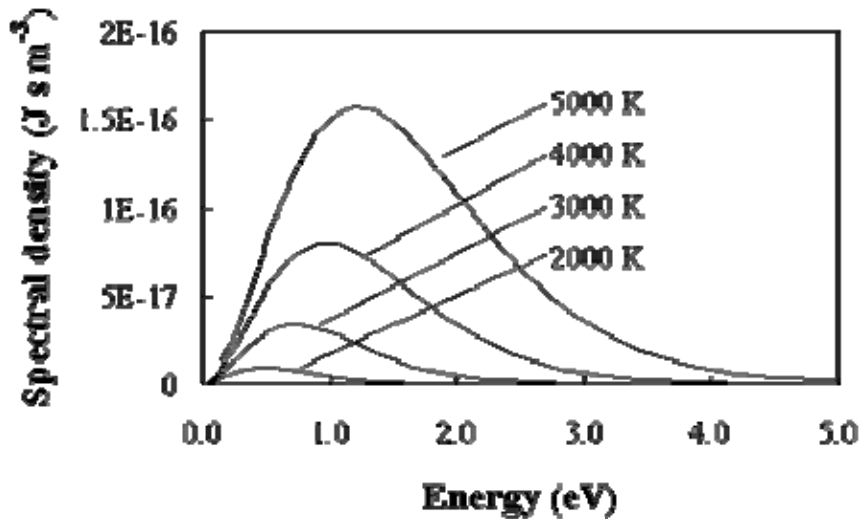
$$\mathbf{E_{ph}} = \mathbf{h\nu} = \hbar\omega \qquad\qquad (1.2.3)$$

Where $\hbar$ is the reduced Planck's constant (= $h/2\pi$), and ω is the radial frequency (= $2\pi \nu$).

The spectral density, u_ω, or the energy density per unit volume and per unit frequency is given by:

$$u_\omega = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{\exp(\frac{\hbar\omega}{kT}) - 1}$$

$$(1.2.4)$$

Where k is Boltzmann's constant and T is the temperature. The spectral density is shown versus energy in Figure 1.2.3.

Figure 1.2.3:
Spectral density of a blackbody at 2000, 3000, 4000 and 5000 K versus energy.

The peak value of the blackbody radiation occurs at 2.82 *kT* and increases with the third power of the temperature. Radiation from the sun closely fits that of a black body at 5800 K.

Example   1.2

The spectral density of the sun peaks at a wavelength of 900 nm. If the sun behaves as a black body, what is the temperature of the sun?
Solution

A wavelength of 900 nm corresponds to a photon energy of:

$$E_{ph} = \frac{hc}{\lambda} = \frac{6.626 \times 10^{-34} \times 3 \times 10^8}{900 \times 10^{-9}} = 2.21 \times 10^{-19} \; Joule$$

Since the peak of the spectral density occurs at 2.82 *kT*, the corresponding temperature equals:

$$T = \frac{E_{ph}}{2.82 k} = \frac{2.21 \times 10^{-19}}{2.82 \times 1.38 \times 10^{-23}} = 5672 \; Kelvin$$

1.2.4 The Bohr model

The spectrum of electromagnetic radiation from an excited hydrogen gas was yet another experiment, which was difficult to explain since it is discreet rather than continuous. The emitted wavelengths were early on associated with a set of discreet energy levels E_n described by:

$$E_n = -\frac{m_0 q^4}{8 \varepsilon_0^2 h^2 n^2}, \text{ with } n = 1, 2, \ldots$$

(1.2.5)

and the emitted photon energies equal the energy difference released when an electron makes a transition from a higher energy E_i to a lower energy E_j.

$$E_{ij} = 13.6 \text{ eV}\left(\frac{1}{j^2} - \frac{1}{i^2}\right), \text{ with } i > j$$

(1.2.6)

The maximum photon energy emitted from a hydrogen atom equals 13.6 eV. This energy is also called one Rydberg or one atomic unit. The electron transitions and the resulting photon energies are further illustrated by Figure 1.2.4.
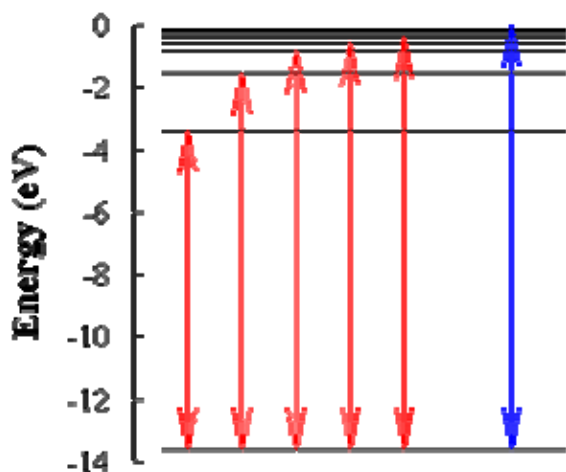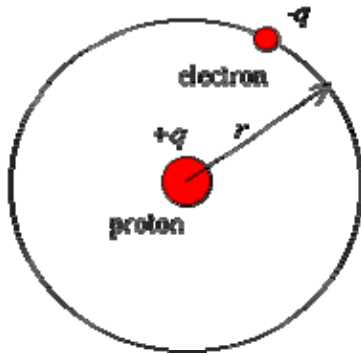

Figure 1.2.4 :
Energy levels and possible electronic transitions in a hydrogen atom. Shown are the first six energy levels, as well as six possible transitions involving the lowest energy level ($n = 1$)

However, there was no explanation why the possible energy values were not continuous. No classical theory based on Newtonian mechanics could provide such spectrum. Further more, there was no theory, which could explain these specific values.

Niels Bohr provided a part of the puzzle. He assumed that electrons move along a circular trajectory around the proton like the earth around the sun, as shown in Figure 1.2.5.

Figure 1.2.5:
Trajectory of an electron in a hydrogen atom as used in the Bohr model.

He also assumed that electrons behave within the hydrogen atom as a wave rather than a particle. Therefore, the orbit-like electron trajectories around the proton are limited to those with a length, which equals an integer number of wavelengths so that

$$2\pi r = n\lambda$$
(1.2.7)

where r is the radius of the circular electron trajectory and n is a positive integer. The Bohr model also assumes that the momentum of the particle is linked to the de Broglie wavelength (equation (1.2.2))

The model further assumes a circular trajectory and that the centrifugal force equals the electrostatic force, or:

$$m\frac{v^2}{r} = \frac{q^2}{4\pi\varepsilon_0 r^2} \qquad (1.2.8)$$

Solving for the radius of the trajectory one finds the Bohr radius, a_0:

$$a_0 = \frac{\varepsilon_0 h^2 n^2}{\pi m_0 q^2} \qquad (1.2.9)$$

and the corresponding energy is obtained by adding the kinetic energy and the potential energy of the particle, yielding:

$$E_n = -\frac{m_0 q^4}{8\varepsilon_0^2 h^2 n^2} \quad \text{where n= 1, 2, 3,} \qquad (1.2.10)$$

Where the potential energy is the electrostatic potential of the proton:

$$V(\mathbf{r}) = -\frac{q^2}{4\pi\varepsilon_o \mathbf{r}} \qquad\qquad (1.2.11)$$

Note that all the possible energy values are negative. Electrons with positive energy are not bound to the proton and behave as free electrons.

The Bohr model does provide the correct electron energies. However, it leaves many unanswered questions and, more importantly, it does not provide a general method to solve other problems of this type. The wave equation of electrons presented in the next section does provide a way to solve any quantum mechanical problem.

1.2.5 Schrödinger's equation

A general procedure to solve quantum mechanical problems was proposed by Erwin Schrödinger. Starting from a classical description of the total energy, E, which equals the sum of the kinetic energy, $K.E.$, and potential energy, V, or:

$$E = T + V = \frac{p^2}{2m} + V(x)$$

(1.2.12)

He converted this expression into a wave equation by defining a wavefunction, Ψ, and multiplied each term in the equation with that wavefunction:

$$E\Psi = \frac{p^2}{2m}\Psi + V(x)\Psi$$

(1.2.13)

To incorporate the de Broglie wavelength of the particle we now introduce the operator, $-\dfrac{\hbar^2 \partial^2}{\partial x^2}$, which provides the square of the momentum, p, when applied to a plane wave:

$$-\hbar^2 \frac{\partial^2 \Psi}{\partial x^2} = \hbar^2 k^2 \Psi = p^2 \Psi \ \text{ for } \Psi = e^{i(kx - \omega t)}$$

(1.2.14)

Where k is the wavenumber, which equals $2\pi/\lambda$. Without claiming that this is an actual proof we now simply replace the momentum squared, p^2, in equation (1.2.13) by this operator yielding the time-independent Schrödinger equation.

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi(x)}{dx^2} + V(x)\Psi(x) = E\Psi(x)$$

(1.2.15)

To illustrate the use of Schrödinger's equation, we present two solutions of Schrödinger's equation, that for an infinite quantum well and that for the hydrogen atom. Prior to that, we discuss the physical interpretation of the wavefunction.

1.2.5.1. Physical interpretation of the wavefunction

The use of a wavefunction to describe a particle, as in the Schrödinger equation, is consistent with the particle-wave duality concept. However, the physical meaning of the wavefunction does not naturally follow. Quantum theory postulates that the wavefunction, $\Psi(x)$, multiplied with its complex conjugate, $\Psi^*(x)$, is proportional to the probability density function, $P(x)$, associated with that particle

$$P(x) = \Psi(x)\Psi^*(x)$$

(1.2.16)

This probability density function integrated over a specific volume provides the probability that the particle described by the wavefunction is within that volume. The probability function is frequently normalized to indicate that the probability of finding the particle somewhere equals 100%. This normalization enables to calculate the magnitude of the wavefunction using:

$$\int_{-\infty}^{\infty} P(x)\,dx = 1$$

(1.2.17)

This probability density function can then be used to find all properties of the particle by using the quantum operators. To find the expected value of a function $f(x,p)$ for the particle described by the wavefunction, one calculates:

$$\langle f(x,p) \rangle = \int_{-\infty}^{\infty} \Psi(x)F(x)\Psi^*(x)\,dx$$

(1.2.18)

Where $F(x)$ is the quantum operator associated with the function of interest. A list of quantum operators corresponding to a selection of common classical variables is provided in Table 1.2.1.

	Classical variable	Quantum operator
Position	x	x
A function which depends only on position	$f(x)$	$F(x)$
Momentum	p	$\dfrac{\hbar}{i}\dfrac{\partial}{\partial x}$
Energy	E	$-\dfrac{\hbar}{i}\dfrac{\partial}{\partial t}$

Table 1.2.1:
Selected classical variables and the corresponding quantum operator.

1.2.5.2. The infinite quantum well

The one-dimensional infinite quantum well represents one of the simplest quantum mechanical structures. We use it here to illustrate some specific properties of quantum mechanical systems. The potential in an infinite well is zero between $x = 0$ and $x = L_x$ and is infinite on either side of the well. The potential and the first five possible energy levels an electron can occupy are shown in Figure 1.2.6:
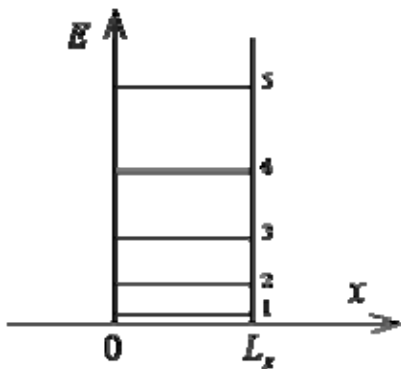
Figure 1.2.6 :
Potential energy of an infinite well, with width L_x. Also indicated are the lowest five energy levels in the well.

The energy levels in an infinite quantum well are calculated by solving Schrödinger's equation 1.2.15 with the potential, $V(x)$, as shown in Figure 1.2.6. As a result one solves the following equation within the well.

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi(x)}{dx^2} = E\Psi(x) \text{ for } 0 < x < L_x$$

(1.2.19)

The general solution to this differential equation is:

$$\Psi(x) = A\sin(\frac{\sqrt{2mE}}{\hbar}x) + B\cos(\frac{\sqrt{2mE}}{\hbar}x) \text{ for } 0 < x < L_x$$

(1.2.20)

Where the coefficients A and B must be determined by applying the boundary conditions. Since the potential is infinite on both sides of the well, the probability of finding an electron outside the well and at the well boundary equals zero. Therefore the wave function must be zero on both sides of the infinite quantum well or:

$$\Psi(0) = 0 \text{ and } \Psi(L_x) = 0$$

(1.2.21)

These boundary conditions imply that the coefficient B must be zero and the argument of the sine function must equal a multiple of pi at the edge of the quantum well or:

$$\frac{\sqrt{2mE_n}}{\hbar}L_x = n\pi, \text{ with } n = 1, 2, ...$$

(1.2.22)

Where the subscript n was added to the energy, E, to indicate the energy corresponding to a specific value of, n. The resulting values of the energy, E_n, are then equal to:

$$E_n = \frac{\hbar^2}{2m^*}(\frac{n}{2L_x})^2, \text{ with } n = 1, 2, ...$$

(1.2.23)

The corresponding normalized wave functions, $\Psi_n(x)$, then equal:

$$\Psi_n(x) = \sqrt{\frac{2}{L_x}}\sin(\frac{\sqrt{2mE_n}}{\hbar}x) \text{ for } 0 < x < L_x$$

(1.2.24)

where the coefficient A was determined by requiring that the probability of finding the electron in the well equals unity or:

$$\int_0^{L_x} \Psi_n(x)\Psi_n^*(x)dx = 1$$

(1.2.25)

The asterisk denotes the complex conjugate.

Note that the lowest possible energy is not zero although the potential is zero within the well. Only discreet energy values are obtained as eigenvalues of the Schrödinger equation. The energy difference between adjacent energy levels increases as the energy increases. An electron occupying one of the energy levels can have a positive or negative spin ($s = 1/2$ or $s = -1/2$). Both quantum numbers, n and s, are the only two quantum numbers needed to describe this system.

The wavefunctions corresponding to each energy level are shown in Figure 1.2.7 (a). Each wavefunction has been shifted by the corresponding energy. The probability density function, calculated as $|\Psi|^2$, provides the probability of finding an electron in a certain location in the well. These probability density functions are shown in Figure 1.2.7 (b) for the first five energy levels. For instance, for $n = 2$ the electron is least likely to be in the middle of the well and at the edges of the well. The electron is most likely to be one quarter of the well width away from either edge.
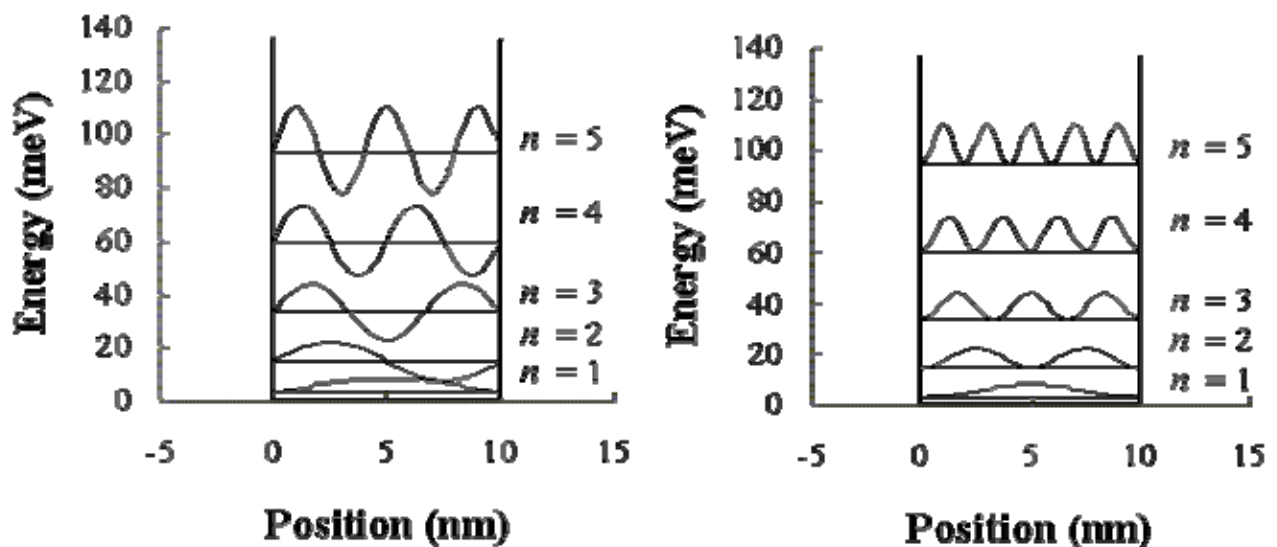
Figure 1.2.7 :
Energy levels, wavefunctions (left) and probability density functions (right) in an infinite quantum well. The figure is calculated for a 10 nm wide well containing an electron with mass m_0. The wavefunctions and the probability density functions are not normalized and shifted by the corresponding electron energy.

Example   1.3

An electron is confined to a 1 micron thin layer of silicon. Assuming that the semiconductor can be adequately described by a one-dimensional quantum well with infinite walls, calculate the lowest possible energy within the material in units of electron volt. If the energy is interpreted as the kinetic energy of the electron, what is the corresponding electron velocity? (The effective mass of electrons in silicon is 0.26 m_0, where $m_0 = 9.11 \times 10^{-31}$ kg is the free electron rest mass).
Solution

The lowest energy in the quantum well equals:

$$E_1 = \frac{\hbar^2}{2m^*} \cdot \left(\frac{1}{2L_x}\right)^2 = \frac{(6.626 \times 10^{-34})^2}{2 \times 0.26 \times 9.11 \times 10^{-31}}\left(\frac{1}{2 \times 10^{-8}}\right)^2$$

= 2.32 x 10^-25 Joules = 1.45 meV

The velocity of an electron with this energy equals:

$$v = \sqrt{\frac{2E_1}{m^*}} = \sqrt{\frac{2 \times 2.32 \times 10^{-25}}{0.26 \times 9.11 \times 10^{-31}}}$$ =1.399 km/s

1.2.5.3. The hydrogen atom

The hydrogen atom represents the simplest possible atom since it consists of only one proton and one electron. Nevertheless, the solution to Schrödinger's equation as applied to the potential of the hydrogen atom is rather complex due to the three-dimensional nature of the problem. The potential, V(r) (equation (1.2.11)), is due to the electrostatic force between the positively charged proton and the negatively charged electron.

$$V(\bar{r}) = \frac{q}{4\pi\varepsilon_0 r}$$

(1.2.26)

The energy levels in a hydrogen atom can be obtained by solving Schrödinger's equation in three dimensions.

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(x,y,z) + V(x,y,z)\Psi(x,y,z) = E\Psi(x,y,z)$$

(1.2.27)

The potential V(x,y,z) is the electrostatic potential, which describes the attractive force between the positively charged proton and the negatively charged electron. Since this potential depends on the distance between the two charged particles one typically assumes that the proton is placed at the origin of the coordinate system and the position of the electron is indicated in polar coordinates by its distance r from the origin, the polar angle θ and the azimuthal angle ϕ.

Schrödinger's equation becomes:

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\bar{r}) + \frac{q}{4\pi\varepsilon_0 r}\Psi(\bar{r}) = E\Psi(\bar{r})$$

(1.2.28)

A more refined analysis includes the fact that the proton moves as the electron circles around it, despite its much larger mass. The stationary point in the hydrogen atom is the center of mass of the two particles. This refinement can be included by replacing the electron mass, m, with the reduced mass, m_r, which includes both the electron and proton mass:

$$\frac{1}{m_r} = \frac{1}{m_{electron}} + \frac{1}{m_{proton}}$$

(1.2.29)

Schrödinger's equation is then solved by using spherical coordinates, resulting in:

$$-\frac{\hbar^2}{2m_r}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial \phi^2}\right]\Psi(r,\theta,\phi)$$

$$+\frac{q}{4\pi\varepsilon_0 r}\Psi(r,\theta,\phi) = E\Psi(r,\theta,\phi)$$

(1.2.30)

In addition, one assumes that the wavefunction, $\Psi(r,\theta,\phi)$, can be written as a product of a radial, angular and azimuthal angular wavefunction, $R(r)$, $\Theta(\theta)$ and $\Phi(\phi)$. This assumption allows the separation of variables, i.e. the reformulation of the problem into three different differential equations, each containing only a single variable, r, θ or ϕ:

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR(r)}{dr}\right) + \left(\frac{2m_r}{\hbar^2}\left[E - \frac{q}{4\pi\varepsilon_0 r}\right] - \frac{A}{r^2}\right)R(r) = 0$$

(1.2.31)

$$\frac{d^2\Phi(\phi)}{d\phi^2} + B\Phi(\phi) = 0$$

(1.2.32)

$$\frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial\Theta(\theta)}{\partial \theta}\right) + \left(A - \frac{B}{\sin^2\theta}\right)\Theta(\theta) = 0$$

(1.2.33)

Where the constants A and B are to be determined. The solution to these differential equations is beyond the scope of this text. Readers are referred to the bibliography for an in depth treatment. We will now examine and discuss the solution.
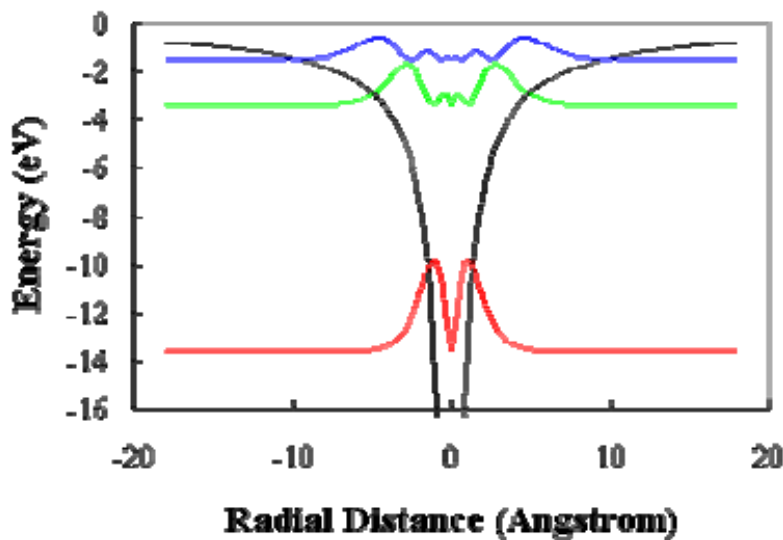
The electron energies in the hydrogen atom as obtained from equation (1.2.31) are:

$$E_n = -\frac{m_r q^4}{8 \varepsilon_0^2 h^2 n^2} \text{, with } n = 1, 2, \ldots$$

(1.2.34)

Where n is the principal quantum number.

This potential as well as the first three probability density functions ($r^2|\Psi|^2$) of the radially symmetric wavefunctions ($l = 0$) is shown in Figure 1.2.8.

Figure 1.2.8 :
Potential energy, $V(x)$, in a hydrogen atom and first three probability densities with $l = 0$. The probability densities are shifted by the corresponding electron energy.

Since the hydrogen atom is a three-dimensional problem, three quantum numbers, labeled n, l, and m, are needed to describe all possible solutions to Schrödinger's equation. The spin of the electron is described by the quantum number s. The energy levels only depend on n, the principal quantum number and are given by equation (1.2.10). The electron wavefunctions however are different for every different set of quantum numbers. While a derivation of the actual wavefunctions is beyond the scope of this text, a list of the possible quantum numbers is needed for further discussion and is therefore provided in Table 1.2.1. For each principal quantum number n, all smaller positive integers are possible values for the angular momentum quantum number l. The quantum number m can take on all integers between l and $-l$, while s can be ½ or -½. This leads to a maximum of 2 unique sets of quantum numbers for all s orbitals ($l = 0$), 6 for all p orbitals ($l = 1$), 10 for all d orbitals ($l = 2$) and 14 for all f orbitals ($l = 3$).

	n	l	m	s
1s	1	0	0	$\frac{1}{2}, -\frac{1}{2}$
2s	2	0	0	$\frac{1}{2}, -\frac{1}{2}$
2p	2	1	1, 0, -1	$\frac{1}{2}, -\frac{1}{2}$
3s	3	0	0	$\frac{1}{2}, -\frac{1}{2}$
3p	3	1	1, 0, -1	$\frac{1}{2}, -\frac{1}{2}$
3d	3	2	2, 1, 0, -1, -2	$\frac{1}{2}, -\frac{1}{2}$
4s	4	0	0	$\frac{1}{2}, -\frac{1}{2}$
4p	4	1	1, 0, -1	$\frac{1}{2}, -\frac{1}{2}$
4d	4	2	2, 1, 0, -1, -2	$\frac{1}{2}, -\frac{1}{2}$
4f	4	3	3, 2, 1, 0, -1, -2, -3	$\frac{1}{2}, -\frac{1}{2}$

Table 1.2.2:
First ten orbitals and corresponding quantum numbers of a hydrogen atom

1.2.6 Pauli exclusion principle

Once the energy levels of an atom are known, one can find the electron configurations of the atom, provided the number of electrons occupying each energy level is known. Electrons are Fermions since they have a half integer spin. They must therefore obey the Pauli exclusion principle. This exclusion principle states that no two Fermions can occupy the same energy level corresponding to a unique set of quantum numbers *n*, *l*, *m* or s. The ground state of an atom is therefore obtained by filling each energy level, starting with the lowest energy, up to the maximum number as allowed by the Pauli exclusion principle.

1.2.7 Electronic configuration of the elements

The electronic configuration of the elements of the periodic table can be constructed using the quantum numbers of the hydrogen atom and the Pauli exclusion principle, starting with the lightest element hydrogen. Hydrogen contains only one proton and one electron. The electron therefore occupies the lowest energy level of the hydrogen atom, characterized by the principal quantum number *n* = 1. The orbital quantum number *l* equals zero and is referred to as an s orbital (not to be confused with the quantum number for spin, *s*). The s orbital can accommodate two electrons with opposite spin, but only one is occupied. This leads to the short-hand notation of $1s^1$ for the electronic configuration of hydrogen as listed in Table 1.2.2.

Helium is the second element of the periodic table. For this and all other atoms one still uses the same quantum numbers as for the hydrogen atom. This approach is justified since all atom cores can be treated as a single charged particle, which yields a potential very similar to that of a proton. While the electron energies are no longer the same as for the hydrogen atom, the electron wavefunctions are very similar and can be classified in the same way. Since helium contains two

electrons it can accommodate two electrons in the 1s orbital, hence the notation $1s^2$. Since the s orbitals can only accommodate two electrons, this orbital is now completely filled, so that all other atoms will have more than one filled or partially-filled orbital. The two electrons in the helium atom also fill all available orbitals associated with the first principal quantum number, yielding a filled outer shell. Atoms with a filled outer shell are called noble gases as they are known to be chemically inert.

Lithium contains three electrons and therefore has a completely filled 1s orbital and one more electron in the next higher 2s orbital. The electronic configuration is therefore $1s^2 2s^1$ or $[He]2s^1$, where [He] refers to the electronic configuration of helium. Beryllium has four electrons, two in the 1s orbital and two in the 2s orbital. The next six atoms also have a completely filled 1s and 2s orbital as well as the remaining number of electrons in the 2p orbitals. Neon has six electrons in the 2p orbitals, thereby completely filling the outer shell of this noble gas.

The next eight elements follow the same pattern leading to argon, the third noble gas. After that the pattern changes as the underlying 3d orbitals of the transition metals (scandium through zinc) are filled before the 4p orbitals, leading eventually to the fourth noble gas, krypton. Exceptions are chromium and zinc, which have one more electron in the 3d orbital and only one electron in the 4s orbital. A similar pattern change occurs for the remaining transition metals, where for the lanthanides and actinides the underlying f orbitals are filled first.

#	Element	Sym								
1	Hydrogen	H	$1s^1$							
2	Helium	He	$1s^2$							
3	Lithium	Li	$1s^2$	$2s^1$						
4	Beryllium	Be	$1s^2$	$2s^2$						
5	Boron	B	$1s^2$	$2s^2$	$2p^1$					
6	Carbon	C	$1s^2$	$2s^2$	$2p^2$					
7	Nitrogen	N	$1s^2$	$2s^2$	$2p^3$					
8	Oxygen	O	$1s^2$	$2s^2$	$2p^4$					
9	Fluorine	F	$1s^2$	$2s^2$	$2p^5$					
10	Neon	Ne	$1s^2$	$2s^2$	$2p^6$					
11	Sodium	Na	$1s^2$	$2s^2$	$2p^6$	$3s^1$				
12	Magnesium	Mg	$1s^2$	$2s^2$	$2p^6$	$3s^2$				
13	Aluminum	Al	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^1$			
14	Silicon	Si	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^2$			
15	Phosphorous	P	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^3$			
16	Sulfur	S	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^4$			
17	Chlorine	Cl	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^5$			
18	Argon	Ar	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$			
19	Potassium	K	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$		$4s^1$	
20	Calcium	Ca	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$		$4s^2$	
21	Scandium	Sc	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^1$	$4s^2$	
22	Titanium	Ti	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^2$	$4s^2$	
23	Vanadium	V	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^3$	$4s^2$	
24	Chromium	Cr	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^5$	$4s^1$	
25	Manganese	Mn	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^5$	$4s^2$	
26	Iron	Fe	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^6$	$4s^2$	
27	Cobalt	Co	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^7$	$4s^2$	
28	Nickel	Ni	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^8$	$4s^2$	
29	Copper	Cu	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^1$	
30	Zinc	Zn	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	
31	Gallium	Ga	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	$4p^1$
32	Germanium	Ge	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	$4p^2$
33	Arsenic	As	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	$4p^3$
34	Selenium	Se	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	$4p^4$
35	Brome	Br	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	$4p^5$
36	Krypton	Kr	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^6$	$3d^{10}$	$4s^2$	$4p^6$

Table 1.2.3:
Electronic configuration of the first thirty-six elements of the periodic table.

Chapter 1: Review of Modern Physics

1.3 Electromagnetic Theory

The analysis of most semiconductor devices includes the calculation of the electrostatic potential within the device as a function of the existing charge distribution. Electromagnetic theory and more specifically electrostatic theory are used to obtain the potential. A short description of the necessary tools, namely Gauss's law and Poisson's equation, is provided below.

1.3.1 Gauss's law

Gauss's law is one of Maxwell's equations ([Appendix 10](#)) and provides the relation between the charge density, ρ, and the electric field, $\mathcal{E}$. In the absence of time dependent magnetic fields the one-dimensional equation is given by:

$$\frac{d\mathcal{E}(x)}{dx} = \frac{\rho(x)}{\varepsilon}$$

(1.3.1)

This equation can be integrated to yield the electric field for a given one-dimensional charge distribution:

$$\mathcal{E}(x_2) - \mathcal{E}(x_1) = \int_{x_1}^{x_2} \frac{\rho(x)}{\varepsilon} \, dx$$

(1.3.2)

Gauss's law as applied to a three-dimensional charge distribution relates the divergence of the electric field to the charge density:

$$\nabla \vec{\mathcal{E}}(x, y, z) = \frac{\rho(x, y, z)}{\varepsilon}$$

(1.3.3)

This equation can be simplified if the field is constant on a closed surface, A, enclosing a charge Q, yielding:

$$\vec{\mathcal{E}} \cdot \vec{A} = \frac{Q}{\varepsilon}$$

(1.3.4)

Example   1.4

Consider an infinitely long cylinder with charge density r, dielectric constant ε_0 and radius r_0. What is the electric field in and around the cylinder?

Solution

Because of the cylinder symmetry one expects the electric field to be only dependent on the radius, r. Applying Gauss's law one finds:

$$\vec{\mathcal{E}} \cdot \vec{A} = \mathcal{E} 2 \pi r L = \frac{Q}{\varepsilon_0} = \frac{\rho \pi r^2 L}{\varepsilon_0}$$

and

$$\vec{\mathcal{E}} \cdot \vec{A} = \mathcal{E} 2 \pi r L = \frac{Q}{\varepsilon_0} = \frac{\rho \pi r_0^2 L}{\varepsilon_0}$$

where a cylinder with length L was chosen to define the surface A, and edge effects were ignored. The electric field then equals:

$$\mathcal{E}(r) = \frac{\rho r}{2 \varepsilon_0}$$

The electric field increases within the cylinder with increasing radius. The electric field decreases outside the cylinder with increasing radius.

1.3.2 Poisson's equation

Gauss's law is one of Maxwell's equations and provides the relation between the charge density, ρ, and the electric field, $\mathcal{E}$. In the absence of time dependent magnetic fields the one-dimensional equation is given by:

$$\frac{d \phi(x)}{dx} = -\mathcal{E}(x)$$

(1.3.5)

The electric field vector therefore originates at a point of higher potential and points towards a point of lower potential.

The potential can be obtained by integrating the electric field as described by:

$$\phi(x_2) - \phi(x_1) = -\int_{x_1}^{x_2} \mathcal{E}(x)\,dx$$

(1.3.6)

At times, it is convenient to link the charge density to the potential by combining equation (1.3.5) with Gauss's law in the form of equation (1.3.1), yielding:

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

(1.3.7)

which is referred to as Poisson's equation.

For a three-dimensional field distribution, the gradient of the potential as described by:

$$\vec{\nabla} \phi(x,y,z) = -\mathcal{E}(x,y,z)$$

(1.3.8)

can be combined with Gauss's law as formulated with equation (1.3.3), yielding a more general form of Poisson's equation:

$$\nabla^2 \phi(x,y,z) = -\frac{\rho(x,y,z)}{\sigma}$$

(1.3.9)

Chapter 1: Review of Modern Physics

1.4. Statistical Thermodynamics

Thermodynamics describes the behavior of systems containing a large number of particles. These systems are characterized by their temperature, volume, number and the type of particles. The state of the system is then further described by its total energy and a variety of other parameters including the entropy. Such a characterization of a system is much simpler than trying to keep track of each particle individually, hence its usefulness. In addition, such a characterization is general in nature so that it can be applied to mechanical, electrical and chemical systems.

The term thermodynamics is somewhat misleading as one deals primarily with systems in thermal equilibrium. These systems have constant temperature, volume and number of particles and their macroscopic parameters do not change over time, so that the dynamics are limited to the microscopic dynamics of the particles within the system.

Statistical thermodynamics is based on the fundamental assumption that all possible configurations of a given system, which satisfy the given boundary conditions such as temperature, volume and number of particles, are equally likely to occur. The overall system will therefore be in the statistically most probable configuration. The entropy of a system is defined as the logarithm of the number of possible configurations. While such definition does not immediately provide insight into the meaning of entropy, it does provide a straightforward analysis since the number of configurations can be calculated for any given system.

Classical thermodynamics provides the same concepts. However, they are obtained through experimental observation. The classical analysis is therefore more tangible compared to the abstract mathematical treatment of the statistical approach.

The study of semiconductor devices requires some specific results, which naturally emerge from statistical thermodynamics. In this section, we review basic thermodynamic principles as well as some specific results. These include the thermal equilibrium concept, the thermodynamic identity, the basic laws of thermodynamics, the thermal energy per particle and the Fermi function.

1.4.1. Thermal equilibrium

A system is in thermal equilibrium if detailed balance is obtained: i.e. every process in the system is exactly balanced by its inverse process so that there is no net effect on the system.

This definition implies that in thermal equilibrium no energy (heat, work or particle energy) is exchanged between the parts within the system or between the system and the environment. Thermal equilibrium is obtained by isolating a system from its environment, removing any internal sources of energy, and waiting for a long enough time until the system does not change any more.

The concept of thermal equilibrium is of interest since various thermodynamic results assume that the system under consideration is in thermal equilibrium. Few systems of interest rigorously satisfy this condition so that we often apply the thermodynamical results to systems that are "close" to thermal equilibrium. Agreement between theories based on this assumption and experiments justify this approach.

1.4.2. Laws of thermodynamics

If two systems are in thermal equilibrium with a third system, they must be in thermal equilibrium with each other.

1. Heat is a form of energy.
2. The second law can be stated either (a) in its classical form or (b) in its statistical form
    a. Heat can only flow from a higher temperature to a lower temperature.
    b. The entropy of a closed system tends to remain constant or increases monotonically over time.

Both forms of the second law could not seem more different. A more rigorous treatment proves the equivalence of both.

3. The entropy of a system approaches a constant as the temperature approaches zero Kelvin.

1.4.3. The thermodynamic identity

The thermodynamic identity states that a change in energy can be caused by adding heat, work or particles. Mathematically this is expressed by:

$$dU = dQ + dW + \mu dN$$
(1.4.1)

where U is the total energy, Q is the heat and W is the work. μ is the energy added to a system when adding one particle without adding either heat or work. This energy is also called the electro-chemical potential. N is the number of particles.

1.4.4. The Fermi energy

The Fermi energy, E_F, is the energy associated with a particle, which is in thermal equilibrium with the system of interest. The energy is strictly associated with the particle and does not consist even in part of heat or work. This same quantity is called the electro-chemical potential, μ, in most thermodynamics texts.

1.4.5. Some useful thermodynamics results

Listed below are two results, which will be used while analyzing semiconductor devices. The actual derivation is beyond the scope of this text.

1. The thermal energy of a particle, whose energy depends quadratically on its velocity, equals $kT/2$ per degree of freedom, where k is Boltzmann's constant. This thermal energy is a kinetic energy, which must be added to the potential energy of the particle, and any other kinetic energy. The thermal energy of a non-relativistic electron, which is allowed to move in three dimensions, equals $3/2\ kT$.

2. Consider an energy level at energy, E, which is in thermal equilibrium with a large system characterized by a temperature T and Fermi energy E_F. The probability that an electron occupies such energy level is given by:

$$f(E) = \frac{1}{1 + e^{(E - E_F)/kT}}$$

(1.4.2)

The function $f(E)$ is called the Fermi function and applies to all particles with half-integer spin. These particles, also called Fermions, obey the Pauli exclusion principle, which states that no two Fermions in a given system can have the exact same set of quantum numbers. Since electrons are Fermions, their probability distribution also equals the Fermi function.

Example   1.5

Calculate the energy relative to the Fermi energy for which the Fermi function equals 5%. Write the answer in units of kT.
Solution

The problems states that:

$$f(E) = \frac{1}{1 + \exp(\frac{E - E_F}{kT})} = 0.05$$

which can be solved yielding:

$$E - E_F = \ln(19)kT = 3kT$$

Chapter 2: Semiconductor Fundamentals

2.1 Introduction

To understand the fundamental concepts of semiconductors, one must apply modern physics to solid materials. More specifically, we are interested in semiconductor crystals. Crystals are solid materials consisting of atoms, which are placed in a highly ordered structure called a lattice. Such a structure yields a periodic potential throughout the material.

Two properties of crystals are of particular interest, since they are needed to calculate the current in a semiconductor. First, we need to know how many fixed and mobile charges are

present in the material. Second, we need to understand the transport of the mobile carriers through the semiconductor.

In this chapter we start from the atomic structure of semiconductors and explain the concepts of energy band gaps, energy bands and the density of states in an energy band. We also show how the current in an almost filled band can more easily be analyzed using the concept of holes. Next, we discuss the probability that energy levels within an energy band are occupied. We will use this probability density to find the density of electrons and holes in a band.

Two transport mechanisms will be considered. The drift of carriers in an electric field and the diffusion of carriers due to a carrier density gradient will be discussed. Recombination mechanisms and the continuity equations are then combined into the diffusion equation. Finally, we present the drift-diffusion model, which combines all the essential elements discussed in this chapter.

Chapter 2: Semiconductor Fundamentals

2.2. Crystals and crystal structures

Solid materials are classified by the way the atoms are arranged within the solid. Materials in which atoms are placed randomly are called amorphous. Materials in which atoms are placed in a high ordered structure are called crystalline. Poly-crystalline materials are materials with a high degree of short-range order and no long-range order. These materials consist of small crystalline regions with random orientation called grains, separated by grain boundaries.

Of primary interest in this text are crystalline semiconductors in which atoms are placed in a highly ordered structure. Crystals are categorized by their crystal structure and the underlying lattice. While some crystals have a single atom placed at each lattice point, most crystals have a combination of atoms associated with each lattice point. This combination of atoms is also called the basis.

The classification of lattices, the common semiconductor crystal structures and the growth of single-crystal semiconductors are discussed in the following sections.

2.2.1 Bravais lattices

The Bravais lattices are the distinct lattice types, which when repeated can fill the whole spa

The lattice can therefore be generated by three unit vectors, $\vec{a}_1, \vec{a}_2, \text{and } \vec{a}_3$ and a set of integers and *m* so that each lattice point, identified by a vector $\vec{r}$, can be obtained from:

$$\vec{r} = k\,\vec{a}_1 + l\,\vec{a}_2 + m\,\vec{a}_3$$

(2.2.1)

The construction of the lattice points based on a set of unit vectors is illustrated by Figure 2.2.

$\vec{r}_{32} = 3\vec{a}_1 + 2\vec{a}_2$
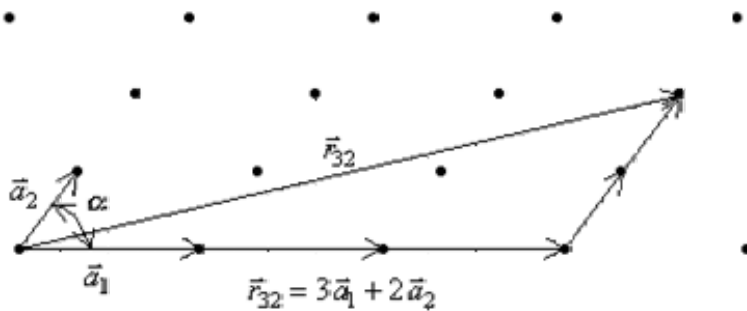
Figure 2.2.1:
The construction of lattice points using unit vectors

In two dimensions, there are five distinct Bravais lattices, while in three dimensions there fourteen. The lattices in two dimensions are the square lattice, the rectangular lattice, centered rectangular lattice, the hexagonal lattice and the oblique lattice as shown in Fig 2.2.2. It is customary to organize these lattices in groups which have the same symmetry. example is the rectangular and the centered rectangular lattice. As can be seen on the figure, the lattice points of the rectangular lattice can be obtained by a combination of the lattice vect . The centered rectangular lattice can be constructed in two ways. It can be obtained by start with the same lattice vectors as those of the rectangular lattice and then adding an additic atom at the center of each rectangle in the lattice. This approach is illustrated by Figure 2.2.2 The lattice vectors generate the traditional unit cell and the center atom is obtained by attach two lattice points to every lattice point of the traditional unit cell. The alternate approach is define a new set of lattice vectors, one identical to and another starting from the same origin ending on the center atom. These lattice vectors generate the so-called primitive cell and dire define the centered rectangular lattice.
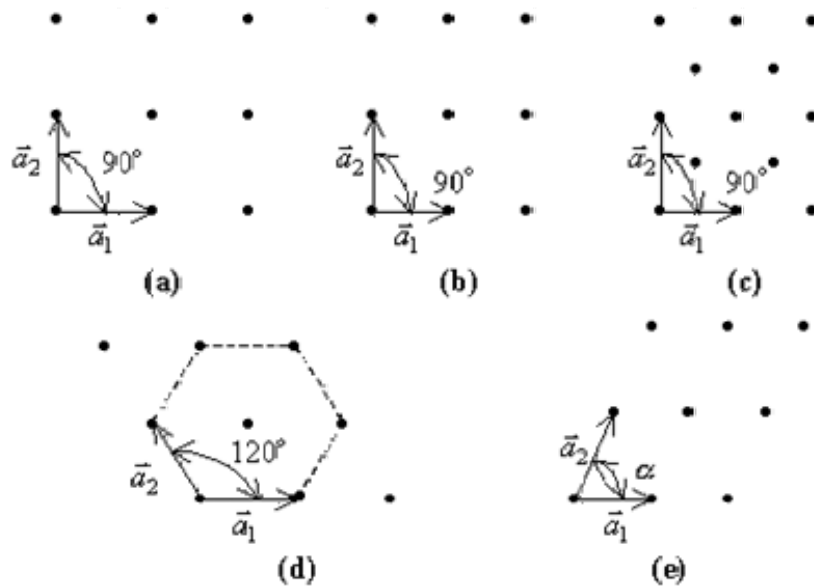
Figure 2.2.2.:
The five Bravais lattices of two-dimensional crystals: (a) cubic, (b) rectangular, (c) cente
rectangular, (d) hexagonal and (e) oblique

These lattices are listed in Table 2.2.1. a_1 and a_2 are the magnitudes of the unit vectors and
the angle between them.

Name	Number of Bravais lattices	Conditions
Square	1	$a_1 = a_2, \alpha = 90°$
Rectangular	2	$a_1 \neq a_2, \alpha = 90°$
Hexagonal	1	$a_1 = a_2, \alpha = 120°$
Oblique	1	$a_1 \neq a_2, \alpha \neq 120°, \alpha \neq 90°$

Table 2.2.1.:
Bravais lattices of two-dimensional crystals

The same approach is used for lattices in three dimensions. The fourteen lattices of th
dimensional crystals are classified as shown in Table 2.2.2, where a_1, a_2 and a_3 are
magnitudes of the unit vectors defining the traditional unit cell and α, β and γ are the ang
between these unit vectors.

Name	Number of Bravais lattices	Conditions
Triclinic	1	$a_1 \neq a_2 \neq a_3 , \alpha \neq \beta \neq \gamma$
Monoclinic	2	$a_1 \neq a_2 \neq a_3 , \alpha = \beta = 90° \neq \gamma$
Orthorhombic	4	$a_1 \neq a_2 \neq a_3 , \alpha = \beta = \gamma = 90°$
Tetragonal	2	$a_1 = a_2 \neq a_3 , \alpha = \beta = \gamma = 90°$
Cubic	3	$a_1 = a_2 = a_3 , \alpha = \beta = \gamma = 90°$
Trigonal	1	$a_1 = a_2 = a_3 , \alpha = \beta = \gamma < 120° \neq 90°$
Hexagonal	1	$a_1 = a_2 \neq a_3 , \alpha = \beta = 90° , \gamma = 120°$

Table 2.2.2.:
Bravais lattices of three-dimensional crystals

The cubic lattices are an important subset of these fourteen Bravais lattices since a large num
of semiconductors are cubic. The three cubic Bravais lattices are the simple cubic lattice,
body-centered cubic lattice and the face-centered cubic lattice as shown in Figure 2.2.3. Since
unit vectors identifying the traditional unit cell have the same size, the crystal structure
completely defined by a single number. This number is the lattice constant, a.

(a)   (b)   (c)
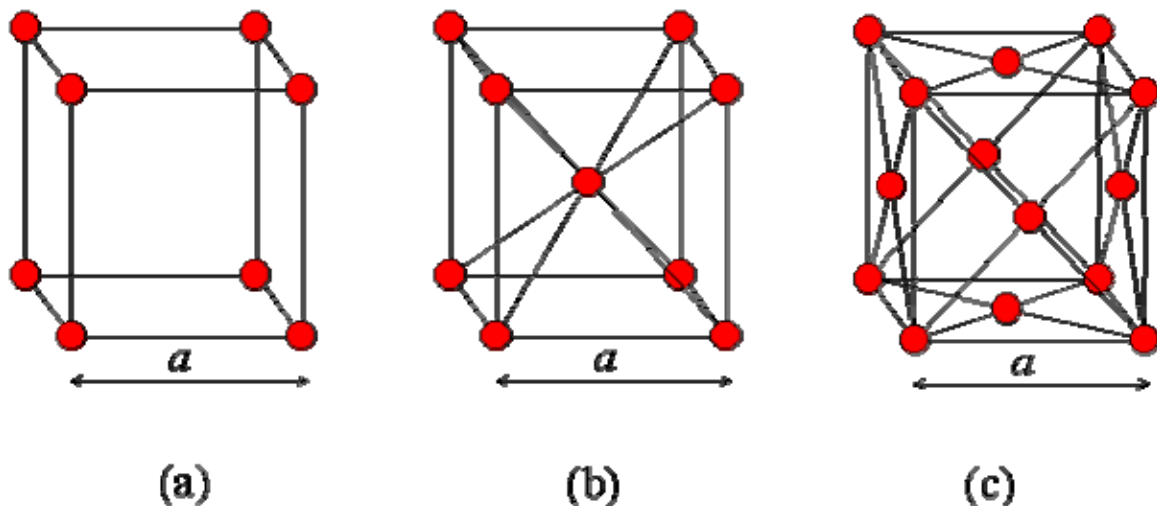
Figure 2.2.3.:
The simple cubic (a), the body-centered cubic (b) and the face centered cubic (c) lattice.

2.2.2 Common semiconductor crystal structures

The most common crystal structure among frequently used semiconductors is the diamo
lattice, shown in Figure 2.2.4. Each atom in the diamond lattice has a covalent bond with f
adjacent atoms, which together form a tetrahedron. This lattice can also be formed from t
face-centered-cubic lattices, which are displaced along the body diagonal of the larger cube
Figure 2.2.4 by one quarter of that body diagonal. The diamond lattice therefore is a fa
centered-cubic lattice with a basis containing two identical atoms.
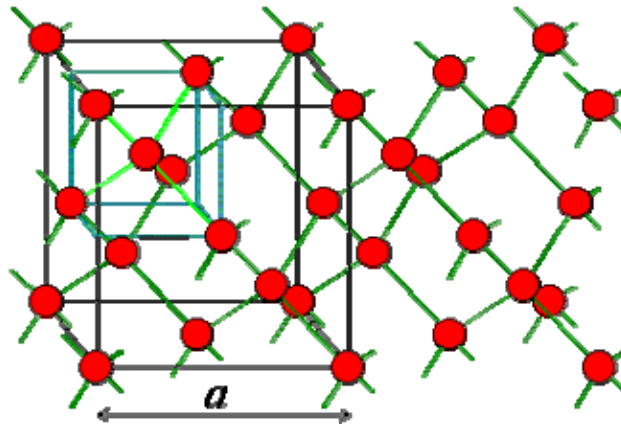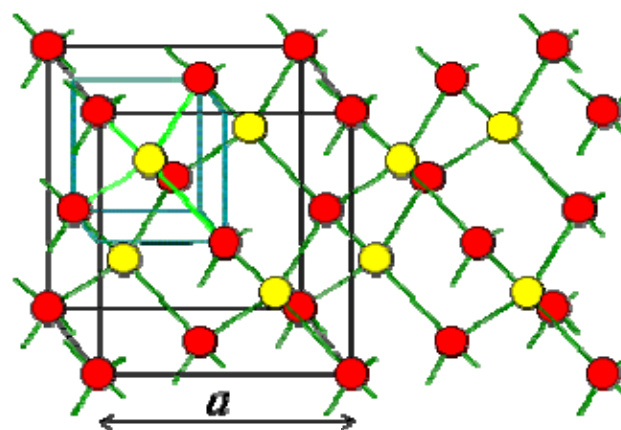
Figure 2.2.4.:
The diamond lattice of silicon and germanium

Compound semiconductors such as GaAs and InP have a crystal structure that is similar to that of diamond. However, the lattice contains two different types of atoms. Each atom still has four covalent bonds, but they are bonds with atoms of the other type. This structure is referred to as the zinc-blende lattice, named after zinc-blende (ZnS) as shown in Figure 2.2.5. Both the diamond lattice and the zinc-blende lattice are cubic lattices. A third common crystal structure is the hexagonal structure also referred to as the wurzite crystal structure, which is the hexagonal form of zinc sulfide (ZnS).

Many semiconductor materials can have more than one crystal structure. A large number of compound semiconductors including GaAs, GaN and ZnS can be either cubic or hexagonal. SiC can be cubic or one of several different hexagonal crystal structures.

The cubic crystals are characterized by a single parameter, the lattice constant a, while the hexagonal structures are characterized in the hexagonal plane by a lattice constant a and by the distance between the hexagonal planes, c.

Figure 2.2.5 :
The zinc-blende crystal structure of GaAs and InP
Example 2.1

Calculate the maximum fraction of the volume in a simple cubic crystal occupied by the ato[...]
Assume that the atoms are closely packed and that they can be treated as hard spheres. T[...]
fraction is also called the packing density.

Solution

The atoms in a simple cubic crystal are located at the corners of the units cell, a cube with sid[...]
Adjacent atoms touch each other so that the radius of each atom equals $a/2$. There are ei[...]
atoms occupying the corners of the cube, but only one eighth of each is within the unit cell[...]
that the number of atoms equals one per unit cell. The packing density is then obtained from:

$$\frac{\text{Volume of atoms}}{\text{Volume of the unit cell}} = \frac{\frac{4}{3}\pi r^3}{a^3} = \frac{\frac{4}{3}\pi(\frac{a}{2})^3}{a^3} = \frac{\pi}{6} = 52\%$$

or about half the volume of the unit cell is occupied by the atoms.

The packing density of four cubic crystals is listed in the table below.

	Radius	Atoms/ unit cell	Packing density
Simple cubic	$\frac{a}{2}$	1	$\frac{\pi}{6} = 52\%$
Body centered cubic	$\frac{\sqrt{3}a}{4}$	2	$\frac{\pi\sqrt{3}}{8} = 68\%$
Face centered cubic	$\frac{\sqrt{2}a}{4}$	4	$\frac{\pi\sqrt{2}}{6} = 74\%$
Diamond	$\frac{\sqrt{3}a}{8}$	8	$\frac{\pi\sqrt{3}}{16} = 34\%$

2.2.3 Growth of semiconductor crystals

Like all crystals, semiconductor crystals can be obtained by cooling the molten semiconduc[...]
material. However, this procedure yields poly-crystalline material since crystals start growing[...]
different locations with a different orientation. Instead when growing single-crystalline sili[...]
one starts with a seed crystal and dips one end into the melt. By controlling the temperat[...]
difference between the seed crystal and the molten silicon, the seed crystal slowly grows. [...]
result is a large single-crystal silicon boule. Such boules have a cylindrical shape, in [...]
because the seed crystal is rotated during growth and in part because of the cylindrical shape[...]

the crucible containing the melt. The boule is then cut into wafers with a diamond saw and further polished to yield the starting material for silicon device fabrication.

Chapter 2: Semiconductor Fundamentals

2.3 Energy bands

Energy bands consisting of a large number of closely spaced energy levels exist in crystalline materials. The bands can be thought of as the collection of the individual energy levels of electrons surrounding each atom. The wave functions of the individual electrons, however, overlap with those of electrons confined to neighboring atoms. The Pauli exclusion principle does not allow the electron energy levels to be the same so that one obtains a set of closely spaced energy levels, forming an energy band. The energy band model is crucial to any detailed treatment of semiconductor devices. It provides the framework needed to understand the concept of an energy bandgap and that of conduction in an almost filled band as described by the empty states.

2.3.1 Free electron model

The free electron model of metals has been used to explain the photo-electric effect (see section 1.2.2). This model assumes that electrons are free to move within the metal but are confined to the metal by potential barriers as illustrated by Figure 2.3.1. The minimum energy needed to extract an electron from the metal equals $q\Phi_M$, where Φ_M is the workfunction. This model is frequently used when analyzing metals. However, this model does not work well for semiconductors since the effect of the periodic potential due to the atoms in the crystal has been ignored.
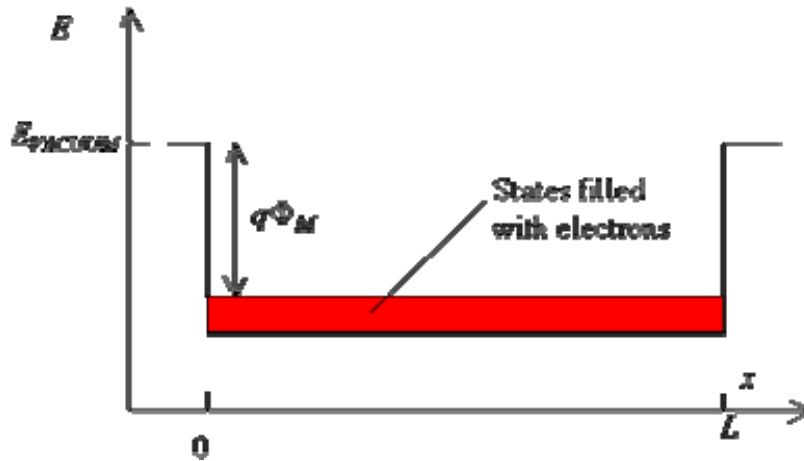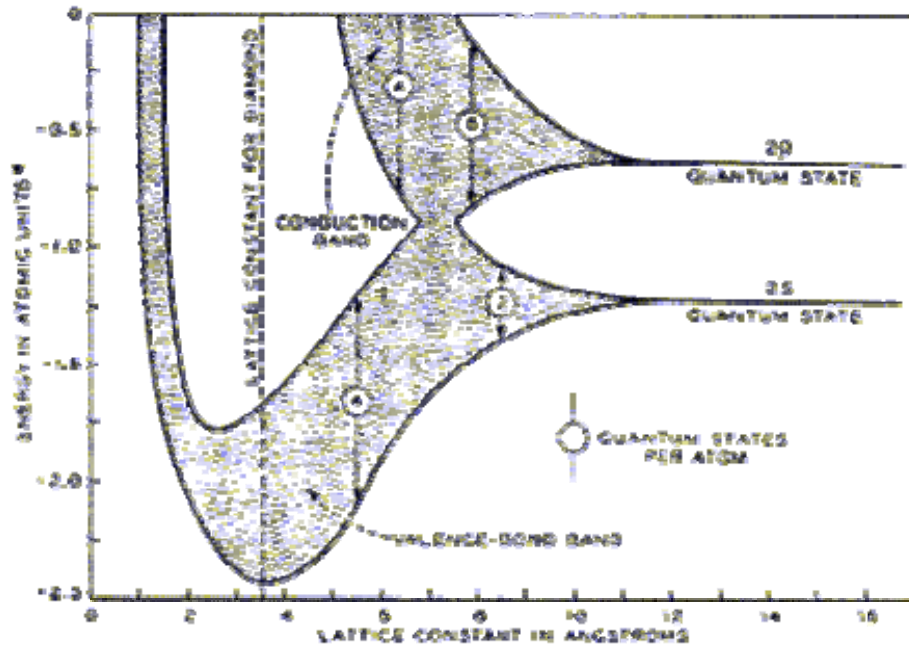
Figure 2.3.1.:
The free electron model of a metal.

2.3.2 Periodic potentials

The analysis of periodic potentials is required to find the energy levels in a semiconductor. T[his]
requires the use of periodic wave functions, called Bloch functions which are beyond the sc[ope]
of this text. The result of this analysis is that the energy levels are grouped in bands, separa[ted]
by energy band gaps. The behavior of electrons at the top and bottom of such a band is simila[r to]
that of a free electron. However, the electrons are affected by the presence of the perio[dic]
potential. The combined effect of the periodic potential is included by adjusting the mass of [the]
electron to a different value. This mass will be referred to as the effective mass.

The effect of a periodic arrangement on the electron energy levels is illustrated by Figure 2.[3.2.]
Shown are the energy levels of electrons in a carbon crystal with the atoms arranged i[n a]
diamond lattice. These energy levels are plotted as a function of the lattice constant, a.

Figure 2.3.2. :
Energy bands for diamond versus lattice constant . One atomic unit equals 1 Rydberg = 13
eV.

Isolated carbon atoms contain six electrons, which occupy the 1s, 2s and 2p orbital in pair
The energy of an electron occupying the 2s and 2p orbital is indicated on the figure. Th
energy of the 1s orbital is not shown. As the lattice constant is reduced, there is an overlap
the electron wavefunctions occupying adjacent atoms. This leads to a splitting of the energ
levels consistent with the Pauli exclusion principle. The splitting results in an energy bar
containing $2N$ states in the 2s band and $6N$ states in the 2p band, where N is the number
atoms in the crystal. A further reduction of the lattice constant causes the 2s and 2p energ
bands to merge and split again into two bands containing $4N$ states each. At zero Kelvin, th
lower band is completely filled with electrons and labeled as the valence band. The upper bar
is empty and labeled as the conduction band.

2.3.3 Energy bands of semiconductors

Complete energy band diagrams of semiconductors are very complex. However, most hav
features similar to that of the diamond crystal discussed in section 2.3.2. In this section, w
first take a closer look at the energy band diagrams of common semiconductors. We the
present a simple diagram containing some of the most important feature and discuss th
temperature dependence of the energy bandgap.

2.3.3.1. Energy band diagrams of common semiconductors

The energy band diagrams of semiconductors are rather complex. The detailed energy ba[nd] diagrams of germanium, silicon and gallium arsenide are shown in Figure <u>2.3.3</u>. The energy[...] plotted as a function of the wavenumber, k, along the main crystallographic directions in [...] crystal, since the band diagram depends on the direction in the crystal. The energy ba[nd] diagrams contain multiple completely-filled and completely-empty bands. In addition, there [...] multiple partially-filled band.
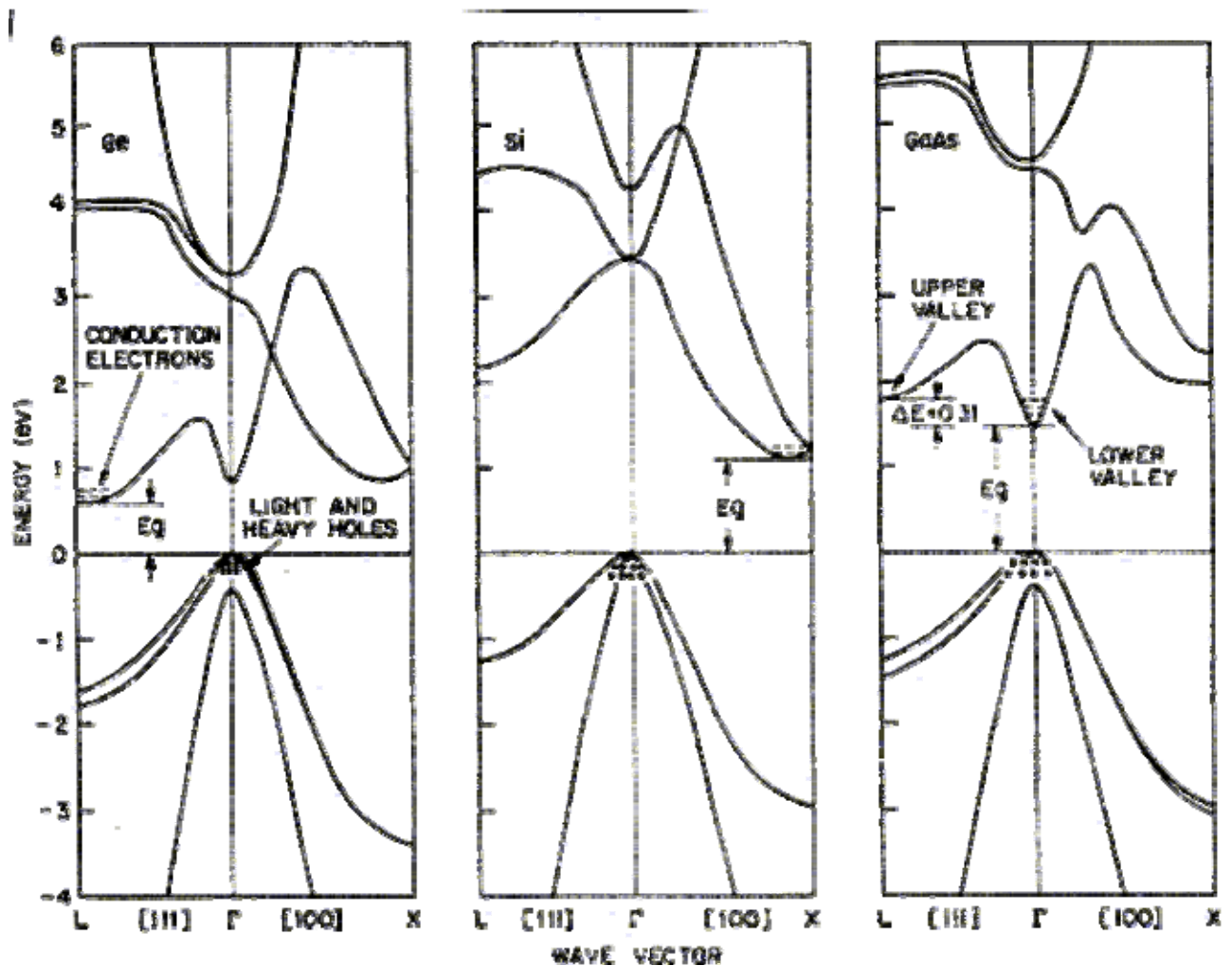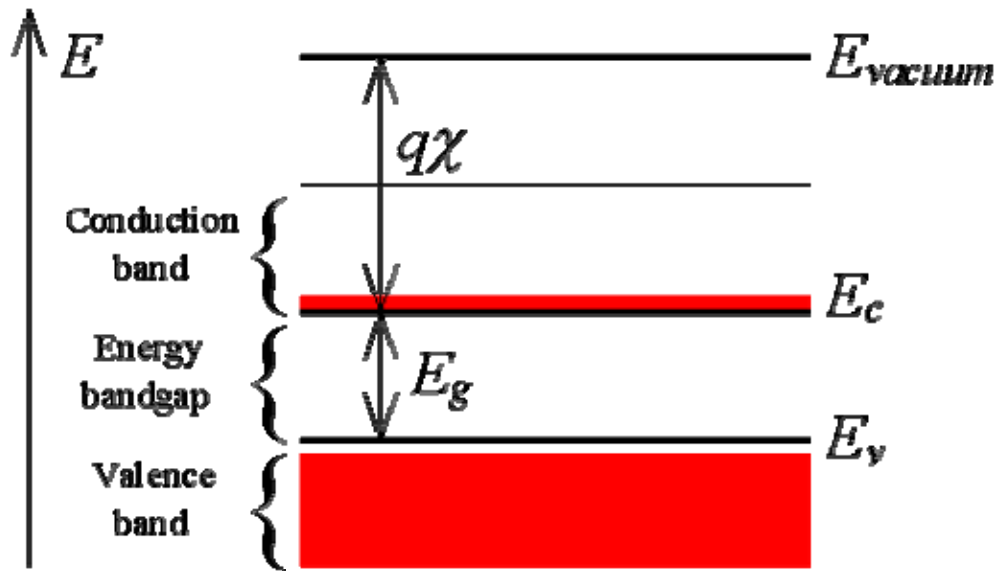

Figure 2.3.3.:
Energy band diagram of (a) germanium, (b) silicon and (c) gallium arsenide

Fortunately, we can simplify the energy band diagram since only the electrons in the high[...] almost-filled band and the lowest almost-empty band dominate the behavior of [...] semiconductor. These bands are indicated on the figure by the + and - signs corresponding to [...] charge of the carriers in those bands.

2.3.3.2. Simple energy band diagram of a semiconductor

The energy band diagrams shown in the previous section are frequently simplified w[hen] analyzing semiconductor devices. Since the electronic properties of a semiconductor [are] dominated by the highest partially empty band and the lowest partially filled band, it is of[ten] sufficient to only consider those bands. This leads to a simplified energy band diagram [for] semiconductors as shown in Figure 2.3.4:

Figure 2.3.4.:
A simplified energy band diagram used to describe semiconductors. Shown are the valence a[nd] conduction band as indicated by the valence band edge, E_v, and the conduction band edge, [E_c.] The vacuum level, E_{vacuum}, and the electron affinity,   , are also indicated on the figure.

The diagram identifies the almost-empty conduction band by a horizontal line. This l[ine] indicates the bottom edge of the conduction band and is labeled E_c. Similarly, the top of [the] valence band is indicated by a horizontal line labeled E_v. The energy band gap is loca[ted] between the two lines, which are separated by the bandgap energy E_g. The distance between [the] conduction band edge, E_c, and the energy of a free electron outside the crystal (called [the] vacuum level labeled E_{vacuum}) is quantified by the electron affinity,   multiplied with [the] electronic charge q.

An important feature of an energy band diagram, which is not included on the simplif[ied] diagram, is whether the conduction band minimum and the valence band maximum occur at [the] same value for the wavenumber. If so, the energy bandgap is called direct. If not, the ene[rgy] bandgap is called indirect. This distinction is of interest for optoelectronic devices as di[rect] bandgap materials provide more efficient absorption and emission of light. For instance, [the] smallest bandgap of germanium and silicon is indirect, while gallium arsenide has a di[rect] bandgap as can be seen on Figure 2.3.3.

2.3.3.3. Temperature dependence of the energy bandgap

The energy bandgap of semiconductors tends to decrease as the temperature is increased. T behavior can be better understood if one considers that the interatomic spacing increases w the amplitude of the atomic vibrations increases due to the increased thermal energy. This ef is quantified by the linear expansion coefficient of a material. An increased interatomic spac decreases the average potential seen by the electrons in the material, which in turn reduces size of the energy bandgap. A direct modulation of the interatomic distance - such as applying compressive (tensile) stress - also causes an increase (decrease) of the bandgap.

The temperature dependence of the energy bandgap, E_g, has been experimentally determi yielding the following expression for E_g as a function of the temperature, T:

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{T + \beta}$$

(2.3.1)

where $E_g(0)$, α and β are the fitting parameters. These fitting parameters are listed germanium, silicon and gallium arsenide in Table 2.3.1:

	Germanium	Silicon	GaAs
$E_g(0)$ (eV)	0.7437	1.166	1.519
α (meV/K)	0.477	0.473	0.541
β (K)	235	636	204

Table 2.3.1.:
Parameters used to calculate the energy bandgap of germanium, silicon and gallium arseni (GaAs) as a function of temperature

A plot of the resulting bandgap versus temperature is shown in Figure 2.3.5 for germaniur silicon and gallium arsenide.
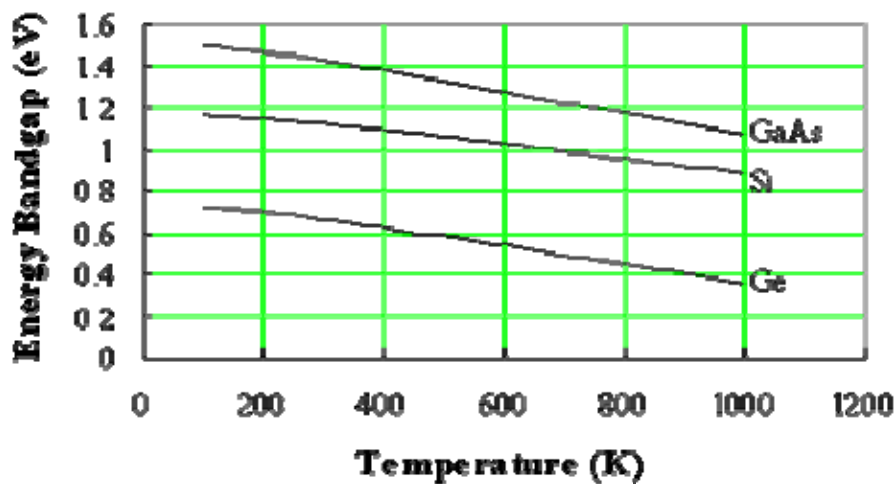
Figure 2.3.5.:
Temperature dependence of the energy bandgap of germanium (Ge), silicon (Si) and galli
arsenide (GaAs).

Example 2.2.

Calculate the energy bandgap of germanium, silicon and gallium arsenide at 300, 400, 500 a
600 K.

Solution

The bandgap of silicon at 300 K equals:

$$E_g(300\ \mathrm{K}) = E_g(0\ \mathrm{K}) - \frac{\alpha T^2}{T + \beta} = 1.166 - \frac{0.473 \times (300)^2}{300 + 636} = 1.12\ \mathrm{eV}$$

Similarly one finds the energy bandgap for germanium and gallium arsenide, as well as at different temperatures, yielding:

	Germanium	Silicon	Gallium Arsenide
T = 300 K	0.66 eV	1.12 eV	1.42 eV
T = 400 K	0.62 eV	1.09 eV	1.38 eV
T = 500 K	0.58 eV	1.06 eV	1.33 eV
T = 600 K	0.54 eV	1.03 eV	1.28 eV

2.3.4 Metals, insulators and semiconductors

Once we know the bandstructure of a given material we still need to find out which ene
levels are occupied and whether specific bands are empty, partially filled or completely filled.

Empty bands do not contain electrons. Therefore, they are not expected to contribute to
electrical conductivity of the material. Partially filled bands do contain electrons as well
available energy levels at slightly higher energies. These unoccupied energy levels ena
carriers to gain energy when moving in an applied electric field. Electrons in a partially fil
band therefore do contribute to the electrical conductivity of the material.

Completely filled bands do contain plenty of electrons but do not contribute to the conductiv
of the material. This is because the electrons cannot gain energy since all energy levels
already filled.

In order to find the filled and empty bands we must find out how many electrons can be pla
in each band and how many electrons are available. Each band is formed due to the splitting
one or more atomic energy levels. Therefore, the minimum number of states in a band equ
twice the number of atoms in the material. The reason for the factor of two is that every ene
level can contain two electrons with opposite spin.

To further simplify the analysis, we assume that only the valence electrons (the electrons in
outer shell) are of interest. The core electrons are tightly bound to the atom and are not allov
to freely move in the material.

Four different possible scenarios are shown in Figure 2.3.6:
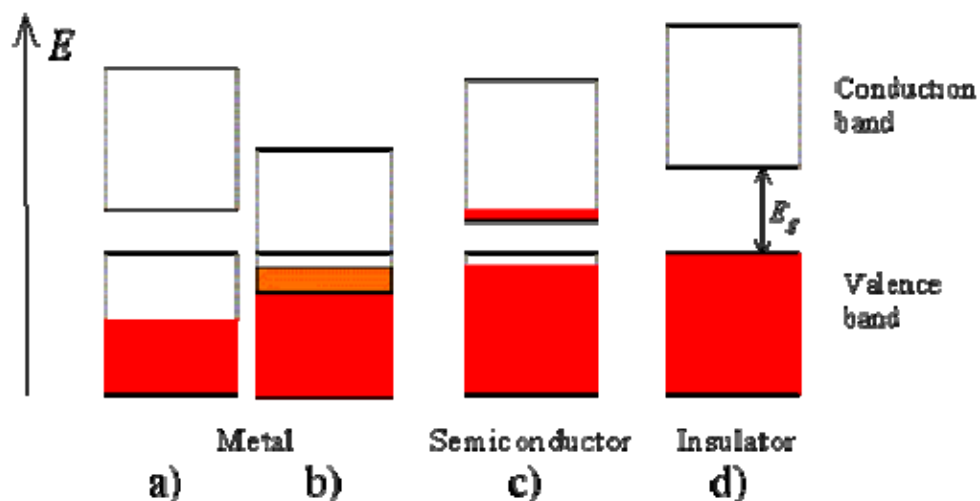
Figure 2.3.6.:
Possible energy band diagrams of a crystal. Shown are a) a half filled band, b) two overlapping
bands, c) an almost full band separated by a small bandgap from an almost empty band and d) a full
band and an empty band separated by a large bandgap.

A half-filled band is shown in Figure 2.3.6 a). This situation occurs in materials consisting of atoms, which contain only one valence electron per atom. Most highly conducting metals including copper, gold and silver satisfy this condition. Materials consisting of atoms that contain two valence electrons can still be highly conducting if the resulting filled band overlaps with an empty band. This scenario is shown in b). No conduction is expected for scenario d) where a completely filled band is separated from the next higher empty band by a larger energy gap. Such materials behave as insulators. Finally, scenario c) depicts the situation in a semiconductor. The completely filled band is now close enough to the next higher empty band that electrons can make it into the next higher band. This yields an almost full band below an almost empty band. We will call the almost full band the valence band since it is occupied by valence electrons. The almost empty band will be called the conduction band, as electrons are free to move in this band and contribute to the conduction of the material.

2.3.5 Electrons and holes in semiconductors

As pointed out in section 2.3.4, semiconductors differ from metals and insulators by the fact that they contain an "almost-empty" conduction band and an "almost-full" valence band. This also means that we will have to deal with the transport of carriers in both bands.

To facilitate the discussion of the transport in the "almost-full" valence band of a semiconductor, we will introduce the concept of holes. It is important for the reader to understand that one could deal with only electrons if one is willing to keep track of all the electrons in the "almost-full" valence band. After all, electrons are the only real particles available in a semiconductor.

The concepts of holes is introduced in semiconductors since it is easier to keep track of the missing electrons in an "almost-full" band, rather than keeping track of the actual electrons in that band. We will now first explain the concept of a hole and then point out how the hole concept simplifies the analysis.

Holes are missing electrons. They behave as particles with the same properties as the electrons would have when occupying the same states except that they carry a positive charge. This definition is illustrated further with Figure 2.3.7, which presents the energy band diagram in the presence of an electric field.
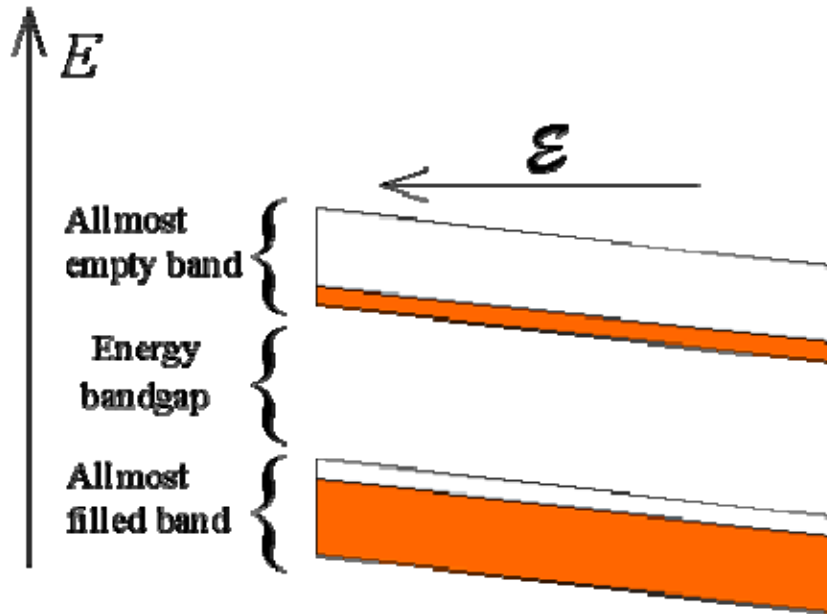
Figure 2.3.7.:
Energy band diagram in the presence of a uniform electric field. Shown are the upper almost-empty band and the lower almost-filled band. The tilt of the bands is caused by an externally applied electric field.

A uniform electric field is assumed which causes a constant gradient of the bands.

The electrons in the almost-empty band are negatively charged particles, which therefore move in a direction, which opposes the direction of the field. Electrons therefore move down hill in the upper band. Electrons in the lower band also move in the same direction. The total current density due to the electrons in the valence band can therefore be written as:

$$J_{vb} = \frac{1}{V} \sum_{\substack{filled \\ states}} (-q) v_i$$

(2.3.2)

where V is the volume of the semiconductor, q is the electronic charge and v is the electron velocity. The sum is taken over all occupied or filled states in the lower band. This equation can be reformulated by first taking the sum over all the states in the lower band and subtracting the current due to the electrons, which are missing in the almost-filled band. This last term therefore represents the sum taken over all the empty states in the lower band, or:

$$J_{vb} = \frac{1}{V} \left(\sum_{\substack{all \\ states}} (-q) v_i - \sum_{\substack{empty \\ states}} (-q) v_i \right)$$

(2.3.3)

The sum over all the states in the lower band has to equal zero since electrons in a completely filled band do not contribute to current, while the remaining term can be written as:

$$J_{vb} = \frac{1}{V} \sum_{\substack{empty \\ states}} (+q)v_i$$

(2.3.4)

which states that the current is due to positively charged particles associated with the empty states in the almost-filled band. We call these particles holes. Keep in mind that there is no real particle associated with a hole. Instead, the combined behavior of all the electrons, which occupy states in the almost-filled band, is the same as that of positively charge particles associated with the unoccupied states.

The reason the concept of holes simplifies the analysis is that the density of states function of a whole band can be rather complex. However, it can be dramatically simplified if only states close to the band edge need to be considered.

2.3.6 The effective mass concept

Electrons with an energy close to a band minimum behave as free electrons. They accelerate in an applied electric field just like a free electron in vacuum. Their wavefunctions are periodic and extend over the size of the material. The presence of the periodic potential, due to the atoms in the crystal without the valence electrons, changes the properties of the electrons. Therefore, the mass of the electron differs from the free electron mass, m_0. Because of the anisotropy of the effective mass and the presence of multiple equivalent band minima, we define two types of effective mass, the effective mass for density of states calculations and the effective mass for conductivity calculations. The effective mass values for electrons and holes are listed together with the value of the smallest energy bandgap in Table 2.3.2. Electrons in gallium arsenide have an isotropic effective mass so that the conductivity effective mass equals the density of states effective mass.

		Germanium	Silicon	GaAs
Smallest energy bandgap at 300 K	E_g (eV)	0 66	1 12	1 424
Electron effective mass for density of states calculations	$\dfrac{m_{e,dos}^*}{m_0}$	0.55	1.08	0.067
Hole effective mass for density of states calculations	$\dfrac{m_{h,dos}^*}{m_0}$	0.37	0.811	0.45
Electron effective mass for conductivity calculations	$\dfrac{m_{e,cond}^*}{m_0}$	0.12	0.26	0.067
Hole effective mass for conductivity calculations	$\dfrac{m_{h,cond}^*}{m_0}$	0.21	0.386	0.34

Table 2.3.2.:
Effective mass of carriers in germanium, silicon and gallium arsenide (GaAs)

Chapter 2: Semiconductor Fundamentals

2.4 Density of states

Before we can calculate the density of carriers in a semiconductor, we have to find the number of available states at each energy. The number of electrons at each energy is then obtained by multiplying the number of states with the probability that a state is occupied by an electron. Since the number of energy levels is very large and dependent on the size of the semiconductor, we will calculate the number of states per unit energy and per unit volume.

2.4.1 Calculation of the density of states

The density of states in a semiconductor equals the density per unit volume and energy of the number of solutions to Schrödinger's equation. We will assume that the semiconductor can be modeled as an infinite quantum well in which electrons with effective mass, m^*, are free to move. The energy in the well is set to zero. The semiconductor is assumed a cube with side L. This assumption does not affect the result since the density of states per unit volume should not depend on the actual size or shape of the semiconductor.

The solutions to the wave equation (equation 1.2.14) where $V(x) = 0$ are sine and cosine functions:

$$\Psi = A \sin(k_x x) + B \cos(k_x x)$$

$$(2.4.1)$$

Where A and B are to be determined. The wavefunction must be zero at the infinite barriers of the well. At $x = 0$ the wavefunction must be zero so that only sine functions can be valid solutions or B must equal zero. At $x = L$, the wavefunction must also be zero yielding the following possible values for the wavenumber, k_x.

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

$$(2.4.2)$$

This analysis can now be repeated in the y and z direction. Each possible solution corresponds to a cube in k-space with size $n\pi/L$ as indicated on Figure 2.4.1.
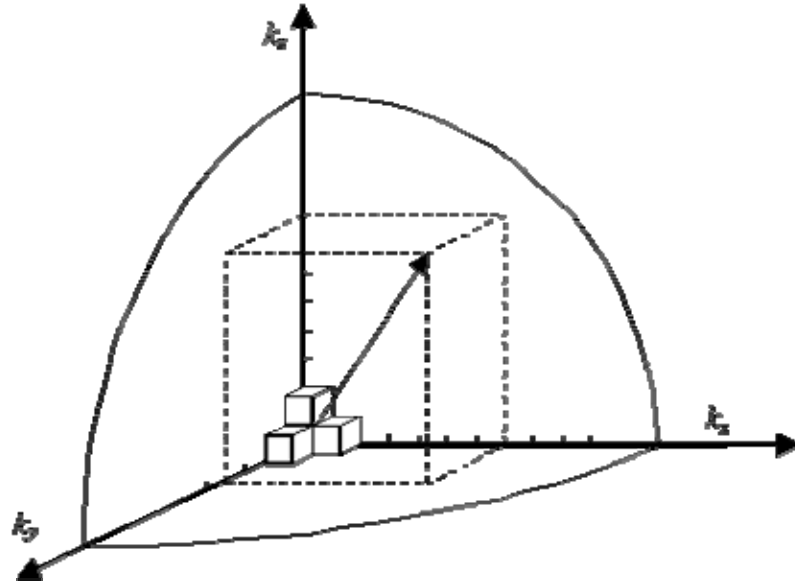
Figure 2.4.1:
Calculation of the number of states with wavenumber less than k.

The total number of solutions with a different value for k_x, k_y and k_z and with a magnitude of the wavevector less than k is obtained by calculating the volume of one eighth of a sphere with radius k and dividing it by the volume corresponding to a single solution, $(\frac{\pi}{L})^3$, yielding:

$$N = 2 \times \frac{1}{8} \times (\frac{L}{\pi})^3 \times \frac{4}{3} \times \pi \times k^3$$
$$(2.4.3)$$

A factor of two is added to account for the two possible spins of each solution. The density per unit energy is then obtained using the chain rule:

$$\frac{dN}{dE} = \frac{dN}{dk}\frac{dk}{dE} = (\frac{L}{\pi})^3 \pi k^2 \frac{dk}{dE}$$
$$(2.4.4)$$

The kinetic energy E of a particle with mass m^* is related to the wavenumber, k, by:

$$E(k) = \frac{\hbar^2 k^2}{2m^*}$$
$$(2.4.5)$$

And the density of states per unit volume and per unit energy, $g(E)$, becomes:

$$g(E) = \frac{1}{L^3}\frac{dN}{dE} = \frac{8\pi\sqrt{2}}{h^3}m^{*3/2}\sqrt{E}, \text{for } E \geq 0$$
$$(2.4.6)$$

The density of states is zero at the bottom of the well as well as for negative energies.

The same analysis also applies to electrons in a semiconductor. The effective mass takes into account the effect of the periodic potential on the electron. The minimum energy of the electron is the energy at the bottom of the conduction band, E_c, so that the density of states for electrons in the conduction band is given by:

$$g_c(E) = \frac{1}{L^3} \frac{dN}{dE} = \frac{8\pi\sqrt{2}}{h^3} m^{*3/2} \sqrt{E - E_c} \text{, for } E \geq E_c$$

$$(2.4.7)$$

Example 2.3

Calculate the number of states per unit energy in a 100 by 100 by 10 nm piece of silicon ($m^* = 1.08$ m_0) 100 meV above the conduction band edge. Write the result in units of eV^{-1}.

Solution

The density of states equals:

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

$$= \frac{8\pi\sqrt{2}\,(1.08\times9.1\times10^{-31})^{3/2}}{(6.626\times10^{-34})^3} \sqrt{0.1\times1.6\times10^{-19}}$$

$$= 1.51\times10^{56} \text{ m}^{-3}\text{J}^{-1}$$

So that the total number of states per unit energy equals:

$$g(E)V = 1.51\times10^{56}\times10^{-22} \text{ J}^{-1} = 2.41\times10^5 \text{ eV}^{-1}$$

Chapter 2: Semiconductor Fundamentals

2.5 Carrier distribution functions

The distribution or probability density functions describe the probability with which one can expect particles to occupy the available energy levels in a given system. Of particular interest is the probability density function of electrons, called the Fermi function. The derivation of such probability density functions belongs in a statistical thermodynamics course. However, given the

importance of the Fermi distribution function, we will carefully examine an example as well as the characteristics of this function. Other distribution functions such as the impurity distribution functions, the Bose-Einstein distribution function and the Maxwell Boltzmann distribution are also provided.

2.5.1 Fermi-Dirac distribution function

The Fermi-Dirac distribution function, also called Fermi function, provides the probability of occupancy of energy levels by Fermions. Fermions are half-integer spin particles, which obey the Pauli exclusion principle. The Pauli exclusion principle postulates that only one Fermion can occupy a single quantum state. Therefore, as Fermions are added to an energy band, they will fill the available states in an energy band just like water fills a bucket. The states with the lowest energy are filled first, followed by the next higher ones. At absolute zero temperature ($T = 0$ K), the energy levels are all filled up to a maximum energy, which we call the Fermi level. No states above the Fermi level are filled. At higher temperature, one finds that the transition between completely filled states and completely empty states is gradual rather than abrupt.

Electrons are Fermions. Therefore, the Fermi function provides the probability that an energy level at energy, E, in thermal equilibrium with a large system, is occupied by an electron. The system is characterized by its temperature, T, and its Fermi energy, E_F. The Fermi function is given by:

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}}$$

(2.5.1)

This function is plotted in Figure 2.5.1 for different temperatures.
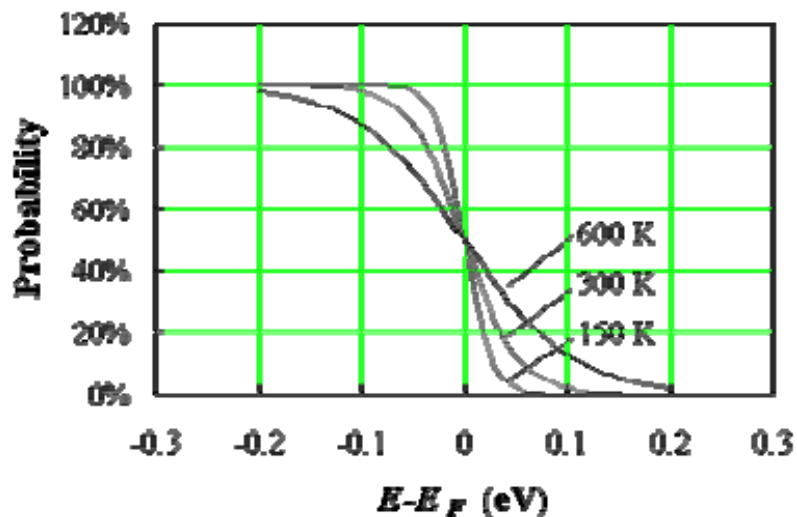
Figure 2.5.1 :
The Fermi function at three different temperatures.

The Fermi function has a value of one for energies, which are more than a few times kT below the Fermi energy. It equals 1/2 if the energy equals the Fermi energy and decreases exponentially for energies which are a few times kT larger than the Fermi energy. While at $T = 0$ K the Fermi

function equals a step function, the transition is more gradual at finite temperatures and more so at higher temperatures.

2.5.2 Example

To better understand the origin of distribution functions, we now consider a specific system with equidistant energy levels at 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, ... eV. Each energy level can contain two electrons. Since electrons are indistinguishable from each other, no more than two electrons (with opposite spin) can occupy a given energy level. This system contains 20 electrons.

The minimum energy of this system corresponds to the situation where all 20 electrons occupy the ten lowest energy levels without placing more than 2 in any given level. This situation occurs at $T =$ 0 K and the total energy equals 100 eV.

Since we are interested in a situation where the temperature is not zero, we arbitrarily set the total energy at 106 eV, which is 6 eV more than the minimum possible energy of this system. This ensures that the thermal energy is not zero so that the system must be at a non-zero temperature.

There are 24 possible and different configurations, which satisfy these particular constraints. Eight of those configurations are shown in Figure 2.5.2, where the filled circles represent the electrons:
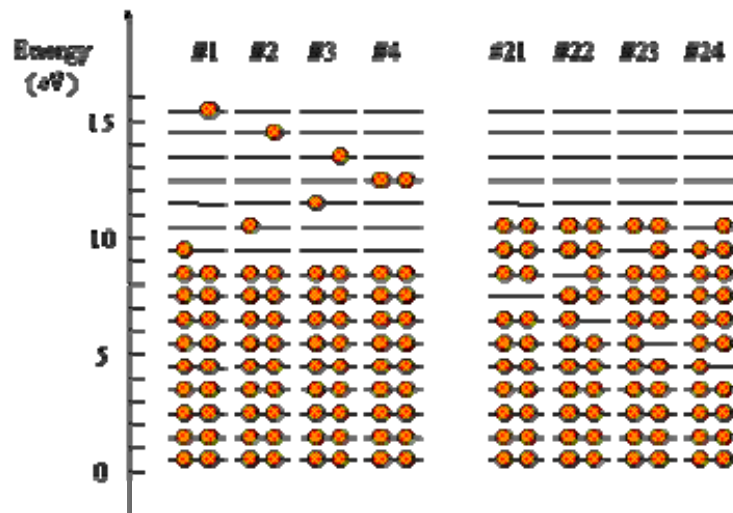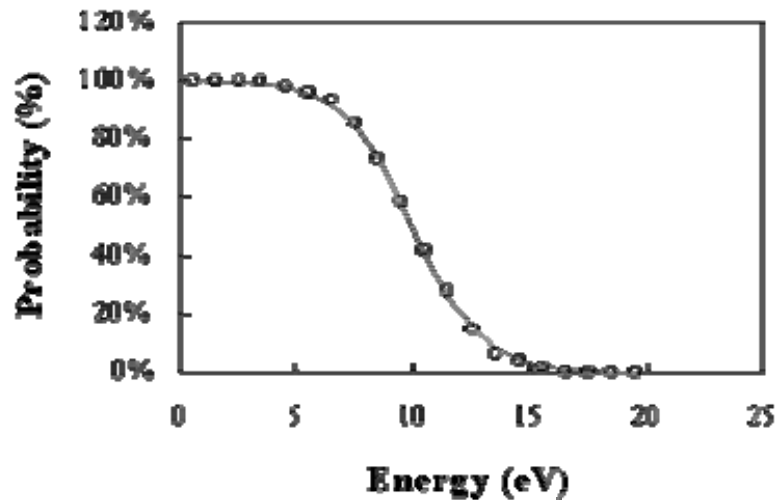

Figure 2.5.2 :
Eight of the 24 possible configurations in which 20 electrons can be placed having a total energy of 106 eV.

We no apply the basis postulate of statistical thermodynamics, namely that all possible configurations are equally likely to occur. The expected configuration therefore equals the average occupancy of all possible configurations.

The average occupancy of each energy level taken over all (and equally probable) 24 configurations is compared in Figure 2.5.3 to the Fermi-Dirac distribution function. A best fit was obtained using a Fermi energy of 9.998 eV and $kT = 1.447$ eV or $T = 16,800$ K. The agreement is surprisingly good considering the small size of this system.

Figure 2.5.3 :
Probability versus energy averaged over the 24 possible configurations (circles) fitted with a Fermi-Dirac function (solid line) using $kT = 1.447$ eV and $E_F = 9.998$ eV.

Based on the construction of the distribution function in this example, one would expect the distribution function to be dependent on the density of states. This is the case for small systems. However, for large systems and for a single energy level in thermal equilibrium with a larger system, the distribution function no longer depends on the density of states. This is very fortunate, since it dramatically simplifies the carrier density calculations. One should also keep in mind that the Fermi energy for a particular system as obtained in section 2.6 does depend on the density of states.

2.5.3 Impurity distribution functions

The distribution function of impurities differs from the Fermi-Dirac distribution function although the particles involved are Fermions. The difference is due to the fact that an ionized donor energy level still contains one electron, which can have either spin (spin up or spin down). The donor energy level cannot be empty since this would leave a doubly positively charged atom, which would have an energy different from the donor energy. The distribution function for donors therefore differs from the Fermi function and is given by:

$$f_{donor}(E_d) = \frac{1}{1 + \frac{1}{2} e^{(E_d - E_F)/kT}}$$
(2.5.2)

The distribution function for acceptors differs also because of the different possible ways to occupy the acceptor level. The neutral acceptor contains no electrons. The ionized acceptor contains one electron, which can have either spin, while the doubly negatively charged state is not allowed since this would require a different energy. This restriction would yield a factor of 2 in front of the exponential term. In addition, one finds that most commonly used semiconductors have a two-fold degenerate valence band, which causes this factor to increase to four, yielding:

$$f_{acceptor}(E_a) = \frac{1}{1 + 4e^{(E_a - E_F)/kT}}$$
(2.5.3)

2.5.4 Other distribution functions and comparison

Other distribution functions include the Bose-Einstein distribution and the Maxwell-Boltzmann distribution. These are briefly discussed below and compared to the Fermi-Dirac distribution function.

The Bose-Einstein distribution function applies to bosons. Bosons are particles with integer spin and include photons, phonons and a large number of atoms. Bosons do not obey the Pauli exclusion principle so that any number can occupy one energy level. The Bose-Einstein distribution function is given by:

$$f_{BE}(E) = \frac{1}{e^{(E - E_F)/kT} - 1}$$
(2.5.4)

This function is only defined for $E > E_F$.

The Maxwell Boltzmann applies to non-interacting particles, which can be distinguished from each other. This distribution function is also called the classical distribution function since it provides the probability of occupancy for non-interacting particles at low densities. Atoms in an ideal gas form a typical example of such particles. The Maxwell-Boltzmann distribution function is given by:

$$f_{MB}(E) = \frac{1}{e^{(E - E_F)/kT}}$$
(2.5.5)

A plot of the three distribution functions, the Fermi-Dirac distribution, the Maxwell-Boltzmann distribution and the Bose-Einstein distribution is shown in Figure 2.5.4.
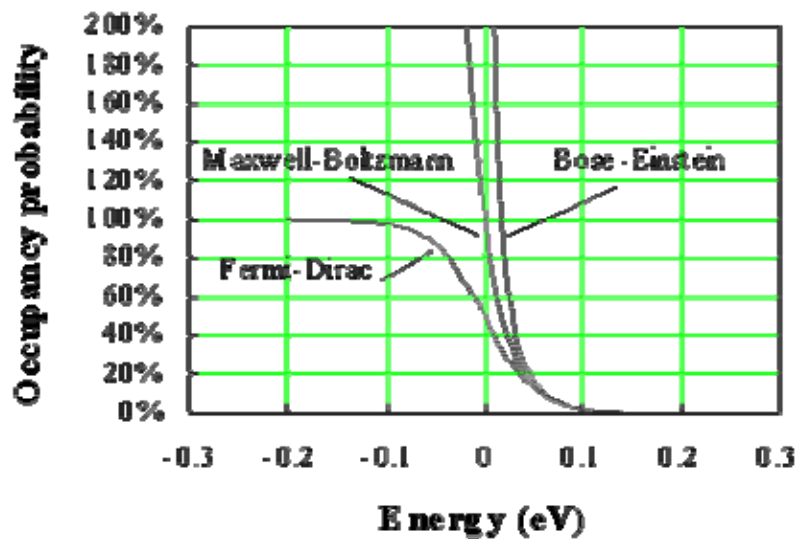
Figure 2.5.4 :
Probability of occupancy versus energy of the Fermi-Dirac, the Bose-Einstein and the Maxwell-Boltzmann distribution. The Fermi energy, E_F, is assumed to be zero.

All three functions are almost equal for large energies (more than a few kT beyond the Fermi energy). The Fermi-Dirac distribution reaches a maximum of 100% for energies, which are a few kT below the Fermi energy, while the Bose-Einstein distribution diverges at the Fermi energy and has no validity for energies below the Fermi energy.

Chapter 2: Semiconductor Fundamentals

2.6. Carrier densities

Now that we have discussed the density of states and the distribution functions, we have all the necessary tools to calculate the carrier density in a semiconductor.

2.6.1 General discussion

The density of electrons in a semiconductor is related to the density of available states and the probability that each of these states is occupied. The density of occupied states per unit volume and energy, $n(E)$, is simply the product of the density of states in the conduction band, $g_c(E)$ and the Fermi-Dirac probability function, $f(E)$, (also called the Fermi function):

$$n(E) = g_c(E) f(E)$$
(2.6.1)

Since holes correspond to empty states in the valence band, the probability of having a hole equals the probability that a particular state is not filled, so that the hole density per unit energy, $p(E)$, equals:

$$p(E) = g_v(E)[1 - f(E)]$$
(2.6.2)

Where $g_v(E)$ is the density of states in the valence band. The density of carriers is then obtained by integrating the density of carriers per unit energy over all possible energies within a band. A general expression is derived as well as an approximate analytic solution, which is valid for non-degenerate semiconductors. In addition, we also present the Joyce-Dixon approximation, an approximate solution useful when describing degenerate semiconductors.

The density of states in a semiconductor was obtained by solving the Schrödinger equation for the particles in the semiconductor. Rather than using the actual and very complex potential in the semiconductor, we use the simple particle-in-a box model, where one assumes that the particle is free to move within the material.

For an electron which behaves as a free particle with effective mass, m^*, the density of states was derived in section 2.4, yielding:

$$g_c(E) = \frac{8\pi\sqrt{2}}{h^3} m_e^{*3/2} \sqrt{E - E_c}, \text{ for } E \geq E_c$$
(2.6.3)

where E_c is the bottom of the conduction band below which the density of states is zero. The density of states for holes in the valence band is given by:

$$g_v(E) = \frac{8\pi\sqrt{2}}{h^3} m_e^{*3/2} \sqrt{E_v - E}, \text{ for } E \leq E_v$$
(2.6.4)

2.6.2. Calculation of the Fermi integral

The carrier density in a semiconductor, is obtained by integrating the product of the density of states and the probability density function over all possible states. For electrons in the conduction band

the integral is taken from the bottom of the conduction band, labeled, E_c, to the top of the conduction band:

$$n = \int_{E_c}^{\text{top of the conduction band}} n(E)dE = \int_{E_c}^{\text{top of the conduction band}} g_c(E)f(E)dE$$

(2.6.5)

Where $g_c(E)$ is the density of states in the conduction band and $f(E)$ is the Fermi function.

This general expression is illustrated with Figure 2.6.1 for a parabolic density of states function with $E_c = 0$. The figure shows the density of states function, $g_c(E)$, the Fermi function, $f(E)$, as well as the product of both, which is the density of electrons per unit volume and per unit energy, $n(E)$. The integral corresponds to the crosshatched area.
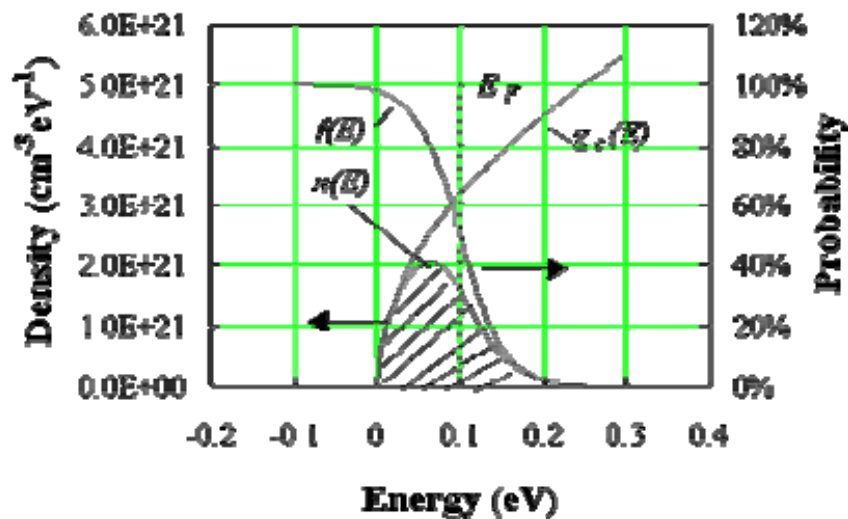
Figure 2.6.1 :
The carrier density integral. Shown are the density of states, $g_c(E)$, the density per unit energy, $n(E)$, and the probability of occupancy, $f(E)$. The carrier density, n_o, equals the crosshatched area.

The actual location of the top of the conduction band does not need to be known as the Fermi function goes to zero at higher energies. The upper limit can therefore be replaced by infinity. We also relabeled the carrier density as no to indicate that the carrier density is the carrier density in thermal equilibrium.

$$n_o = \int_{E_c}^{\infty} g_c(E)f(E)dE$$

(2.6.6)

Using equations (2.6.3) and (2.5.1) this integral becomes:

$$n_o = \int_{E_c}^{\infty} \frac{8\pi\sqrt{2}}{h^3} m_e^{*3/2} \sqrt{E - E_c} \frac{1}{1 + e^{\frac{E - E_F}{kT}}} dE$$

While this integral can not be solved analytically at non-zero temperatures, we can obtain either a numeric solution or an approximate analytical solution. Similarly for holes one obtains:

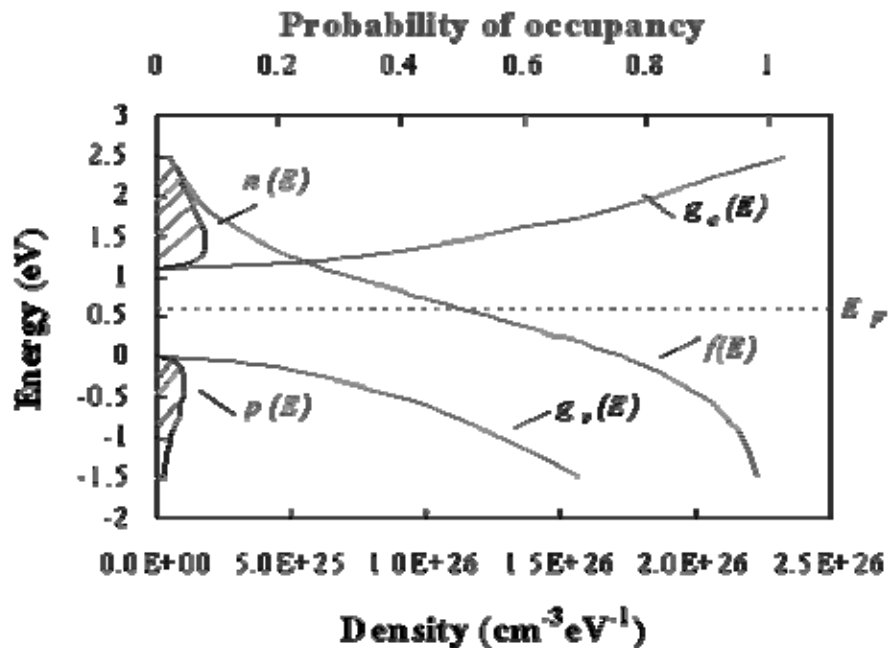$$p_o = \int_{-\infty}^{E_v} g_v(E)[1 - f(E)]dE$$

(2.6.8)

and

$$p_o = \int_{-\infty}^{E_v} \frac{8\pi\sqrt{2}}{h^3} m_h^{*3/2} \sqrt{E_v - E} \frac{1}{1 + e^{\frac{E_F - E}{kT}}} dE$$

(2.6.9)

The calculation of the electron and hole density in a semiconductor is further illustrated by Figure 2.6.2.

Figure 2.6.2 :
The density of states and carrier densities in the conduction and valence band. Shown are the electron and hole density per unit energy, $n(E)$ and $p(E)$, the density of states in the conduction and valence band, $g_c(E)$ and $g_v(E)$ and the probability of occupancy, $f(E)$. The crosshatched area indicates the electron and hole densities.

Indicated on the figure are the density of states in the conduction and valence band, the Fermi distribution function and the electron and hole densities per unit energy. The crosshatched areas indicate the thermal-equilibrium carrier densities. From the figure, one can easily see that the electron density will increase as the Fermi energy is increased. The hole density decreases with

increasing Fermi energy. As the Fermi energy is decreased, the electron density decreases and the hole density increases.

2.6.2.1 Carrier density at zero Kelvin

Equation (2.6.7) can be solved analytically at $T = 0$ K, since the Fermi function at $T = 0$ K equals one for all energies below the Fermi energy and 0 for all energies larger than the Fermi energy. Equation (2.6.7) can therefore be simplified to:

$$n_o = \int_{E_c}^{E_F} g_c(E)dE \text{ at } T = 0 \text{ K}$$

(2.6.10)

and integration yields:

$$n_o = \frac{2}{3}\frac{\sqrt{2}}{\pi^2}\left(\frac{qm^*}{\hbar^2}\right)^{3/2}(E_F - E_c)^{3/2}, \text{ for } E_F \geq E_c$$

(2.6.11)

This expression can be used to approximate the carrier density in heavily degenerate semiconductors provided $kT << (E_F - E_c) > 0$

2.6.2.2 Non-degenerate semiconductors

Non-degenerate semiconductors are defined as semiconductors for which the Fermi energy is at least $3kT$ away from either band edge. The reason we restrict ourselves to non-degenerate semiconductors is that this definition allows the Fermi function to be replaced by a simple exponential function, i.e. the Maxwell-Boltzmann distribution function. The carrier density integral can then be solved analytically yielding:

$$n_o \cong \int_{E_c}^{\infty} \frac{8\pi\sqrt{2}}{h^3}m_e^{*3/2}\sqrt{E - E_c}\,e^{\frac{E_F - E}{kT}}dE = N_c e^{\frac{E_F - E_c}{kT}}$$

(2.6.12)

with

$$N_c = 2[\frac{2\pi m_e^* kT}{h^2}]^{3/2}$$

(2.6.13)

where N_c is the effective density of states in the conduction band. Similarly for holes, one can approximate the hole density integral as:

$$p_o \cong \int_{-\infty}^{E_v} \frac{8\pi\sqrt{2}}{h^3}m_h^{*3/2}\sqrt{E_v - E}\,e^{\frac{E - E_F}{kT}}dE = N_v e^{\frac{E_v - E_F}{kT}}$$

(2.6.14)

with

$$N_v = 2[\frac{2 \pi m_h^* kT}{h^2}]^{3/2}$$
(2.6.15)

where N_v is the effective density of states in the valence band.

Example 2.4
Calculate the effective densities of states in the conduction and valence bands of germanium, silicon and gallium arsenide at 300 K.
Solution

The effective density of states in the conduction band of germanium equals: where the effective mass for density of states was used (Appendix 3). Similarly one finds the effective densities for silicon and gallium arsenide and those of the valence band:

$$N_c = 2(\frac{2 \pi m_e^* kT}{h^2})^{3/2}$$
$$= 2(\frac{2 \pi 0.55 \times 9.11 \times 10^{-31} \times 1.38 \times 10^{-23} \times 300}{(6.626 \times 10^{-34})^2})^{3/2}$$
$$= 1.02 \times 10^{25} \text{ m}^{-3} = 1.02 \times 10^{19} \text{ cm}^{-3}$$

Calculate the effective densities of states in the conduction and valence bands of germanium, silicon and gallium arsenide at 300 K.

	Germanium	Silicon	Gallium Arsenide
N_c (cm^{-3})	1.02×10^{19}	2.81×10^{19}	4.35×10^{17}
N_v (cm^{-3})	5.64×10^{18}	1.83×10^{19}	7.57×10^{18}

Note that the effective density of states is temperature dependent and can be obtain from:

$$N_c(T) = N_c(300\text{K})(\frac{T}{300})^{3/2}$$

where $N_c(300 \text{ K})$ is the effective density of states at 300 K.

2.6.2.3 Degenerate semiconductors

A useful approximate expression applicable to degenerate semiconductors was obtained by Joyce and Dixon and is given by:

$$\frac{E_F - E_c}{kT} \cong \ln \frac{n_o}{N_c} + \frac{1}{\sqrt{8}} \frac{n_o}{N_c} - (\frac{3}{16} - \frac{\sqrt{3}}{9})(\frac{n_o}{N_c})^2 + \dots$$

(2.6.16)

for electrons and by:

$$\frac{E_v - E_F}{kT} \cong \ln \frac{p_o}{N_v} + \frac{1}{\sqrt{8}} \frac{p_o}{N_v} - (\frac{3}{16} - \frac{\sqrt{3}}{9})(\frac{p_o}{N_v})^2 + \dots$$

(2.6.17)

for holes.

2.6.3. Intrinsic semiconductors

Intrinsic semiconductors are semiconductors, which do not contain impurities. They do contain electrons as well as holes. The electron density equals the hole density since the thermal activation of an electron from the valence band to the conduction band yields a free electron in the conduction band as well as a free hole in the valence band. We will identify the intrinsic hole and electron density using the symbol n_i, and refer to it as the intrinsic carrier density.

2.6.3.1 Intrinsic carrier density

Intrinsic semiconductors are usually non-degenerate, so that the expressions for the electron (2.6.12) and hole (2.6.14) densities in non-degenerate semiconductors apply. Labeling the Fermi energy of intrinsic material as E_i, we can then write two relations between the intrinsic carrier density and the intrinsic Fermi energy, namely:

$$n_i = n_o \big|_{(E_F = E_i)} = N_c e^{(E_i - E_c)/kT}$$

(2.6.18)

It is possible to eliminate the intrinsic Fermi energy from both equations, simply by multiplying both equations and taking the square root. This provides an expression for the intrinsic carrier density as a function of the effective density of states in the conduction and valence band, and the bandgap energy $E_g = E_c - E_v$.

$$n_i = \sqrt{N_c N_v} \, e^{-E_g/2kT}$$

$$(2.6.19)$$

The temperature dependence of the intrinsic carrier density is dominated by the exponential dependence on the energy bandgap. In addition, one has to consider the temperature dependence of the effective densities of states and that of the energy bandgap. A plot of the intrinsic carrier density versus temperature is shown in Figure 2.6.3. The temperature dependence of the effective masses was ignored.
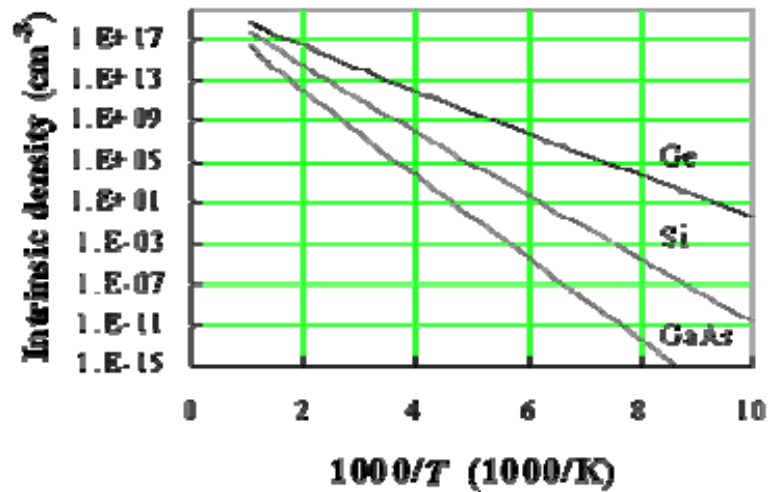
Figure 2.6.3 :
Intrinsic carrier density versus temperature in gallium arsenide (GaAs), silicon (Si) and germanium (Ge).
Example 2.4b
Calculate the intrinsic carrier density in germanium, silicon and gallium arsenide at 300, 400, 500 and 600 K.
Solution

The intrinsic carrier density in silicon at 300 K equals:

$$n_i \, (300 \, \text{K}) = \sqrt{N_c N_v} \, \exp(\frac{E_g}{2kT})$$
$$= \sqrt{2.81 \times 10^{19} \times 1.83 \times 10^{19}} \, \exp(\frac{1.12}{2 \times 0.0258})$$
$$= 8.72 \times 10^9 \, \text{cm}^{-3}$$

Similarly one finds the intrinsic carrier density for germanium and gallium arsenide at different temperatures, yielding:

	Germanium	Silicon	Gallium Arsenide
300 K	2.02×10^{13}	8.72×10^{9}	2.03×10^{6}
400 K	1.38×10^{15}	4.52×10^{12}	5.98×10^{9}
500 K	1.91×10^{16}	2.16×10^{14}	7.98×10^{11}
600 K	1.18×10^{17}	3.07×10^{15}	2.22×10^{13}

Note that the values at 300 K as calculated in example 2.4 are not identical to those listed in Appendix 3. This is due to an accumulation of assumptions in the derivation. The numbers in Appendix 3 are obtained from careful measurements and should therefore be used instead of those calculated in example 2.4.

2.6.3.2 Mass action law

Using the same approach as in section 2.6.3.1, one can prove that the product of the electron and hole density equals the square of the intrinsic carrier density for any non-degenerate semiconductor. By multiplying the expressions for the electron and hole densities in a non-degenerate semiconductor, as in equations (2.6.12) and (2.6.14), one obtains:

$$n_o \cdot p_o = N_c N_v e^{(E_v - E_c)/kT} = n_i^2$$
(2.6.20)

This property is referred to as the mass action law. It is a powerful relation, which enables to quickly find the hole density if the electron density is known or vice versa. This relation is only valid for non-degenerate semiconductors in thermal equilibrium

2.6.3.3 Intrinsic Fermi energy

The above equations for the intrinsic electron and hole density can be solved for the intrinsic Fermi energy, yielding:

$$E_i = \frac{E_c + E_v}{2} + \frac{1}{2} kT \ln(\frac{N_v}{N_c})$$
(2.6.21)

The intrinsic Fermi energy is typically close to the midgap energy, half way between the conduction and valence band edge. The intrinsic Fermi energy can also be expressed as a function of the effective masses of the electrons and holes in the semiconductor. For this we use equations (2.6.13) and (2.6.15) for the effective density of states in the conduction and valence band, yielding:

$$E_i = \frac{E_c + E_v}{2} + \frac{3}{4} kT \ln(\frac{m_h^*}{m_e^*})$$
(2.6.22)

2.6.3.4 Intrinsic material as reference

Dividing the expressions for the carrier densities (2.6.12) and (2.6.14), by the one for the intrinsic density (2.6.18) allows to write the carrier densities as a function of the intrinsic density, n_i, and the intrinsic Fermi energy, E_i, or:

$$n_o = n_i e^{(E_F - E_i)/kT}$$

(2.6.23)

and

$$p_o = n_i e^{(E_i - E_F)/kT}$$

(2.6.24)

We will use primarily these two equations to find the electron and hole density in a semiconductor in thermal equilibrium. The same relations can also be rewritten to obtain the Fermi energy from either carrier density, namely:

$$E_F = E_i + kT \ln \frac{n_o}{n_i}$$

(2.6.25)

and

$$E_F = E_i - kT \ln \frac{p_o}{n_i}$$

(2.6.26)

2.6.4. Doped semiconductors

2.6.4.1 Dopants and impurities
2.6.4.2 Ionization energy model
2.6.4.3 Analysis of non-degenerately doped semiconductors
2.6.4.4 General analysis

Doped semiconductors are semiconductors, which contain impurities, foreign atoms incorporated into the crystal structure of the semiconductor. Either these impurities can be unintentional, due to lack of control during the growth of the semiconductor, or they can be added on purpose to provide free carriers in the semiconductor.

The generation of free carriers requires not only that impurities are present, but also that the impurities give off electrons to the conduction band in which case they are called donors. If they give off holes to the valence band, they are called acceptors (since they effectively accept an electron from the filled valence band). The ionization of shallow donors and acceptors are illustrated by Figure 2.6.4. Indicated are the donor and acceptor energies, E_d and E_a. The donor energy level is filled prior to ionization. Ionization causes the donor to be emptied, yielding an electron in the conduction band and a positively charged donor ion. The acceptor energy is empty prior to ionization. Ionization of the acceptor corresponds to the empty acceptor level being filled by an electron from the filled valence band. This is equivalent to a hole given off by the acceptor atom to the valence band.
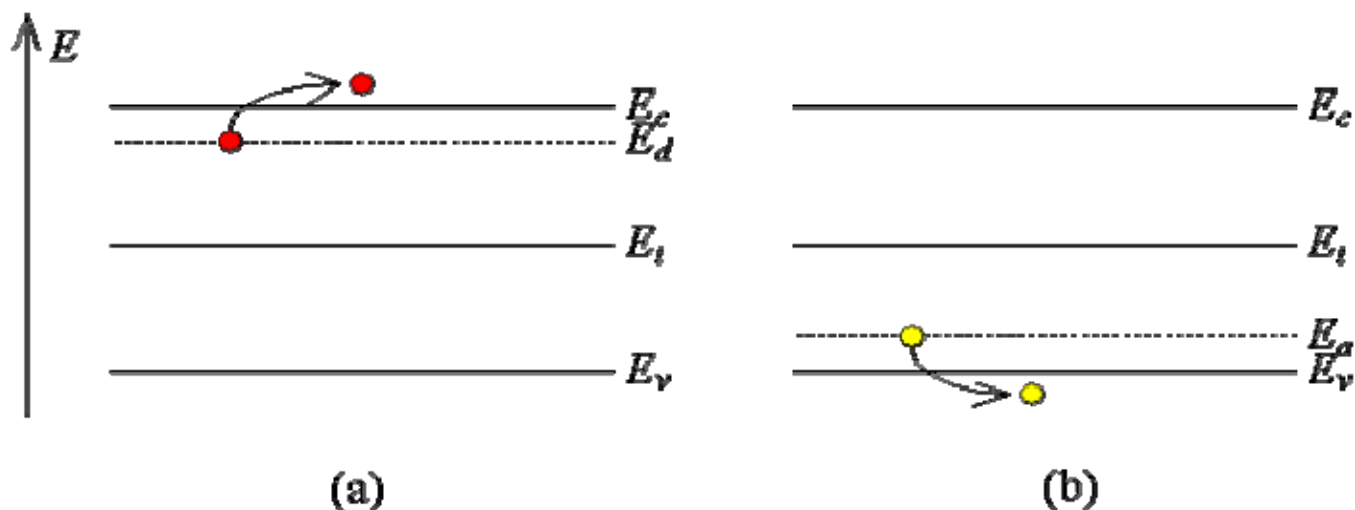
Figure 2.6.4 :
Ionization of a) a shallow donor and b) a shallow acceptor

A semiconductor doped with impurities, which are ionized (meaning that the impurity atoms either have donated or accepted an electron), will therefore contain free carriers. Shallow impurities are impurities, which require little energy - typically around the thermal energy, kT, or less - to ionize. Deep impurities require energies much larger than the thermal energy to ionize so that only a fraction of the impurities present in the semiconductor contribute to free carriers. Deep impurities, which are more than five times the thermal energy away from either band edge, are very unlikely to ionize. Such impurities can be effective recombination centers, in which electrons and holes fall and annihilate each other. Such deep impurities are also called traps.

Ionized donors provide free electrons in a semiconductor, which is then called n-type, while ionized acceptors provide free holes in a semiconductor, which we refer to as being a p-type semiconductor.

2.6.4.1 Dopants and impurities

The ionization of the impurities is dependent on the thermal energy and the position of the impurity level within the energy band gap as described by the impurity distribution functions discussed in section 2.5.3.

Shallow impurities readily ionize so that the free carrier density equals the impurity concentration. For shallow donors this implies that the electron density equals the donor concentration, or:

$$N_d^+ \cong N_d$$
(2.6.27)

While for shallow acceptors the hole density equals the acceptor concentration, or:

$$N_a^- \cong N_a$$
(2.6.28)

If a semiconductor contains both shallow donors and shallow acceptors it is called compensated since equal amounts of donor and acceptor atoms compensate each other, yielding no free carriers. The presence of shallow donors and shallow acceptors in a semiconductor cause the electrons given off by the donor atoms to fall into the acceptor state, which ionizes the acceptor atoms without yielding a free electron or hole. The resulting carrier density in compensated material, which contains both shallow donors and shallow acceptors, is approximately equal to the difference between the donor and acceptor concentration if the donor concentration is larger, yielding n-type material, or:

$$n_o \cong N_d^+ - N_a^-, \text{ if } N_d^+ - N_a^- \gg n_i$$
(2.6.29)

If the acceptor concentration is larger than the donor concentration, the hole density of the resulting p-type material equals the difference between the acceptor and donor concentration, or:

$$p_o \cong N_a^- - N_d^+, \text{ if } N_a^- - N_d^+ \gg n_i$$
(2.6.30)

2.6.4.2 Ionization energy model

The energy required to remove an electron from a donor atom can be approximated using a hydrogen-like model. After all, the donor atom consists of a positively charged ion and an electron just like the proton and electron of the hydrogen atom. The difference however is that the average distance, r, between the electron and the donor ion is much larger since the electron occupies one of the outer orbitals. This is illustrated by Figure 2.6.5.
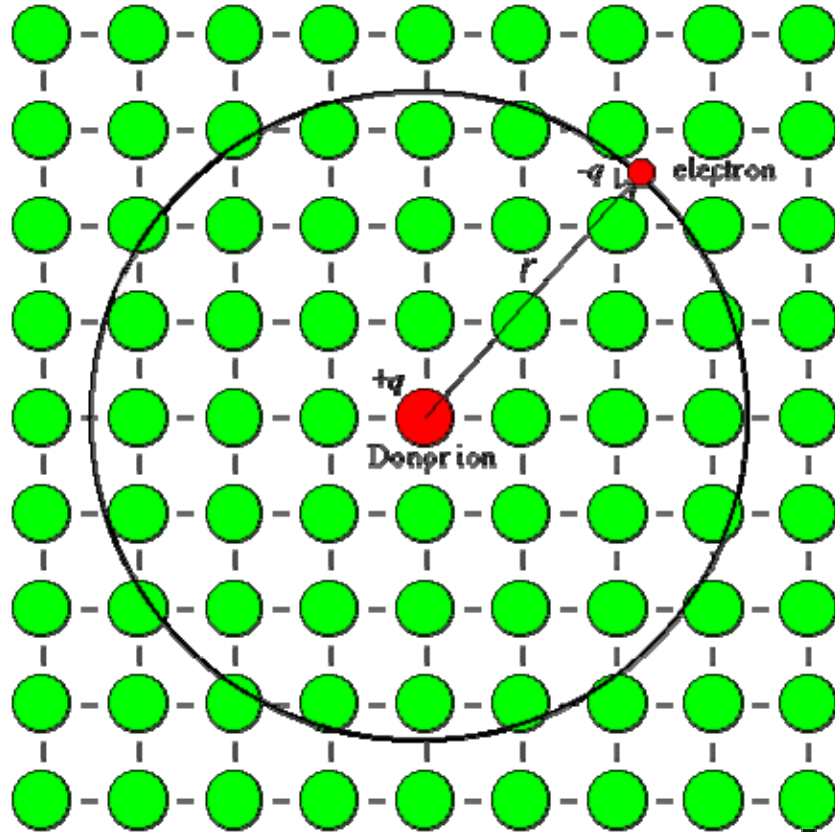
Figure 2.6.5:
Trajectory of an electron bound to a donor ion within a semiconductor crystal.

For shallow donors, this distance, r, is much larger than the inter-atomic spacing of the semiconductor crystal. The ionization energy, E_d, can be estimated by modifying equation (1.2.10), which describes the electron energy in a Hydrogen atom, yielding:

$$E_c - E_d = 13.6 \frac{m_{cond}^*}{m_0 \, \varepsilon_r^2} \, eV$$

(2.6.31)

where m_{cond}^* is the effective mass for conductivity calculations and er is the relative dielectric constant of the semiconductor. The ionization energy is calculated as the difference between the energy of a free electron and that of an electron occupying the lowest energy level, E_1.

Example 2.5
Calculate the ionization energy for shallow donors and acceptors in germanium and silicon using the hydrogen-like model.
Solution

Using the effective mass for conductivity calculations (Appendix 3) one finds the ionization energy for shallow donors in germanium to be:

$$E_c - E_d = 13.6\frac{m^*_{cond}}{m_0\,\varepsilon^2_r}\,eV = 13.6\frac{0.12}{16^2}\,eV = 6.4\,meV$$

The calculated ionization energies for donors and acceptors in germanium and silicon are provided below.

	Germanium	Silicon
donors	6.4 meV	13.8 meV
acceptors	11 2 meV	20.5 meV

Note that the actual ionization energies differ from these values and depend on th actual donor atom.

2.6.4.3 Analysis of non-degenerately doped semiconductors

The calculation of the electron density starts by assuming that the semiconductor is neutral, so that there is a zero charge density in the material. This is a reasonable assumption since a net charge density would result in an electric field. This electric field would move any mobile charge so that it eliminates any charge imbalance.

The charge density in a semiconductor depends on the free electron and hole density and on the ionized impurity densities. Ionized donors, which have given off an electron, are positively charged. Ionized acceptors, which have accepted an electron, are negatively charged. The total charge density is therefore given by:

$$\rho = q(p_o - n_o + N^+_d - N^-_a) = 0$$
(2.6.32)

The hole concentration in thermal equilibrium can be written as a function of the electron density by using the mass action law ([2.6.20](#)). This yields the following relation between the electron density and the ionized impurity densities:

$$n_o = \frac{n^2_i}{n_o} + N^+_d - N^-_a$$
(2.6.33)

Note that the use of the mass action law restricts the validity of this derivation to non-degenerate semiconductors as defined in section [2.6.2.2](#). Solving this quadratic equation yields a solution for the electron density, namely:

$$n_o = \frac{N^+_d - N^-_a}{2} + \sqrt{(\frac{N^+_d - N^-_a}{2})^2 + n^2_i}$$
(2.6.34)

The same derivation can be repeated for holes, yielding:

$$p_o = \frac{N_a^- - N_d^+}{2} + \sqrt{\left(\frac{N_a^- - N_d^+}{2}\right)^2 + n_i^2}$$

(2.6.35)

The above expressions provide the free carrier densities for compensated semiconductors assuming that all donors and acceptors are ionized.
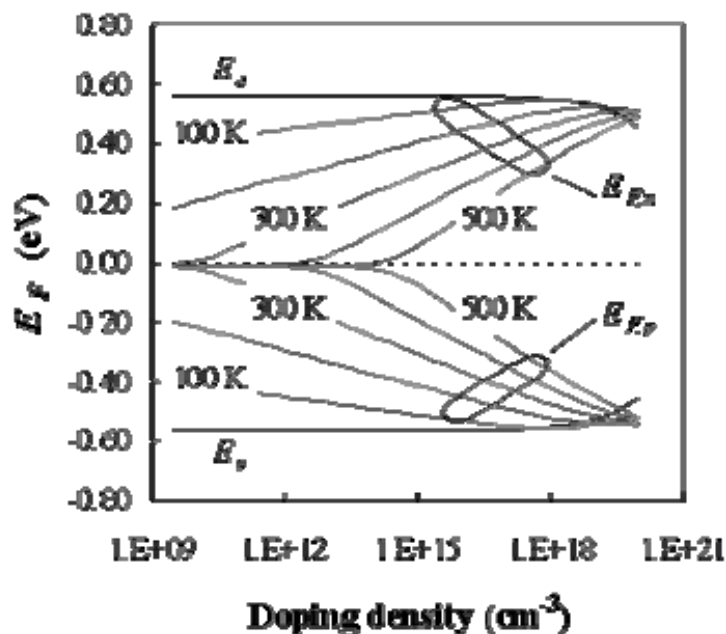
From the carrier densities, one then obtains the Fermi energies using equations (2.6.25) and (2.6.26) which are repeated below:

$$E_F = E_i + kT \ln \frac{n_o}{n_i}$$

(2.6.25)

or

$$E_F = E_i - kT \ln \frac{p_o}{n_i}$$

(2.6.26)

The Fermi energies in n-type and p-type silicon as a function of doping density is shown in Figure 2.6.6 for different temperatures:

Figure 2.6.6 :
Fermi energy of n-type and p-type silicon, $E_{F,n}$ and $E_{F,p}$, as a function of doping density at 100, 200, 300, 400 and 500 K. Shown are the conduction and valence band edges, E_c and E_v. The midgap energy is set to zero.

Figure 2.6.6 illustrates how the Fermi energies vary with doping density. The Fermi energy varies linearly, when plotting the density on a logarithmic scale, up to a doping density of 10^{18} cm^{-3}. This simple dependence requires that the semiconductor is neither intrinsic nor degenerate and that all the dopants are ionized. For compensated material, containing only shallow dopants, one uses the net doping density, $|N_d - N_a|$.

Example 2.6a
A germanium wafer is doped with a shallow donor density of $3n_i/2$. Calculate the electron and hole density.
Solution

The electron density is obtained from equation (2.6.34) and the hole density is obtained using the mass action law:

Example 2.6b
A silicon wafer is doped with a shallow acceptor doping of 10^{16} cm^{-3}. Calculate the electron and hole density.
Solution

Since the acceptor doping is much larger than the intrinsic density and much smaller than the effective density of states, the hole density equals: The electron density is then obtained using the mass action law The approach described in example 2.6a yields the same result.
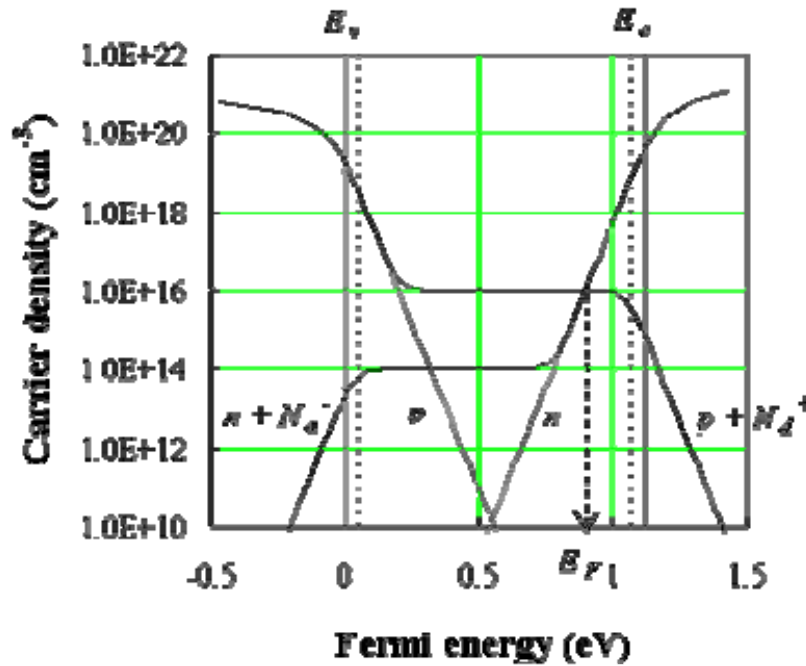
2.6.4.4 General analysis

A more general analysis takes also into account the fact that the ionization of the impurities is not 100%, but instead is given by the impurity distribution functions provided in section 2.5.3.

The analysis again assumes that there is no net charge in the semiconductor (charge neutrality). This also means that the total density of positively charged particles (holes and ionized donors) must equals the total density of negatively charged particles (electrons and ionized acceptors) yielding:

$$p_o + N_d^+ = n_o + N_a^-$$
(2.6.36)

The electron and hole densities are then written as a function of the Fermi energy. For non-degenerate semiconductors one uses equations (2.6.12) and (2.6.14), while the ionized impurity densities equal the impurity density multiplied with the probability of occupancy for the acceptors and one minus the probability of occupancy for the donors. The Joyce-Dixon approximation, described in section 2.6.2.3 is used to calculate the degenerate carrier densities.

A graphical solution to equation (2.6.36) above can be obtained by plotting both sides of the equation as a function of the Fermi energy as illustrated in Figure 2.6.7.
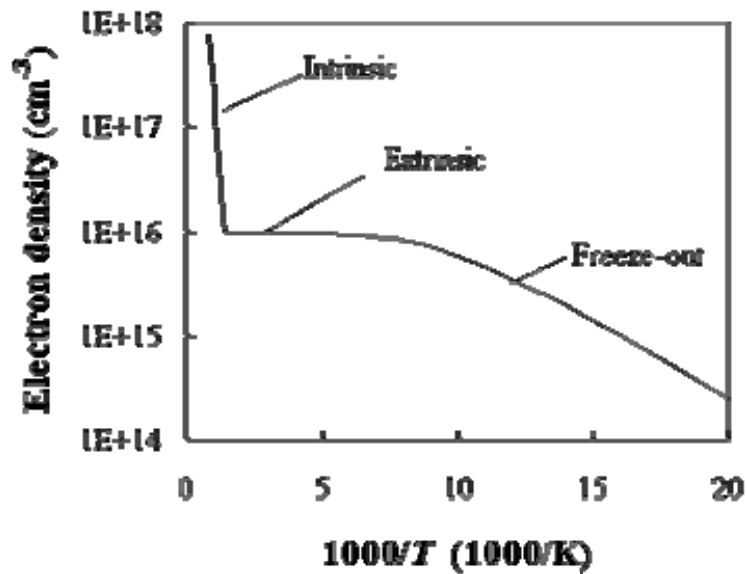
Figure 2.6.7 :
Graphical solution of the Fermi energy based on the general analysis. The value of the Fermi energy and the free carrier density is obtained at the intersection of the two curves, which represent the total positive and total negative charge in the semiconductor. N_a equals 10^{16} cm^{-3} and N_d equals 10^{14} cm^{-3}.

Figure 2.6.7 shows the positive and negative charge densities as well as the electron and hole densities as a function of the Fermi energy. The dotted lines indicate the position of the acceptor and donor energies. The Fermi energy is obtained at the intersection of both curves as indicated by the arrow.

This graphical solution is a very useful tool to explore the Fermi energy as a function of the doping densities, ionization energies and temperature.

Operation of devices over a wide temperature range requires a detailed knowledge of the carrier density as a function of temperature. At intermediate temperatures the carrier density approximately equals the net doping, $|N_a - N_d|$. Semiconductors, which satisfy this condition, are also called extrinsic semiconductors. The free carrier density increases at high temperatures for which the intrinsic density approaches the net doping density and decreases at low temperatures due to incomplete ionization of the dopants. The carrier density and Fermi energy are shown in Figure 2.6.8 for silicon doped with 10^{16} cm^{-3} donors and 10^{15} cm^{-3} acceptors:

Figure 2.6.8 :
Electron density and Fermi energy as a function of temperature in silicon with $N_d = 10^{16}$ cm^{-3}, $N_a = 10^{14}$ cm^{-3} and $E_c - E_d = E_a - E_v = 50$ meV. The activation energy at 70 K equals 27.4 meV.

At high temperatures, the carrier density equals the intrinsic carrier concentration, while at low temperatures the carrier density is dominated by the ionization of the donors.

The temperature dependence is related to an activation energy by fitting the carrier density versus $1/T$ on a semi-logarithmic scale to a straight line of the form $n_o(T) = C \exp(-E_A/kT)$, where C is a constant. At high temperatures this activation energy equals half the bandgap energy or $E_A = E_g/2$.

The temperature dependence at low temperatures is somewhat more complex as it depends on whether or not the material is compensated. Figure 2.6.8 was calculated for silicon containing both donors and acceptors. At 70 K the electron density is below the donor density but still larger than the acceptor density. Under such conditions the activation energy, E_A, equals half of the ionization energy of the donors or $(E_c - E_d)/2$. At lower temperatures where the electron density is lower than the acceptor density, the activation energy equals the ionization energy or $E_c - E_d$. This behavior is explained by the fact that the Fermi energy in compensated material is fixed at the donor energy. The donors levels are always partially empty as electrons are removed from the donor atoms to fill the acceptor energy levels. If the acceptor density is smaller than the electron density - as is true for uncompensated material - the Fermi energy does change with temperature and the activation energy approaches half of the ionization energy.

Lightly doped semiconductors suffer from freeze-out at relatively high temperature. Higher-doped semiconductors freeze-out at lower temperatures. Highly-doped semiconductors do not contain a single donor energy, but rather an impurity band which overlaps with the conduction or valence band. The overlap of the two bands results in free carriers even at zero Kelvin. Degenerately doping a semiconductor therefore eliminates freeze-out effects.

2.6.5. Non-equilibrium carrier densities

Up until now, we have only considered the thermal equilibrium carrier densities, n_o and p_o. However most devices of interest are not in thermal equilibrium. Keep in mind that a constant ambient constant temperature is not a sufficient condition for thermal equilibrium. In fact, applying a non-zero voltage to a device or illuminating it with light will cause a non-equilibrium condition, even if the temperature is constant.

To describe a system that is not in thermal equilibrium we assume that each of the carrier distributions is still in equilibrium with itself. Such assumption is justified on the basis that electrons readily interact with each other and interact with holes only on a much longer time scale. As a result the electron density can still be calculated using the Fermi-Dirac distribution function, but with a different value for the Fermi energy. The total carrier density for a non-degenerate semiconductor is then described by:

$$n = n_o + \delta n = n_i \exp(\frac{F_n - E_i}{kT})$$
(2.6.37)

Where δ_n is the excess electron density and F_n is the quasi-Fermi energy for the electrons. Similarly, the hole density can be expressed as:

$$p = p_o + \delta p = n_i \exp(\frac{E_i - F_p}{kT})$$
(2.6.38)

Where δp is the excess hole density and F_p is the quasi-Fermi energy for the holes.

Example 2.7
A piece of germanium doped with 10^{16} cm^{-3} shallow donors is illuminated with light generating 10^{15} cm^{-3} excess electrons and holes. Calculate the quasi-Fermi energies relative to the intrinsic energy and compare it to the Fermi energy in the absence of illumination.
Solution

The carrier densities when illuminating the semiconductor are:

$$n = n_o + \delta n = 10^{16} + 10^{15} = 1.1 \times 10^{16} \text{ cm}^{-3}$$

$$p = p_o + \delta p \simeq 10^{15} \text{ cm}^{-3}$$

and the quasi-Fermi energies are:

$$F_n - E_i = kT \ln \frac{n}{n_i} = 0.0259 \times \ln \frac{1.1 \times 10^{16}}{2 \times 10^{13}} = 163 \text{meV}$$

$$F_p - E_i = -kT \ln \frac{p}{n_i} = 0.0259 \times \ln \frac{1 \times 10^{15}}{2 \times 10^{13}} = -101 \text{meV}$$

In comparison, the Fermi energy in the absence of light equals

$$E_F - E_i = kT \ln \frac{n_o}{n_i} = 0.0259 \times \ln \frac{10^{16}}{2 \times 10^{13}} = 161\,\text{meV}$$

which is very close to the quasi-Fermi energy of the majority carriers.

Chapter 2: Semiconductor Fundamentals

2.7. Carrier Transport

A motion of free carriers in a semiconductor leads to a current. This motion can be caused by an electric field due to an externally applied voltage, since the carriers are charged particles. We will refer to this as carrier drift. In addition, carriers also move from regions where the carrier density is high to regions where the carrier density is low. This carrier transport mechanism is due to the thermal energy and the associated random motion of the carriers. We will refer to this transport mechanism as carrier diffusion. The total current in a semiconductor equals the sum of the drift and the diffusion current.

As one applies an electric field to a semiconductor, the electrostatic force causes the carriers to first accelerate and then reach a constant average velocity, v, due to collisions with impurities and lattice vibrations. The ratio of the velocity to the applied field is called the mobility. The velocity saturates at high electric fields reaching the saturation velocity. Additional scattering occurs when carriers flow at the surface of a semiconductor, resulting in a lower mobility due to surface or interface scattering mechanisms.

Diffusion of carriers is obtained by creating a carrier density gradient. Such gradient can be obtained by varying the doping density in a semiconductor or by applying a thermal gradient.

Both carrier transport mechanisms are related since the same particles and scattering mechanisms are involved. This leads to a relationship between the mobility and the diffusion constant called the Einstein relation.
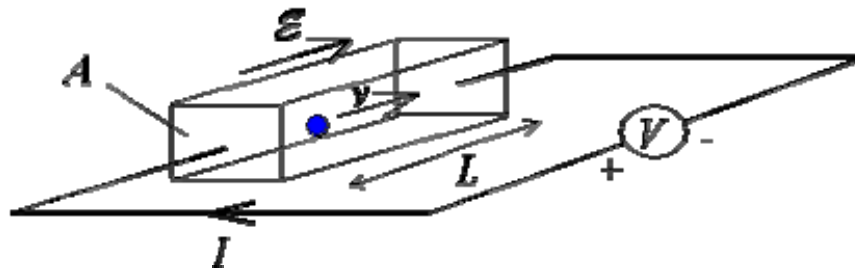
2.7.1. Carrier drift

The motion of a carrier drifting in a semiconductor due to an applied electric field is illustrated in Figure 2.7.1. The field causes the carrier to move with a velocity, v.

Figure 2.7.1 :
Drift of a carrier due to an applied electric field.

Assuming that all the carriers in the semiconductor move with the same velocity, the current can be expressed as the total charge in the semiconductor divided by the time needed to travel from one electrode to the other, or:

$$I = \frac{Q}{t_r} = \frac{Q}{L/v}$$

(2.7.1)

where t_r is the transit time of a particle, traveling with velocity, v, over the distance L. The current density can then be rewritten as a function of either the charge density, ρ, or the density of carriers, n in the semiconductor:

$$\vec{J} = \frac{Q}{AL}\vec{v} = \rho\vec{v} = qn\vec{v}$$

(2.7.2)

Carriers however do not follow a straight path along the electric field lines, but instead bounce around in the semiconductor and constantly change direction and velocity due to scattering. This behavior occurs even when no electric field is applied and is due to the thermal energy of the electrons. Electrons in a non-degenerate and non-relativistic electron gas have a thermal energy, which equals $kT/2$ per particle per degree of freedom. A typical thermal velocity at room temperature is around 10^7 cm/s, which exceeds the typical drift velocity in semiconductors. The carrier motion in the semiconductor in the absence and in the presence of an electric field can therefore be visualized as in Figure 2.7.2.
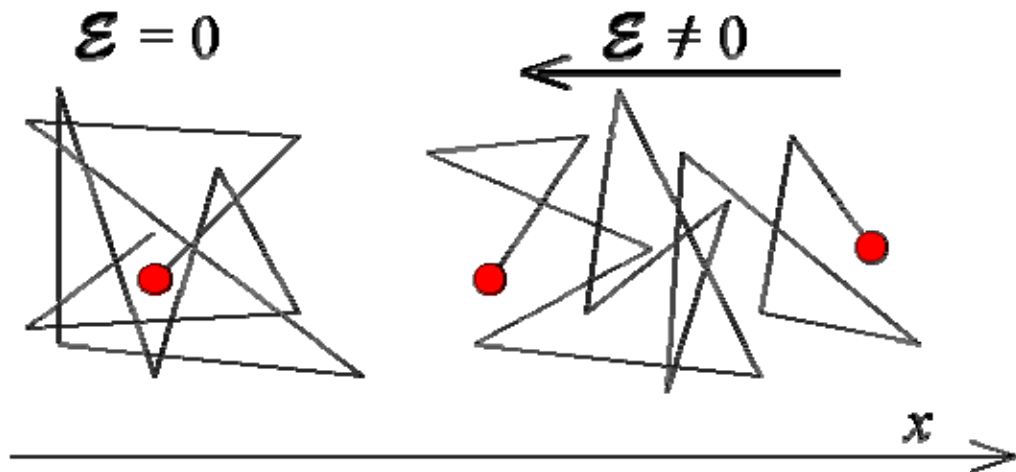
Figure 2.7.2 :
Random motion of carriers in a semiconductor with and without an applied electric field.

In the absence of an applied electric field, the carrier exhibits random motion and the carriers move quickly through the semiconductor and frequently change direction. When an electric field is applied, the random motion still occurs but in addition, there is on average a net motion along the direction of the field.

We now analyze the carrier motion considering only the average velocity, $\langle \vec{v} \rangle$ of the carriers. Applying Newton's law, we state that the acceleration of the carriers is proportional to the applied force:

$$\vec{F} = m\vec{a} = m\frac{d\langle \vec{v} \rangle}{dt}$$

(2.7.3)

The force consists of the difference between the electrostatic force and the scattering force due to the loss of momentum at the time of scattering. This scattering force equals the momentum divided by the average time between scattering events, so that:

$$\vec{F} = q\vec{\mathcal{E}} - \frac{m\langle \vec{v} \rangle}{\tau}$$

(2.7.4)

Combining both relations yields an expression for the average particle velocity:

$$q\vec{\mathcal{E}} = m\frac{d\langle \vec{v} \rangle}{dt} + \frac{m\langle \vec{v} \rangle}{\tau}$$

(2.7.5)

We now consider only the steady state situation in which the particle has already accelerated and has reached a constant average velocity. Under such conditions, the velocity is proportional to the applied electric field and we define the mobility as the velocity to field ratio:

$$\mu = \frac{\Delta |\bar{v}|}{|\bar{\mathcal{E}}|} = \frac{q\,\tau}{m}$$

(2.7.6)

The mobility of a particle in a semiconductor is therefore expected to be large if its mass is small and the time between scattering events is large.

The drift current, described by (2.7.2), can then be rewritten as a function of the mobility, yielding:

$$\bar{J} = qn\,\mu_n\bar{\mathcal{E}}$$

(2.7.7)

Throughout this derivation, we simply considered the mass, *m*, of the particle. However in order to incorporate the effect of the periodic potential of the atoms in the semiconductor we must use the effective mass, m^*, rather than the free particle mass:

$$\mu = \frac{q\,\tau}{m^*}$$

(2.7.8)

Example 2.8

Electrons in undoped gallium arsenide have a mobility of 8,800 cm^2/V-s. Calculate the average time between collisions. Calculate the distance traveled between two collisions (also called the mean free path). Use an average velocity of 10^7 cm/s.

Solution

The collision time, τ_c, is obtained from:

$$\tau_c = \frac{\mu_n m_e^*}{q} = \frac{0.88 \times 0.067 \times 9.1 \times 10^{-31}}{1.6 \times 10^{-19}} = 0.34 \, ps$$

where the mobility was first converted in MKS units.

The mean free path, *l*, equals:

$$l = v_{average}\tau_c = 10^7 \times 0.34 \times 10^{-12} = 34 \, nm$$

2.7.1.1 Impurity scattering

By impurities, we mean foreign atoms in the solid, which are efficient scattering centers especially when they have a net charge. Ionized donors and acceptors in a semiconductor are a common example of such impurities. The amount of scattering due to electrostatic forces between the carrier and the ionized impurity depends on the interaction time and the number of impurities. Larger impurity concentrations result in a lower mobility. The dependence on the interaction time helps to explain the temperature dependence. The interaction time is directly linked to the relative velocity of the carrier and the impurity, which is related to the thermal velocity of the carriers. This thermal velocity increases with the ambient temperature so that the interaction time increases. Thereby, the amount of scattering decreases, resulting in a mobility increase with temperature. To first order, the

mobility due to impurity scattering is proportional to $T^{3/2}/N_I$, where N_I is the density of charged impurities.

2.7.1.2 Lattice scattering

Scattering by lattice waves includes the absorption or emission of either acoustical or optical phonons. Since the density of phonons in a solid increases with temperature, the scattering time due to this mechanism will decrease with temperature as will the mobility. Theoretical calculations reveal that the mobility in non-polar semiconductors, such as silicon and germanium, is dominated by acoustic phonon interaction. The resulting mobility is expected to be proportional to $T^{-3/2}$, while the mobility due to optical phonon scattering only is expected to be proportional to $T^{-1/2}$. Experimental values of the temperature dependence of the mobility in germanium, silicon and gallium arsenide are provided in Table 2.7.1.

	Germanium	Silicon	Gallium Arsenide
Electron mobility	$\propto T^{-1.7}$	$\propto T^{-2.4}$	$\propto T^{-1.0}$
Hole mobility	$\propto T^{-2.3}$	$\propto T^{-2.2}$	$\propto T^{-2.1}$

Table 2.7.1 :
Temperature dependence of the mobility in germanium, silicon and gallium arsenide due to phonon scattering

2.7.1.3 Surface scattering

The surface and interface mobility of carriers is affected by the nature of the adjacent layer or surface. Even if the carrier does not transfer into the adjacent region, its wavefunction does extend over 1 to 10 nanometer, so that there is a non-zero probability for the particle to be in the adjacent region. The net mobility is then a combination of the mobility in both layers. For carriers in the inversion layer of a MOSFET, one finds that the mobility can be up to three times lower than the bulk value. This is due to the distinctly lower mobility of electrons in the amorphous silicon. The presence of charged surface states further reduces the mobility just as ionized impurities would.
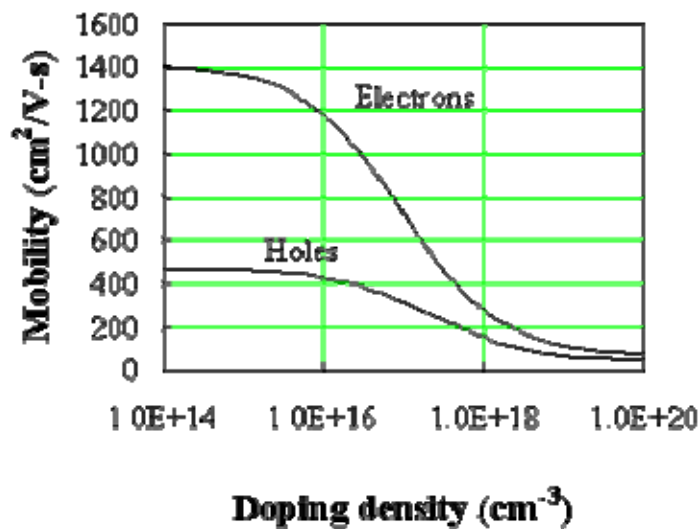
2.7.2. Carrier Mobility

2.7.2.1 Doping dependence

The mobility of electrons and holes in silicon at room temperature is shown in Figure2.7.3.

Figure 2.7.3 :
Electron and hole mobility versus doping density for silicon

The electron mobility and hole mobility have a similar doping dependence: For low doping concentrations, the mobility is almost constant and is primarily limited by phonon scattering. At higher doping concentrations, the mobility decreases due to ionized impurity scattering with the ionized doping atoms. The actual mobility also depends on the type of dopant. Figure 2.7.3 is for phosphorous and boron doped silicon.

Note that the mobility is linked to the total number of ionized impurities or the sum of the donor and acceptor densities. The free carrier density, as described in section 2.6.4.1 is to first order related to the difference between the donor and acceptor concentration.

The minority carrier mobility also depends on the total impurity density. The minority-carrier mobility can be approximated by the majority-carrier mobility in a material with the same number of impurities. The mobility at a particular doping density is obtained from the following empiric expression:

$$\mu = \mu_{min} + \frac{\mu_{max} - \mu_{min}}{1 + (\frac{N}{N_r})^{\alpha}}$$

(2.7.9)

	Arsenic	Phosphorous	Boron
μ_{min} (cm²/V-s)	52.2	68.5	44.9
μ_{max} (cm²/V-s)	1417	1414	470.5
N_r (cm⁻³)	9.68×10^{16}	9.20×10^{16}	2.23×10^{17}
α	0.68	0.711	0.719

Table 2.7.2 :
Parameters for calculation of the mobility as a function of the doping density

The resulting mobilities in units of cm²/V-s are listed for different doping densities in Table 2.7.3.

N	Arsenic	Phosphorous	Boron
10^{15} cm^{-3}	1359	1362	462
10^{16} cm^{-3}	1177	1184	429
10^{17} cm^{-3}	727	721	317
10^{18} cm^{-3}	284	277	153
10^{19} cm^{-3}	108	115	71

Table 2.7.3 :
Mobility in silicon for different doping densities

2.7.2.2 Conductivity and Resistivity

The conductivity of a material is defined as the current density divided by the applied electric field. Since the current density equals the product of the charge of the mobile carriers, their density and velocity, it can be expressed as a function of the electric field using the mobility. To include the contribution of electrons as well as holes to the conductivity, we add the current density due to holes to that of the electrons, or:

$$J = qnv_e + qpv_h = q(n\mu_n + p\mu_p)\mathcal{E}$$

(2.7.10)

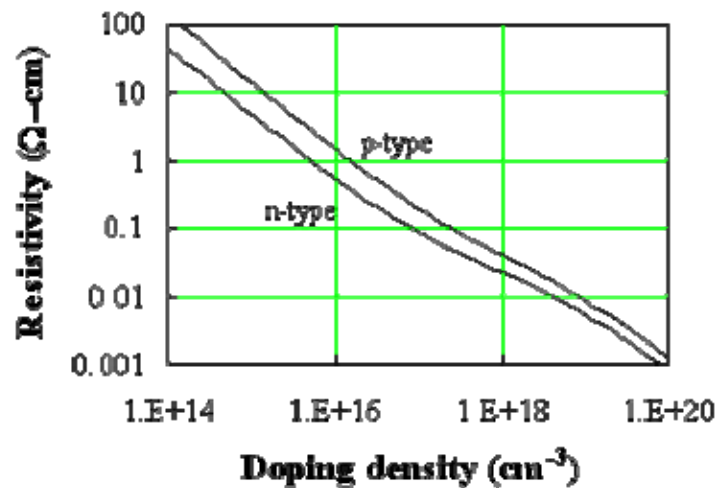The conductivity due to electrons and holes is then obtained from:

$$\sigma = \frac{\Delta J}{\mathcal{E}} = q(n\mu_n + p\mu_p)$$

(2.7.11)

The resistivity is defined as the inverse of the conductivity, namely:

$$\rho = \frac{1}{\sigma} = \frac{1}{q(\mu_n n + \mu_p p)}$$

(2.7.12)

The resulting resistivity as calculated with equation (2.7.12) is shown in Figure 2.7.4.

Figure 2.7.4 :
Resistivity of n-type and p-type silicon versus doping density

The sheet resistance concept is used to characterize both wafers and thin doped layers, since it is typically easier to measure the sheet resistance rather than the resistivity of the material. The sheet resistance of a uniformly-doped layer with resistivity, r, and thickness, t, is given by their ratio:

$$R_s = \frac{\rho}{t}$$
(2.7.13)

While the unit of the sheet resistance is Ohms, one refers to it as Ohms per square. This nomenclature comes in handy when the resistance of a rectangular piece of material with length, L, and width W must be obtained. It equals the product of the sheet resistance and the number of squares or:

$$R = R_s \frac{L}{W}$$
(2.7.14)

where the number of squares equals the length divided by the width. Figure 2.7.5 provides, as an example, the sheet resistance of a 14 mil thick silicon wafer which is n-type or p-type.
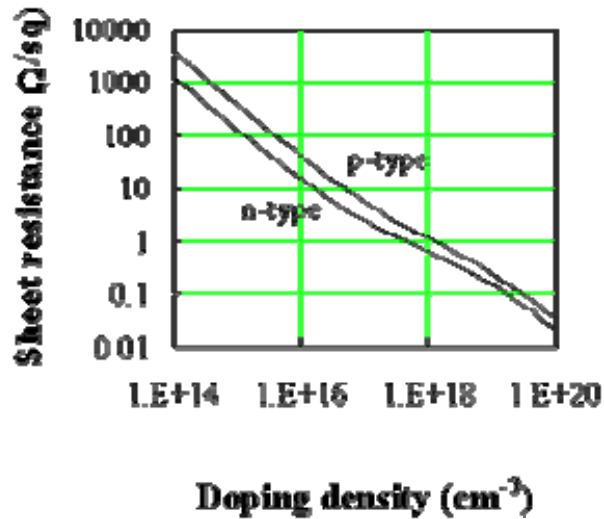
Figure 2.7.5 :
Sheet resistance of a 14 mil thick n-type and p-type silicon wafer versus doping density.

Example 2.9

A piece of silicon doped with arsenic ($N_d = 10^{17}$ cm^{-3}) is 100 µm long, 10 µm wide and 1 µm thick. Calculate the resistance of this sample when contacted one each end.

Solution

The resistivity of the silicon equals:

$$\rho = \frac{1}{q n \mu_n} = \frac{1}{1.6 \times 10^{-19} \times 10^{17} \times 727} = 0.086 \ \Omega cm$$

where the mobility was obtained from Table 2.7.3.

The resistance then equals:

$$R = \rho \frac{L}{Wt} = 0.086 \times \frac{100 \times 10^{-4}}{10 \times 10^{-4} \times 10^{-4}} = 8.6 \ k\Omega$$

An alternate approach is to first calculate the sheet resistance, R_s:

$$R_s = \frac{\rho}{t} = \frac{0.086}{10^{-4}} = 860 \ \Omega / square$$

From which one then obtains the resistance:

$$R = R_s \frac{L}{W} = 860 \times \frac{100 \times 10^{-4}}{10 \times 10^{-4}} = 8.6 \ k\Omega$$

2.7.3. Velocity saturation

The linear relationship between the average carrier velocity and the applied field breaks down when high fields are applied. As the electric field is increased, the average carrier velocity and the average carrier energy increases as well. When the carrier energy increases beyond the optical phonon energy, the probability of emitting an optical phonon increases abruptly. This mechanism causes the carrier velocity to saturate with increasing electric field. For carriers in silicon and other materials, which do not contain accessible higher bands, the velocity versus field relation can be described by:

$$v(\mathcal{E}) = \frac{\mu \mathcal{E}}{1 + \frac{\mu \mathcal{E}}{v_{sat}}}$$

(2.7.15)

The maximum obtainable velocity, v_{sat}, is referred to as the saturation velocity.

2.7.4. Carrier diffusion

Carrier diffusion is due to the thermal energy, kT, which causes the carriers to move at random even when no field is applied. This random motion does not yield a net motion of carriers nor does it yield a net current in material with a uniform carrier density as any carrier which leaves a specific location is on average replace by another one. However if a carrier gradient is present, the diffusion process will attempt to make the carrier density uniform: carriers diffuse from regions where the density is high to regions where the density is low. The diffusion process is not unlike the motion of sand on a vibrating table; hills as well as valleys are smoothed out over time.

In this section we will first derive the expression for the current due to diffusion and then combine it with the drift current to obtain the total drift-diffusion current.
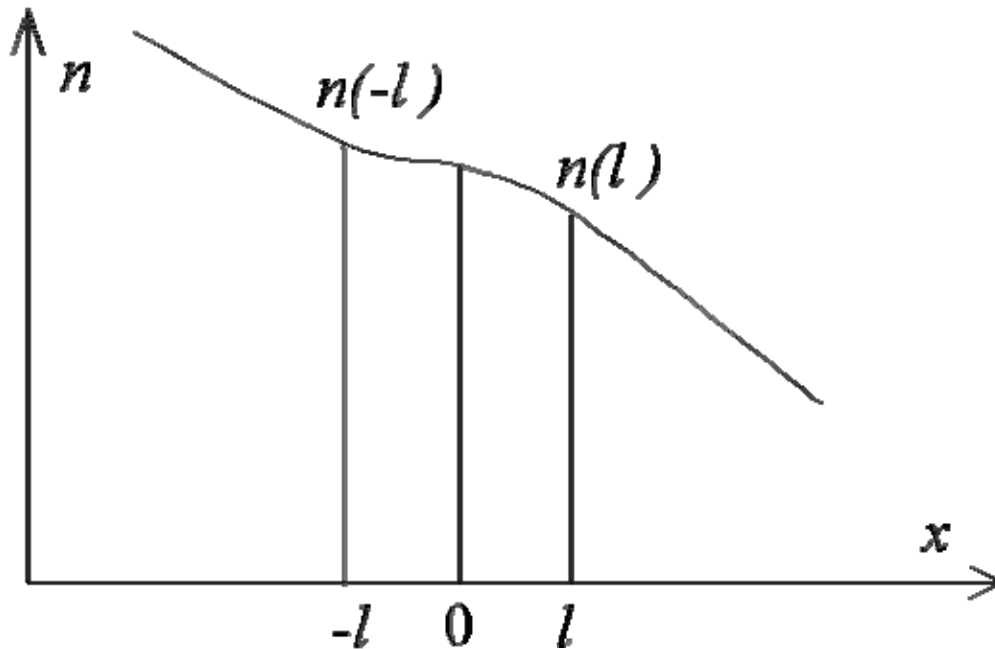
2.7.4.1 Diffusion current

The derivation is based on the basic notion that carriers at non-zero temperature (Kelvin) have an additional thermal energy, which equals $kT/2$ per degree of freedom. It is the thermal energy, which drives the diffusion process. At $T = 0$ K there is no diffusion.

While one should recognize that the random nature of the thermal energy would normally require a statistical treatment of the carriers, we instead will use average values to describe the process. Such approach is justified on the basis that a more elaborate statistical approach yields the same results. To further simplify the derivation, we will derive the diffusion current for a one-dimensional semiconductor in which carriers can only move along one direction.

We now introduce the average values of the variables of interest, namely the thermal velocity, v_{th}, the collision time, τ_c, and the mean free path, l. The thermal velocity is the average velocity of the carriers going in the positive or negative direction. The collision time is the time during which carriers will move with the same velocity before a collision occurs with an atom or with another carrier. The mean free path is the average length a carrier will travel between collisions. These three averages are related by:

$$v_{th} = \frac{l}{\tau}$$
(2.7.16)

Consider now the situation illustrated with Figure 2.7.6.

Figure 2.7.6 :
Carrier density profile used to derive the diffusion current expression

Shown is a variable carrier density, $n(x)$. Of interest are the carrier densities which are one mean free path away from $x = 0$, since the carriers which will arrive at $x = 0$ originate either at $x = -l$ or $x = l$. The flux at $x = 0$ due to carriers which originate at $x = -l$ and move from left to right equals:

$$\Phi_{n,left \rightarrow right} = \frac{1}{2}v_{th} n(x = -l)$$
(2.7.17)

where the factor 1/2 is due to the fact that only half of the carriers move to the left while the other half moves to the right. The flux at $x = 0$ due to carriers, which originate at $x = +l$ and move from right to left, equals:

$$\Phi_{n,right \rightarrow left} = \frac{1}{2}v_{th} n(x = l)$$
(2.7.18)

The total flux of carriers moving from left to right at $x = 0$ therefore equals:

$$\Phi_n = \Phi_{n,left \rightarrow right} - \Phi_{n,right \rightarrow left} = \frac{1}{2}v_{th}[n(x = -l) - n(x = l)]$$
(2.7.19)

Where the flux due to carriers moving from right to left is subtracted from the flux due to carriers moving from left to right. Given that the mean free path is small we can write the difference in densities divided by the distance between $x = -l$ and $x = l$ as the derivative of the carrier density:

$$\Phi_n = -l v_{th} \frac{n(x-l) - n(x--l)}{2l} = -l v_{th} \frac{dn}{dx}$$

(2.7.20)

The electron diffusion current equals this flux times the charge of an electron, or:

$$J_n = -q\Phi_n = q l v_{th} \frac{dn}{dx}$$

(2.7.21)

Typically, we will replace the product of the thermal velocity and the mean free path by a single parameter, namely the diffusion constant, D_n.

$$J_n = q D_n \frac{dn}{dx}$$

(2.7.22)

Repeating the same derivation for holes yields:

$$J_p = -q D_p \frac{dp}{dx}$$

(2.7.23)

We now further explore the relation between the diffusion constant and the mobility. At first, it seems that there should be no relation between the two since the driving force is distinctly different: diffusion is caused by thermal energy while an externally applied field causes drift. However one essential parameter in the analysis, namely the collision time, τ_c, should be independent of what causes the carrier motion.

We now combine the relation between the velocity, mean free path and collision time,

$$v_{th} = \frac{l}{\tau}$$

(2.7.24)

with the result from thermodynamics, stating that electrons carry a thermal energy which equals $kT/2$ for each degree of freedom. Applied to a one-dimensional situation, this leads to:

$$\frac{kT}{2} = \frac{m^* v_{th}^2}{2}$$

(2.7.25)

We now use these relations to rewrite the product of the thermal velocity and the mean free path as a function of the carrier mobility:

$$l v_{th} = \frac{m^* v_{th}^2}{q} \frac{q \, \tau}{m^*} = \frac{kT}{q} \mu$$

$$(2.7.26)$$

Using the definition of the diffusion constant we then obtain the following expressions which are often referred to as the Einstein relations:

$$D_n = \mu_n \frac{kT}{q} = \mu_n V_t$$

$$(2.7.27)$$

$$D_p = \mu_p \frac{kT}{q} = \mu_p V_t$$

$$(2.7.28)$$

Example 2.10

The hole density in an n-type silicon wafer ($N_d = 10^{17}$ cm^{-3}) decreases linearly from 10^{14} cm^{-3} to 10^{13} cm^{-3} between $x = 0$ and $x = 1$ μm. Calculate the hole diffusion current density.

Solution

The hole diffusion current density equals:

$$J_p = q D_p \frac{dp}{dx} = 1.6 \times 10^{-19} \times 8.2 \times \frac{9 \times 10^{13}}{10^{-4}} = 1.18 \text{ A/cm}^2$$

where the diffusion constant was calculated using the Einstein relation:

$$D_p = V_t \mu_p = 0.0259 \times 317 = 8.2 \text{ cm}^2/\text{s}$$

and the hole mobility in the n-type wafer was obtained from Table 2.7.3 as the hole mobility in a p-type material with the same doping density.

2.7.4.2 Total current

The total electron current is obtained by adding the current due to diffusion to the drift current, yielding:

$$J_n = q n \mu_n \mathcal{E} + q D_n \frac{dn}{dx}$$

$$(2.7.29)$$

and similarly for holes:

$$J_p = q p \mu_p \mathcal{E} - q D_p \frac{dp}{dx}$$

$$(2.7.30)$$

The total current is the sum of the electron and hole current densities multiplied with the area, A, perpendicular to the direction of the carrier flow:

$$I_{total} = A(J_n + J_p)$$
(2.7.31)

Chapter 2: Semiconductor Fundamentals

2.8. Carrier recombination and generation

Recombination of electrons and holes is a process by which both carriers annihilate each other: electrons occupy - through one or multiple steps - the empty state associated with a hole. Both carriers eventually disappear in the process. The energy difference between the initial and final state of the electron is released in the process. This leads to one possible classification of the recombination processes. In the case of radiative recombination, this energy is emitted in the form of a photon. In the case of non-radiative recombination, it is passed on to one or more phonons and in Auger recombination it is given off in the form of kinetic energy to another electron. Another classification scheme considers the individual energy levels and particles involved. These different processes are further illustrated with Figure 2.8.1.
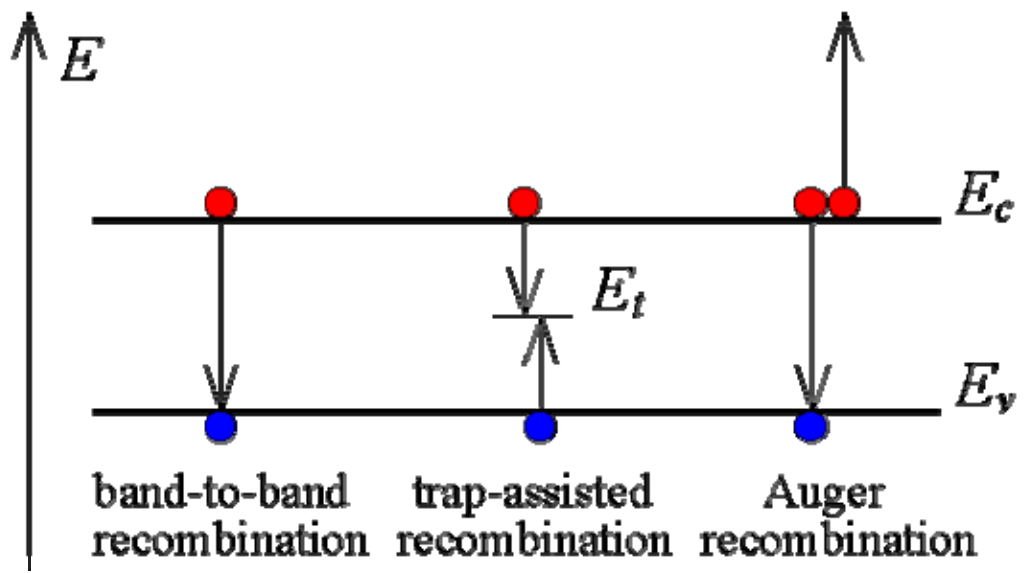
Figure 2.8.1 :
Carrier recombination mechanisms in semiconductors

Band-to-band recombination occurs when an electron falls from its conduction band state into the empty valence band state associated with the hole. This band-to-band transition is typically also a radiative transition in direct bandgap semiconductors.

Trap-assisted recombination occurs when an electron falls into a "trap", an energy level within the bandgap caused by the presence of a foreign atom or a structural defect. Once the trap is filled it cannot accept another electron. The electron occupying the trap, in a second step, falls into an empty valence band state, thereby completing the recombination process. One can envision this process as a two-step transition of an electron from the conduction band to the valence band or as the annihilation of the electron and hole, which meet each other in the trap. We will refer to this process as Shockley-Read-Hall (SRH) recombination.

Auger recombination is a process in which an electron and a hole recombine in a band-to-band transition, but now the resulting energy is given off to another electron or hole. The involvement of a third particle affects the recombination rate so that we need to treat Auger recombination differently from band-to-band recombination.

Each of these recombination mechanisms can be reversed leading to carrier generation rather than recombination. A single expression will be used to describe recombination as well as generation for each of the above mechanisms.

In addition, there are generation mechanisms, which do not have an associated recombination mechanism: generation of carriers by light absorption or a high-energy electron/particle beam. These processes are referred to as ionization processes. Impact ionization, which is the generation mechanism, associated with Auger recombination also belongs to this category. The generation mechanisms are illustrated with Figure 2.8.2.
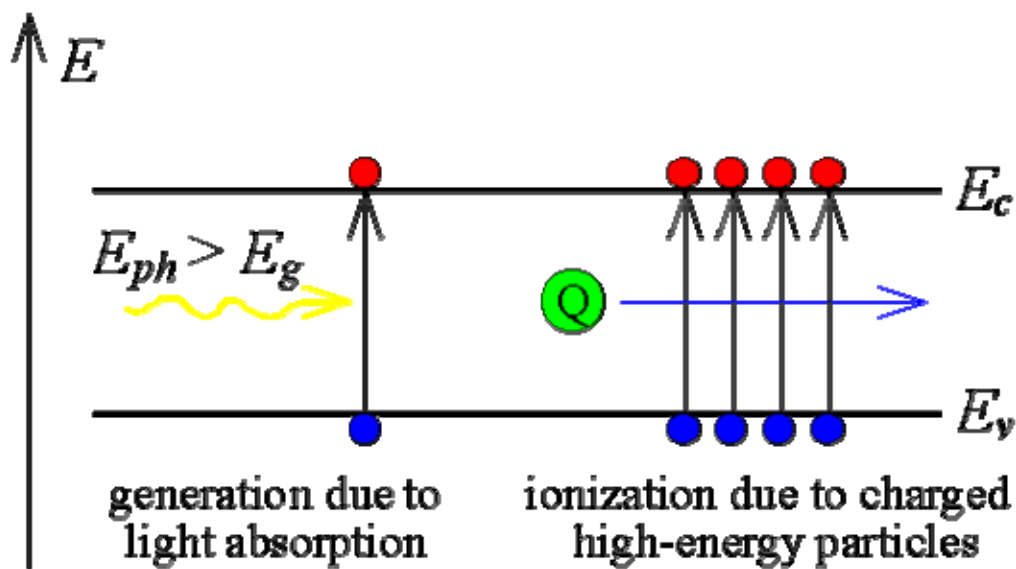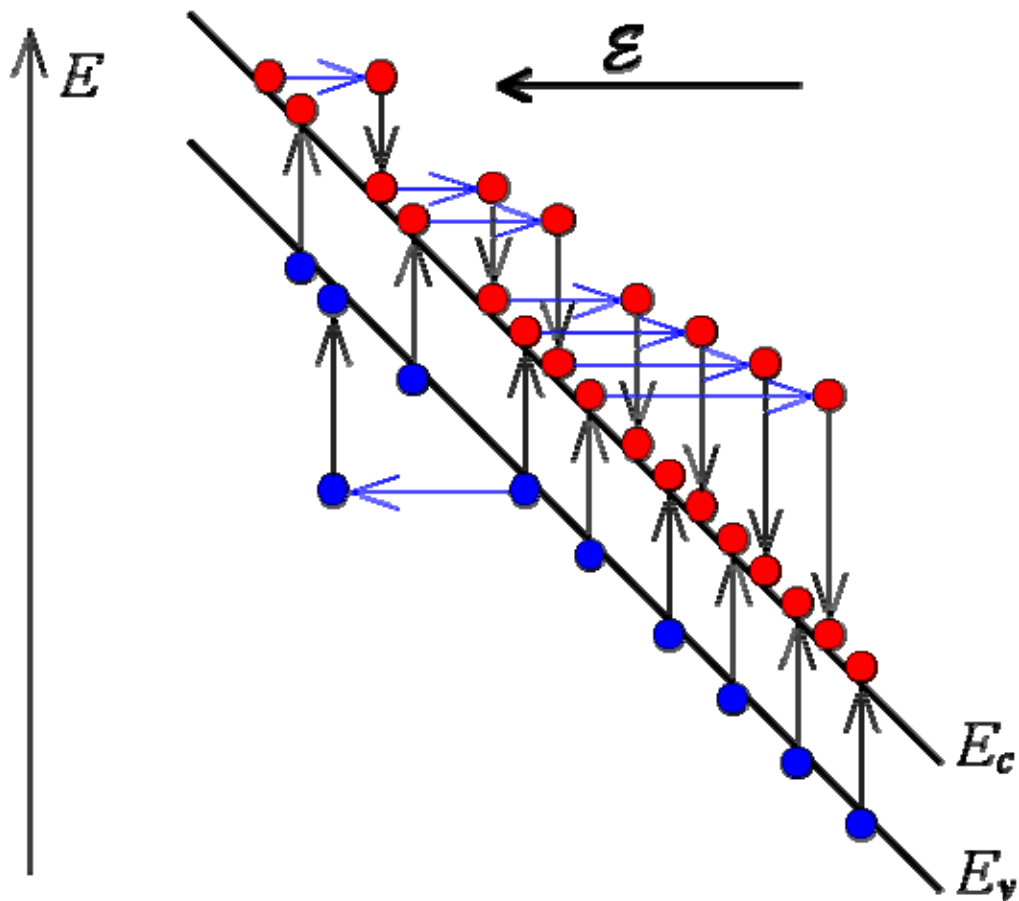

Figure 2.8.2 :
Carrier generation due to light absorption and ionization due to high-energy particle beams

Carrier generation due to light absorption occurs if the photon energy is large enough to lift an electron from the valence band into an empty conduction band state, generating one electron-hole

pair. The photon energy needs to be larger than the bandgap energy to satisfy this condition. The photon is absorbed in this process and the excess energy, E_{ph} - E_g, is added to the electron and the hole in the form of kinetic energy.

Carrier generation or ionization due to a high-energy beam consisting of charged particles is similar except that the available energy can be much larger than the bandgap energy so that multiple electron-hole pairs can be formed. The high-energy particle gradually loses its energy and eventually stops. This generation mechanism is used in semiconductor-based nuclear particle counters. As the number of ionized electron-hole pairs varies with the energy of the particle, one can also use such detector to measure the particle energy.

Finally, there is a generation process called impact ionization, the generation mechanism that is the counterpart of Auger recombination. Impact ionization is caused by an electron/hole with an energy, which is much larger/smaller than the conduction/valence band edge. The detailed mechanism is illustrated with Figure 2.8.3.

Figure 2.8.3:
Impact ionization and avalanche multiplication of electrons and holes in the presence of a large electric field.

The excess energy is given off to generate an electron-hole pair through a band-to-band transition. This generation process causes avalanche multiplication in semiconductor diodes under high reverse bias: As one carrier accelerates in the electric field it gains energy. The kinetic energy is given off to an electron in the valence band, thereby creating an electron-hole pair. The resulting

two electrons can create two more electrons which generate four more causing an avalanche multiplication effect. Electrons as well as holes contribute to avalanche multiplication.

2.8.1. Simple recombination-generation model

A simple model for the recombination-generation mechanisms states that the recombination-generation rate is proportional to the excess carrier density. It acknowledges the fact that no recombination takes place if the carrier density equals the thermal equilibrium value. The resulting expression for the recombination of electrons in a p-type semiconductor is given by:

$$U_n = R_n - G_n = \frac{n_p - n_{p0}}{\tau_n}$$
(2.8.1)

and similarly for holes in an n-type semiconductor:

$$U_p = R_p - G_p = \frac{p_n - p_{n0}}{\tau_p}$$
(2.8.2)

where the parameter τ can be interpreted as the average time after which an excess minority carrier recombines.

We will show for each of the different recombination mechanisms that the recombination rate can be simplified to this form when applied to minority carriers in a "quasi-neutral" semiconductor. The above expressions are therefore only valid under these conditions. The recombination rates of the majority carriers equals that of the minority carriers since in steady state recombination involves an equal number of holes and electrons. Therefore, the recombination rate of the majority carriers depends on the excess-minority-carrier-density as the minority carriers limit the recombination rate.

Recombination in a depletion region and in situations where the hole and electron density are close to each other cannot be described with the simple model and the more elaborate expressions for the individual recombination mechanisms must be used.

2.8.2. Band-to-band recombination

Band-to-band recombination depends on the density of available electrons and holes. Both carrier types need to be available in the recombination process. Therefore, the rate is expected to be proportional to the product of n and p. Also, in thermal equilibrium, the recombination rate must equal the generation rate since there is no net recombination or generation. As the product of n and p equals n_i^2 in thermal equilibrium, the net recombination rate can be expressed as:

$$U_{b-b} = b(np - n_i^2)$$
(2.8.3)

where b is the bimolecular recombination constant.

2.8.3. Trap assisted recombination

The net recombination rate for trap-assisted recombination is given by:

$$U_{SHR} = \frac{pn - n_i^2}{p + n + 2n_i \cosh\left(\dfrac{E_i - E_t}{kT}\right)} N_t v_{th} \sigma$$

(2.8.4)

This expression can be further simplified for $p \gg n$ to:

$$U_n = R_n - G_n = \frac{n_p - n_{p0}}{\tau_n}$$

(2.8.5)

and for $n \gg p$ to:

$$U_p = R_p - G_p = \frac{p_n - p_{n0}}{\tau_p}$$

(2.8.6)

were

$$\tau_n = \tau_p = \frac{1}{N_t v_{th} \sigma}$$

(2.8.7)

2.8.4. Surface recombination

Recombination at semiconductor surfaces and interfaces can have a significant impact on the behavior of devices. This is because surfaces and interfaces typically contain a large number of recombination centers because of the abrupt termination of the semiconductor crystal, which leaves a large number of electrically active dangling bonds. In addition, the surfaces and interfaces are more likely to contain impurities since they are exposed during the device fabrication process. The net recombination rate due to trap-assisted recombination and generation is given by:

$$U_{s,SHR} = \frac{pn - n_i^2}{p + n + 2n_i \cosh\left(\dfrac{E_i - E_{st}}{kT}\right)} N_{st} v_{th} \sigma_s$$

(2.8.8)

This expression is almost identical to that of Shockley-Hall-Read recombination. The only difference is that the recombination is due to a two-dimensional density of traps, N_{ts}, as the traps only exist at the surface or interface.

This equation can be further simplified for minority carriers in a quasi-neutral region. For instance for electrons in a quasi-neutral p-type region $p \gg n$ and $p \gg n_i$ so that for $E_i = E_{st}$, it can be simplified to:

$$U_{s,n} = R_{s,n} - G_{s,n} = v_s(n_p - n_{p0})$$
(2.8.9)

where the recombination velocity, v_s, is given by:

$$v_s = N_{st} v_{th} \sigma_s$$
(2.8.10)

2.8.5. Auger recombination

Auger recombination involves three particles: an electron and a hole, which recombine in a band-to-band transition and give off the resulting energy to another electron or hole. The expression for the net recombination rate is therefore similar to that of band-to-band recombination but includes the density of the electrons or holes, which receive the released energy from the electron-hole annihilation:

$$U_{Auger} = \Gamma_n n(np - n_i^2) + \Gamma_p p(np - n_i^2)$$
(2.8.11)

The two terms correspond to the two possible mechanisms.

2.8.6. Generation due to light

Carriers can be generated in semiconductors by illuminating the semiconductor with light. The energy of the incoming photons is used to bring an electron from a lower energy level to a higher energy level. In the case where an electron is removed from the valence band and added to the conduction band, an electron-hole pair is generated. A necessary condition for this to happen is that the energy of the photon, E_{ph}, is larger than the bandgap energy, E_g. As the energy of the photon is given off to the electron, the photon no longer exists.

If each absorbed photon creates one electron-hole pair, the electron and hole generation rates are given by:

$$G_{p,light} = G_{n,light} = \alpha \frac{P_{opt}(x)}{E_{ph} A}$$
(2.8.12)

where α is the absorption coefficient of the material at the energy of the incoming photon. The absorption of light in a semiconductor causes the optical power to decrease with distance. This effect is described mathematically by:

$$\frac{d P_{opt}(x)}{dx} = -\alpha P_{opt}(x)$$
(2.8.13)

The calculation of the generation rate of carriers therefore requires first a calculation of the optical power within the structure from which the generation rate can then be obtained using (2.8.12).

Example 2.11
Calculate the electron and hole densities in an n-type silicon wafer ($N_d = 10^{17}$ cm^{-3}) illuminated uniformly with 10 mW/cm^2 of red light ($E_{ph} = 1.8$ eV). The absorption coefficient of red light in silicon is 10^{-3} cm^{-1}. The minority carrier lifetime is 10 μs.
Solution

The generation rate of electrons and holes equals:

$$G_n = G_p = \alpha \frac{P_{opt}}{E_{ph}A} = 10^{-3} \frac{10^{-2}}{1.8 \times 1.6 \times 10^{-19}} = 3.5 \times 10^{13} \ cm^{-3}s^{-1}$$

where the photon energy was converted into Joules. The excess carrier densities are then obtained from:

$$\delta n = \delta p = \tau_p G_p = 10 \times 10^{-6} \times 3.5 \times 10^{13} = 3.5 \times 10^{8} \ cm^{-3}s^{-1}$$

The excess carrier densities are then obtained from: So that the electron and hole densities equal:

$$n = n_o + \delta n = 10^{17} + 3.5 \times 10^{13} = 10^{17} \ cm^{-3}s^{-1}$$

# Chapter 2: Semiconductor Fundamentals

2.9. Continuity equation

2.9.1. Derivation

The continuity equation describes a basic concept, namely that a change in carrier density over time is due to the difference between the incoming and outgoing flux of carriers plus the generation and minus the recombination. The flow of carriers and recombination and generation rates are illustrated with Figure 2.9.1.
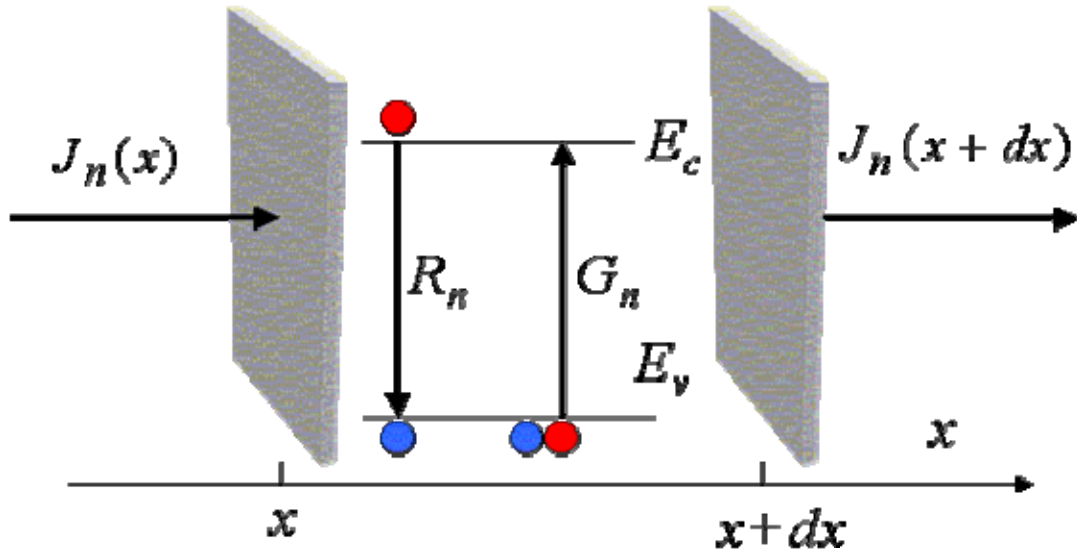
Figure 2.9.1 :
Electron currents and possible recombination and generation processes

The rate of change of the carriers between x and $x + dx$ equals the difference between the incoming flux and the outgoing flux plus the generation and minus the recombination:

$$\frac{\partial n(x,t)}{\partial t} A\, dx = (\frac{J_n(x)}{-q} - \frac{J_n(x+dx)}{-q}) A + (G_n(x,t) - R_n(x,t))\, A\, dx$$

(2.9.1)

where $n(x,t)$ is the carrier density, A is the area, $G_n(x,t)$ is the generation rate and $R_n(x,t)$ is the recombination rate. Using a Taylor series expansion,

$$J_n(x + dx) = J_n(x) + \frac{d\, J_n(x)}{dx} dx$$

(2.9.2)

this equation can be formulated as a function of the derivative of the current:

$$\frac{\partial n(x,t)}{\partial t} = \frac{1}{q} \frac{\partial J_n(x,t)}{\partial x} + G_n(x,t) - R_n(x,t)$$

(2.9.3)

and similarly for holes one finds:

$$\frac{\partial p(x,t)}{\partial t} = -\frac{1}{q} \frac{\partial J_p(x,t)}{\partial x} + G_p(x,t) - R_p(x,t)$$

(2.9.4)

A solution to these equations can be obtained by substituting the expression for the electron and hole current, (2.7.29) and (2.7.30). This then yields two partial differential equations as a function

of the electron density, the hole density and the electric field. The electric field itself is obtained from Gauss's law.

$$\frac{\partial n(x,t)}{\partial t} =$$

$$\mu_n n \frac{\partial \mathcal{E}(x,t)}{\partial x} + \mu_n \mathcal{E} \frac{\partial n(x,t)}{\partial x} + D_n \frac{\partial^2 n(x,t)}{\partial x^2} + G_n(x,t) - R_n(x,t)$$

(2.9.5)

$$\frac{\partial p(x,t)}{\partial t} =$$

$$- \mu_p p \frac{\partial \mathcal{E}(x,t)}{\partial x} - \mu_p \mathcal{E} \frac{\partial p(x,t)}{\partial x} + D_p \frac{\partial^2 p(x,t)}{\partial x^2} + G_p(x,t) - R_p(x,t)$$

(2.9.6)

A generalization in three dimensions yields the following continuity equations for electrons and holes:

$$\frac{\partial n(x,y,z,t)}{\partial t} = \frac{1}{q} \nabla J_n(x,y,z,t) + G_n(x,y,z,t) - R_n(x,y,z,t)$$

(2.9.7)

$$\frac{\partial p(x,y,z,t)}{\partial t} = -\frac{1}{q} \nabla J_p(x,y,z,t) + G_p(x,y,z,t) - R_p(x,y,z,t)$$

(2.9.8)

2.9.2. The diffusion equation

In the quasi-neutral region - a region containing mobile carriers, where the electric field is small - the current is due to diffusion only. In addition, we can use the simple recombination model for the net recombination rate. This leads to the time-dependent diffusion equations for electrons in p-type material and for holes in n-type material:

$$\frac{\partial n(x,t)}{\partial t} = D_n \frac{\partial^2 n_p(x,t)}{\partial x^2} - \frac{n_p(x,t) - n_{p0}}{\tau_n}$$

(2.9.9)

$$\frac{\partial p(x,t)}{\partial t} = D_p \frac{\partial^2 p_n(x,t)}{\partial x^2} - \frac{p_n(x,t) - p_{n0}}{\tau_p}$$

(2.9.10)

2.9.3. Steady state solution to the diffusion equation

In steady state, the partial derivatives with respect to time are zero, yielding:

$$0 = D_n \frac{d^2 n_p(x)}{dx^2} - \frac{n_p(x) - n_{p0}}{\tau_n}$$

(2.9.11)

$$0 = D_p \frac{d^2 p_n(x)}{dx^2} - \frac{p_n(x) - p_{n0}}{\tau_p}$$

(2.9.12)

The general solution to these second order differential equations are:

$$n_p(x \le -x_p) = n_{p0} + Ce^{-(x+x_p)/L_n} + De^{(x+x_p)/L_n}$$

(2.9.13)

$$p_n(x \ge x_n) = p_{n0} + Ae^{-(x-x_n)/L_p} + Be^{(x-x_n)/L_p}$$

(2.9.14)

where L_n and L_p are the diffusion lengths given by:

$$L_n = \sqrt{D_n \tau_n}$$

(2.9.15)

$$L_p = \sqrt{D_p \tau_p}$$

(2.9.16)

The diffusion constants, D_n and D_p, are obtained using the Einstein relations (2.7.27) and (2.7.28). The diffusion equations can also be written as a function of the excess carrier densities, δn and δp, which are related to the total carrier densities, n and p, and the thermal equilibrium densities, n_0 and p_0, by:

$$n = n_0 + \delta n$$

(2.9.17)

$$p = p_0 + \delta p$$

(2.9.18)

yielding:

$$0 = \frac{d^2(\delta n_p)}{dx^2} - \frac{\delta n_p}{L_n^2}$$

(2.9.19)

$$0 = \frac{d^2(\delta p_n)}{dx^2} - \frac{\delta p_n}{L_p^2}$$

(2.9.20)

The diffusion equation will be used to calculate the diffusion current in p-n junctions and bipolar transistors.

Chapter 2: Semiconductor Fundamentals

2.10. The drift-diffusion model

The drift-diffusion model of a semiconductor is frequently used to describe semiconductor devices. It contains all the features described in this chapter.

Starting with Chapter 3, we will apply the drift-diffusion model to a variety of different devices. To facilitate this analysis, we present here a simplified drift-diffusion model, which contains all the essential features. This model results in a set of ten variables and ten equations.

The assumptions of the simplified drift-diffusion model are:

Full ionization: all dopants are assumed to be ionized (shallow dopants)

Non-degenerate: the Fermi energy is assumed to be at least 3 kT below/above the conduction/valence band edge.

Steady state: All variables are independent of time

Constant temperature: The temperature is constant throughout the device.

The ten variables are the following:

ρ, the charge density

n, the electron density

p, the hole density

$\mathcal{E}$, the electric field

ϕ, the potential

E_i, the intrinsic energy

F_n, the electron quasi-Fermi energy

F_p, the hole quasi-Fermi energy

J_n, the electron current density

J_p, the hole current density

The ten equations are:

Charge density equation

$$\rho = q(p - n + N_d^+ - N_a^-)$$
(2.10.1)

Electric field and potential equations

$$\frac{d\mathcal{E}}{dx} = \frac{\rho}{\varepsilon}$$
(2.10.2)

$$\frac{d\phi}{dx} = -\mathcal{E}$$
(2.10.3)

$$\frac{dE_i}{dx} = q\mathcal{E}$$
(2.10.4)

Carrier density equations

$$n = n_i e^{(F_n - E_i)/kT}$$
(2.10.5)

$$p = n_i e^{(E_i - F_p)/kT}$$
(2.10.6)

Drift and diffusion current equations

$$J_n = qn\mu_n\mathcal{E} + qD_n\frac{dn}{dx}$$
(2.10.7)

$$J_p = qp\mu_p\mathcal{E} - qD_p\frac{dp}{dx}$$
(2.10.8)

Continuity equation in steady state with SHR recombination

$$0 = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{np - n_i^2}{n + p + 2n_i \cosh(\frac{E_t - E_i}{kT})}\frac{1}{\tau}$$
(2.10.9)

$$0 = -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{np - n_i^2}{n + p + 2n_i \cosh(\frac{E_t - E_i}{kT})}\frac{1}{\tau}$$
(2.10.10)