Chapter 4

Carrier Drift and Diffusion

[version 3.1 (updated: July 31, 2012)] ©Prentice Hall

Outline

This chapter deals with two key processes by which carriers flow in semiconductors: drift and diffusion. Carrier drift originates from the electrostatic force that an electric field exerts over electrons and holes. Diffusion is the movement of particles in response to a concentration gradient. Carrier drift and diffusion are at the heart of device operation. Electrons, for example, drift from source to drain in an n-channel MOSFET. The magnitude of the current and the time that carriers take to traverse the device are determined by the physics of drift in the channel. In npn bipolar transistors, a combination of electron diffusion and drift in the base takes electrons from the emitter to the collector.

In order to understand the physics of carrier drift and diffusion in semiconductors it is essential to develop an appreciation for the random motion of electrons and holes at finite temperatures. In the time scale of interest in most microelectronic devices, carriers undergo many collisions with the lattice, with impurities, and with each other. An important implication of this is that carriers are unlikely to maintain energy distributions that are too different from equilibrium over any meaningful periods of time. For most time scales of interest, the carriers can be considered to be in a "quasi-equilibrium" state in which, with some care, the concept of Fermi level remains useful. In these situations we define two quasi-Fermi levels, one for electrons and another one for holes. These allow us to graph simple and intuitive energy band diagrams that effectively convey a great deal of information about device operation.

Finally, in this chapter we also learn to compute the equilibrium carrier concentrations in situations where dopant distributions are non-uniform in space. In these instances, it is important to realize that a dynamic balance between drift and diffusion exists with no net carrier flow. Non-uniformly doped regions are pervasive and for the most part unavoidable in modern devices. Often, device characteristics are tailored through the engineering of the doping profiles. The aversion that reasonably extrinsic semiconductors have to the presence of volume charge implies that in many practical circumstances, the majority carrier concentration closely tracks the doping distribution. In other words, the semiconductor remains *quasi-neutral*. This greatly simplifies the treatment of non-uniformly doped regions.

This is a chapter rich in concepts that will be extensively used throughout the rest of the book. It is important to understand this material well before proceeding much further.

4.1 Thermal motion

In this section we look at the movement of carriers in a semiconductor in thermal equilibrium. We first look at a very idealized situation of a perfectly periodic solid. We then introduce lattice vibrations and other perturbations to this picture and consider their impact on carrier movement.

4.1.1 Thermal velocity

Let us consider a semiconductor in thermal equilibrium at a finite temperature. Let us assume a pure semiconductor with a perfect crystalline periodicity in which nothing changes in space (uniform situation) and all the surfaces are very far away. In Ch. 2 we learned how to compute the electron and hole concentrations under a variety of conditions. In this chapter we take an interest in the movement of electrons and holes in the lattice.

If we think of a perfectly ideal semiconductor in which electrons and holes can move around without bumping into anything, at what velocity would they be traveling? We can certainly answer right away that in thermal equilibrium, by definition, the average velocity of the ensemble of electrons or holes has to be zero. This does not mean that the velocity of an individual electron or hole with a certain energy above or below their respective band edges is also zero. In fact, we have already discussed that the kinetic energy of carriers increases as they occupy states further away from the corresponding band edge. This means that their velocity increases too.

In general, the relationship between velocity (or momentum) and energy for carriers in a semiconductor is not simple. It can be computed in band calculations and the result is very specific to each semiconductor. It is beyond the scope of this book to discuss the energy-momentum (E-k) diagrams, as these relationships are often referred to. For the purposes of this book, we can go quite far with a simple formalism that works reasonably well for most semiconductors close to the bottom of the bands. In this approach, as in classical mechanics, the carrier kinetic energy increases quadratically with velocity. This suggests that the electron velocity as a function of energy above the conduction band edge follows a simple square-root expression given by:

$$v_e(E) = \sqrt{\frac{2(E - E_c)}{m_{ce}^*}} \tag{4.1}$$

A similar equation can be written for holes with the kinetic energy increasing downwards from the valence band edge.

In Eq. 4.1 m_{ce}^* is known as the *electron conductivity effective mass.* This is the effective mass that connects velocity and kinetic energy of an electron. The conductivity effective mass is in general different from the free electron mass because in a semiconductor electrons are not in vacuum but are immersed in the periodic potential of the lattice. Also, m_{ce}^* is different from the density of states effective mass discussed in Ch. 2. Appendix B at the end of the book lists m_{ce}^*/m_o and m_{ch}^*/m_o for Si and GaAs.

Eq. 4.1 gives the magnitude of the velocity of an electron with a certain energy in the conduction band. Electrons higher up in the conduction band move faster than those close to the band edge. The average magnitude of the velocity for an electron population in thermal equilibrium receives the name of *thermal velocity.* We can derive an expression for the thermal velocity by appropriately averaging the velocities of all electrons in the conduction band in the following way:

$$v_{the} = \frac{\int_{E_c}^{\infty} v_e(E) n_o(E)\, dE}{\int_{E_c}^{\infty} n_o(E)\, dE} \tag{4.2}$$

Here, $n_o(E)$ represents the electron distribution in energy in the conduction band as given by Eq. 2.21). Using this equation as well as Eq. 4.1, we can rewrite 4.2, as:

$$v_{the} = \sqrt{\frac{2}{m^*_{ce}} \frac{\int_{E_c}^{\infty} \frac{(E-E_c)}{1+\exp\frac{E-E_F}{kT}}\, dE}{\int_{E_c}^{\infty} \frac{\sqrt{E-E_c}}{1+\exp\frac{E-E_F}{kT}}\, dE}} \tag{4.3}$$

Under Maxwell-Boltzmann statistics and redefining variables as in Sec. 2.4.2, we can rewrite this expression as:

$$v_{the} \simeq \sqrt{\frac{2kT}{m^*_{ce}} \frac{\int_0^{\infty} \eta e^{-\eta} d\eta}{\int_0^{\infty} \eta^{1/2} e^{-\eta} d\eta}} \tag{4.4}$$

The integrals in Eq. 4.4 belong to a class called the Gamma function. This is defined in Advanced Topic AT4.1 where some properties of this function are also given. Using the results shown there, we can rewrite 4.4 as:

$$v_{the} = \sqrt{\frac{2kT}{m^*_{ce}} \frac{\Gamma(2)}{\Gamma(\frac{3}{2})}} = \sqrt{\frac{8}{\pi} \frac{kT}{m^*_{ce}}} \tag{4.5}$$

A similar equation applies to holes. The temperature dependence that emerges in this equation arises from the electron distribution in energy. The higher the temperature, the higher the average energy of the electrons in the conduction band and the higher their average velocity.

Even though we have been discussing a highly idealized situation, the thermal velocity is an important benchmark for carrier transport in semiconductors that will make frequent appearances in this book. If one puts numbers in Eq. 4.3 for electrons in Si at room temperature ($m^*_{ce} = 0.28\ m_o$), one finds that the electron thermal velocity is about $2.0 \times 10^7\ cm/s$. For holes ($m^*_{ch} = 0.41\ m_o$), the thermal velocity is $1.7 \times 10^7\ cm/s$. Holes are somehow slower than electrons.

4.1.2 Scattering

A real semiconductor is quite different from the ideal unperturbed, perfectly periodic lattice that we discussed in the previous section. As we detailed below, there are many perturbations to this idealized picture that have the effect of causing collisions, also called *scattering*, to freely moving carriers. A result of this is that carriers change directions and velocity frequently in a movement that is called *Brownian motion*. On average, just as before, carriers do not get anywhere.

Brownian motion is sketched in space in Fig. 4.1a) and in energy in Fig. 4.1b). The movement of the carriers between collisions is graphed in the energy picture by horizontal (constant energy) lines. Some collisions change the carrier's energy. These events are represented in the energy diagram by a step change in energy. Other collisions only change the carrier's velocity but not its energy. These cannot be represented very clearly in the flattened energy view.

There are many scattering mechanisms. At finite temperatures, the semiconductor atoms vibrate about their equilibrium positions disrupting the otherwise perfect crystal potential. These

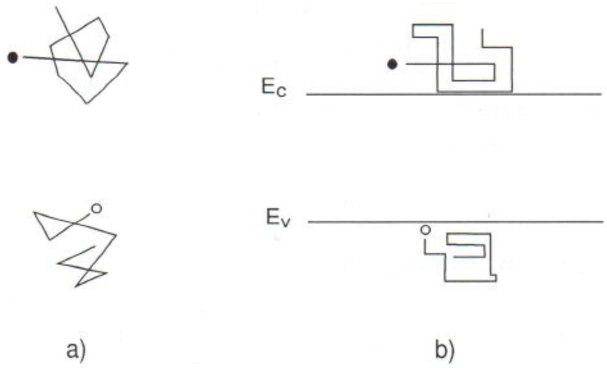


Figure 4.1: Schematic diagram of carrier thermal or brownian motion in a semiconductor: a) in real space, b) in the energy band diagram.

perturbations behave as scattering centers for the carriers. In the phonon view of lattice vibrations introduced in Ch. 1, we can think of *lattice scattering* as a mechanical collision between the carriers and massive particles called phonons, hence the name **phonon scattering**. Since the phonons have fairly small energy, the energy exchanged in these collisions is relatively small. The change in the direction of the carrier velocity can however be large. In fact, phonon scattering is an effective randomizer of carrier velocity. A key feature of phonon scattering is that it increases with temperature. As the temperature is increased, the higher energy phonon modes are proportionally more populated. Or, in simpler terms, the higher the temperature, the greater the vibrations of the lattice atoms about their equilibrium positions, and the more frequently carriers will scatter.

The introduction of an electrically charged impurity in a semiconductor lattice also upsets the perfect periodicity of the crystal potential. The Coulombic interaction of a carrier with a charged impurity constitutes an effective scattering mechanism. This is called **ionized impurity scattering**. Naturally, ionized impurity scattering is more important the higher the impurity concentration. Due to the Coulombic nature of the interaction, ionized impurity scattering does not change the particle's energy but it affects its velocity.

Although phonon scattering and ionized impurity scattering are the most important scattering mechanisms, there are others to be aware of. Any crystalline imperfection that results in a disruption of the perfect periodic potential becomes a potential source of collision events. Such are, for instance, interstitial atoms and lattice vacancies. Even neutral foreign atoms in a substitutional position impose slight perturbations to the crystalline periodic potential as a result of their electronic arrangement around the nucleus which is always different from that of the host atom. This is called **neutral impurity scattering**.

Carriers can also collide with other carriers in a process known as **carrier-carrier scattering**. This is particularly important when carrier concentrations are high. In a collision event between two identical carriers, such as an electron with another electron, the carriers can exchange energy and momentum. Since this process does not impact the average behavior of the ensemble of carriers, we need not be concerned with it here. Electron-hole collisions do take place but, in most circumstances, too infrequently to concern us here.

In surface-type devices where carrier transport takes place very close to the semiconductor surface such as in the case of the MOSFET, **surface roughness scattering** also plays a key role. A surface represents a dramatic perturbation of the crystal periodicity of a semiconductor lattice. As a result it is an effective source of scattering events for carriers in its vicinity. We will deal with it when we treat transport in an inversion layer.

The main effect of all these scattering events is to change the direction of the carrier momentum and not so much its energy. Intuitively, carriers are simply deflected but keep their energy intact. Only in phonon scattering events does a carrier change its energy by a small amount, the phonon energy. A carrier can acquire energy by absorbing a phonon or lose energy by emitting a phonon. It is through the energy exchanges associated with phonon scattering that carriers achieve thermal equilibrium with the lattice.

Scattering in a semiconductor is characterized by the *scattering time*, τ_c. This is defined as the average time between collisions. In general, the scattering time depends on the semiconductor itself, the scattering mechanism in question, and temperature. Models for scattering time associated with the various scattering mechanisms have been developed from first principles. Understanding and working with these models is beyond our purpose in this book. Fortunately, a first-order understanding of the physics of scattering can bring us very far in our quest for designing and modeling microelectronic devices. We follow this approach here.

We can get quite far by realizing that, in general, the scattering time depends on the carrier energy. Scattering becomes more likely the more states a carrier can scatter into. Therefore, the frequency of scattering should increase with the density of states. In this simplistic model, the scattering time goes as the inverse of the density of states. For electrons, we then have:

$$\tau_{ce}(E) = \frac{M}{(E - E_c)^{1/2}} \tag{4.6}$$

where M is a constant. A similar equation applies to holes with the energy measured from the top of the valence band and increasing downwards.

We can compute the average scattering time in thermal equilibrium following an approach similar to that of the thermal velocity in Eq. 4.2. In doing this, we also obtain expressions in terms of the Gamma function. The result is:

$$\tau_{ce} = <\tau_{ce}(E)> = \frac{\int_{E_c}^{\infty} \tau_{ce}(E) n_o(E)\, dE}{\int_{E_c}^{\infty} n_o(E)\, dE} \simeq \frac{2}{\sqrt{\pi k T}} M \tag{4.7}$$

We refer to this average value, τ_{ce}, as simply the scattering time. The temperature dependence that emerges here is due to the fact that at higher temperatures, the average energy of electrons in the conduction band increases and so does the frequency of scattering events. The value of M in a specific situation depends on many factors (nature of semiconductor, doping level, temperature, etc). We will learn below how to derive τ_{ce} from actual measurements in devices.

We complete the picture of thermal motion and scattering by introducing the notion of *mean free path*, l_c. This is the average distance traveled by a carrier between collisions. For a given

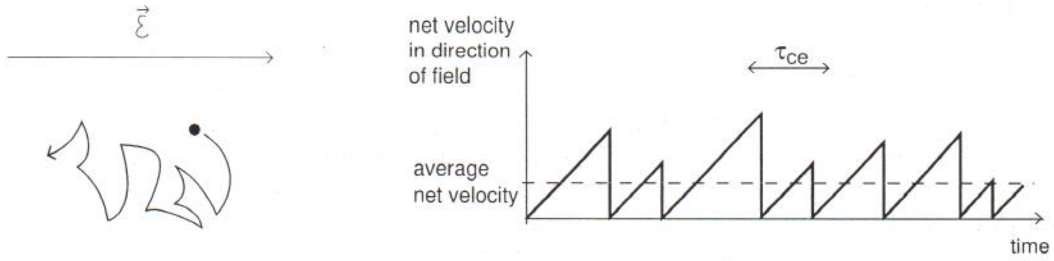


Figure 4.2: Left: schematic diagram of drift motion of an electron in an electric field. Right: time evolution of net electron velocity in the direction of the electric field.

carrier with a certain energy, the mean free path is the product of its thermal velocity and the corresponding scattering time. Working with electrons, since the thermal velocity goes as $(E - E_c)^{1/2}$ and the scattering time goes as $(E - E_c)^{-1/2}$, the mean free path ends up energy independent. Hence, the mean free path for an entire electron distribution in thermal equilibrium can simply be obtained by multiplying Eqs. 4.5 and 4.6:

$$l_{ce} = \sqrt{\frac{2}{m^*_{ce}}}M = \frac{\pi}{4}v_{the}\tau_{ce} \tag{4.8}$$

In this last expression, we have substituted M from Eq. 4.7 and used v_{the} in Eq. 4.3. A similar equation applies for holes.

As we will learn to appreciate, the mean-free path is an important length scale in semiconductors. Traditionally, device dimensions have been much larger than the mean-free path and carriers suffer many collisions as they travel through devices. It is under this assumption that the transport formulation introduced in this chapter is valid. Today, advanced MOSFETs are getting so small that this no longer is a good assumption. A new "ballistic" transport formulation is required in this instance.

4.2 Drift

Since electrons and holes are charged particles, the application of an electric field onto a semiconductor affects them. As a carrier flies in between two scattering events in the presence of an electric field, it becomes subject to a Coulombic force. As a result of this, it picks up a component of velocity in the direction of the electric field. Every collision that a carrier suffers effectively randomizes its velocity but in between collisions the carrier is under the influence of the electric field. The impact of this is a relatively slow "drift" in the direction of the electric field. This is sketched for an electron in Fig. 4.2. The average value of the *net* carrier velocity that results is called the *drift velocity*. The net current that is produced is called the *drift current*. We examine both in some detail in the next two subsections.

4.2.1 Drift velocity

The drift velocity is the the average net velocity that carriers acquire in an electric field. Let us develop a simple model for electrons. The electrostatic force that an electron experiences between collisions is $-q\mathcal{E}$, where $\mathcal{E}$ is the electric field and q is the absolute value of the electron charge. This force imposes a constant acceleration in the electron given by $a = -q\mathcal{E}/m^*_{ce}$. The initial velocity in the direction of the field is, on average, zero since a collision completely randomizes the electron velocity. The maximum velocity that the electron acquires in the direction of the field during a scattering time is then:

$$v_e^{drift} = -\frac{q\mathcal{E}\tau_{ce}}{m^*_{ce}} \tag{4.9}$$

Since the velocity ramps up linearly between collisions, the average magnitude of the velocity is actually half of this. However, our calculation here was made assuming that all collisions were separated by a time τ_{ce}. The fact is that shorter times occur more often than longer times but make a proportionally smaller contribution to the average drift velocity. When one takes proper account of this, Eq. 4.9 is obtained. A similar equation can be derived for holes (with a positive sign).

The drift velocity in Eq. 4.9 is proportional to the electric field $\mathcal{E}$. The proportionality constant is called the *mobility* and is usually represented as μ, that is:

$$\mu_e = \frac{q\tau_{ce}}{m^*_{ce}} \tag{4.10}$$

A similar equation applies to holes. The mobility is defined such that it is positive for both electrons and holes. In terms of mobility, the electron and hole drift velocities can thus be expressed as:

$$v_e^{drift} = -\mu_e\mathcal{E} \tag{4.11}$$

$$v_h^{drift} = \mu_h\mathcal{E} \tag{4.12}$$

The mobility represents the "ease" by which a carrier drifts in an electric field. The mobility is directly proportional to the scattering time and inversely proportional to the effective mass. These dependencies make sense. The longer the time between scattering events, the faster it drifts between collisions. The heavier the carrier is, the higher its inertia and the less velocity it picks up between collisions.

The mobility is a very important physical parameter in semiconductor device engineering. The fact that the mobility is directly proportional to the scattering time implies that its value depends on the strength of the various scattering mechanisms. Fig. 4.3 shows the carrier mobilities in Si at room temperature as a function of doping level. For low enough doping levels, phonon scattering dominates and the mobility is independent of doping. The value of the phonon-scattering limited

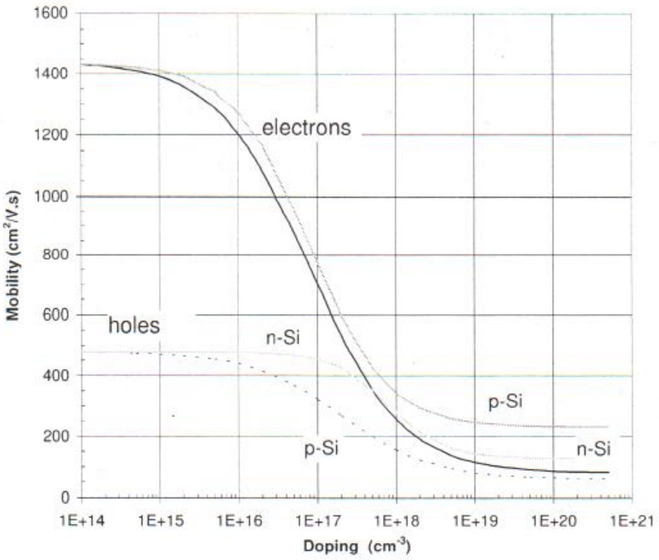


Figure 4.3: Minority and majority carrier mobility in Si at 300 K.

mobility is about a factor of three larger for electrons than for holes in Si. This ratio is about 20 for GaAs. Analytical expressions for the dependencies shown in Fig. 4.3 are given in Appendix E.

As the doping level goes up, ionized impurity scattering becomes important and the mobility decreases very quickly. Ionized impurity scattering is not a strong function of the specific dopant atom provided that it is of the same "kind". That is, electrons in Si doped with the same concentration of P, As or Sb have nearly identical mobilities. This is not the case if the type of dopant is reversed. For example, at high doping levels, the electron mobility in n-type Si is different from that of electrons in p-type Si for an identical doping level. In fact, at high doping levels and for both electrons and holes, the mobility is about a factor of two higher when the carrier is a minority carrier than when it is a majority carrier. This is because in ionized impurity scattering, *attractive scattering* is more effective than *repulsive scattering*. This arises from the peculiar shape of the potential (the "Yukawa" potential) associated with an ionized dopant. This effect is irrelevant at low doping levels where phonon scattering dominates. For low enough doping, the minority and majority carrier mobilities converge. Using the correct mobility for minority carriers is crucial to computing certain important parameters such as the base transit time in bipolar transistors.

As the temperature is reduced, phonon scattering becomes less important while ionized impurity scattering is enhanced. This implies that a reduction of temperature increases the mobility for low doping levels while it decreases it for high doping levels.

Exercise 4.1: *Estimate the electron scattering time and the mean-free path in an n-type Si sample with $N_D = 10^{17}\ cm^{-3}$ at room temperature.*

For $N_D = 10^{17}\ cm^{-3}$, Fig. 4.3 gives $\mu_e = 700\ cm^2/V \cdot s$. Solving for τ_{ce} in Eq. 4.10, we get:

$$\tau_{ce} = \frac{\mu_e m^*_{ce}}{q} = \frac{700\ cm^2/V \cdot s \times 0.28 \times 5.69 \times 10^{-16}\ eV \cdot s^2/cm^2}{1\ e} = 1.1 \times 10^{-13} s = 0.1\ ps$$

The mean-free path can be computed using Eq. 4.8:

$$l_{ce} = \frac{\pi}{4} v_{the} \tau_{ce} = 2.0 \times 10^7\ cm/s \times 1.1 \times 10^{-13} s = 1.7 \times 10^{-6}\ cm = 17\ nm$$

This is substantially smaller than typical device dimensions although the gate length of modern MOSFETs is approaching this value.

4.2.2 Velocity saturation

The linear relationship between the drift velocity and the electric field that we derived above, Eq. 4.9, is found experimentally to break down for high fields. This equation was obtained under the assumption of "near" thermal equilibrium or "quasi-equilibrium." That is, although thermal equilibrium is disturbed by the application of an electric field, the disruption is minor and the rates of the various scattering mechanisms remain unchanged from their thermal equilibrium values. This is a reasonable assumption if the carriers do not acquire too much energy from the electric field during their free flight between two collisions, or in other terms, when the drift velocity is much smaller than the thermal velocity, $v^{drift} \ll v_{th}$.

At high fields, this assumption fails. Electrons and holes can acquire substantial energy from a strong electric field. This has the effect of enhancing phonon scattering, in particular, phonon emission. For high fields, the scattering time decreases proportionally to $1/\mathcal{E}$. This leads to a saturation of the average drift velocity, as sketched in Fig. 4.4. The asymptotic value of velocity is called *saturated drift velocity* or simply saturation velocity and is represented as v_{sat}. By definition, v_{sat} is always a positive quantity. For a given semiconductor, v_{sat} only depends on temperature (weakly) and is independent of the doping level. In Si at room temperature, v_{sat} for electrons is $1 \times 10^7\ cm/s$, while for holes it is $6 \times 10^6\ cm/s$.

The overall behavior of the drift velocity with electric field is relatively well represented by an equation of the form:

$$v^{drift} = \mp \frac{\mu \mathcal{E}}{1 + |\frac{\mu \mathcal{E}}{v_{sat}}|} \tag{4.13}$$

where the minus sign applies to electrons and the plus sign to holes. Equation 4.13 has the correct limits. For low fields it reduces to $v^{drift} = \mp \mu \mathcal{E}$. For high fields, it saturates to v_{sat}.

An estimate of the magnitude of the electric field, $\mathcal{E}_c$ beyond which saturation effects are significant can be obtained from Eq. 4.13 by equating both terms in the denominator:

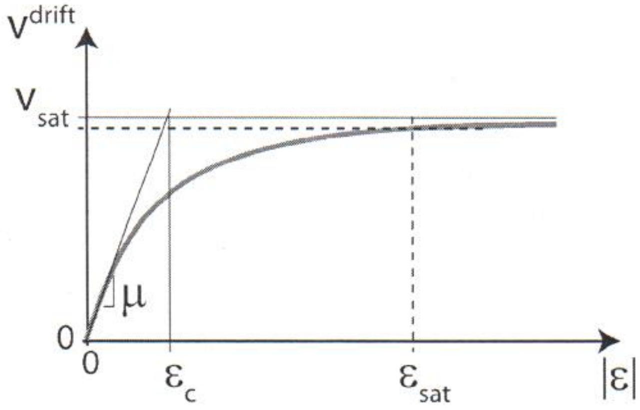


Figure 4.4: Sketch of velocity-field characteristics showing velocity saturation at high electric fields.

$$\mathcal{E}_c = \frac{v_{sat}}{\mu} \tag{4.14}$$

Reaching fields of the order of $\mathcal{E}_{sat}$ is not difficult in modern devices. For example, in a region of a Si device with a mobility of 500 $cm^2/V \cdot s$, $\mathcal{E}_{sat}$ is 2×10^4 V/cm. It only takes a voltage of 2 V applied across a length of 1 μm to reach this electric field (if it is uniformly distributed in space). The occurrence of velocity saturation, particularly for electrons, is very common in today's microelectronic devices.

For carriers to approach $v^{drift} \simeq v_{sat}$, the field has to be much higher than $\mathcal{E}_c$. If we denote as $\mathcal{E}_{sat}$ the field required to attain 90% of v_{sat}, for example, then we have that $\mathcal{E}_{sat} = 9\mathcal{E}_c$.

The saturation field depends on mobility and is therefore a function of doping level. This is clearly seen in Fig. 4.5 which shows the evolution of electron drift velocity in Si at room temperature for different doping levels. As the doping level increases, the mobility decreases and as a result, the magnitude of the field that is required to reach velocity saturation increases too. In consequence, v_{sat} effects are less likely to show up the higher the doping level. Holes display a similar behavior. Advanced Topic AT4.2 discusses in more detail velocity saturation and other "hot"-carrier effects.

4.2.3 Drift current

Carrier drift in response to an electric field produces an electric current. Once the drift velocity is known, it is easy to derive an expression for the drift current. Before doing that, let us make two important definitions that apply to any situation in which there is a net flow of particles from one point to another. *Particle flux* is defined as the number of particles that cross a surface of unit area placed normal to the particle flow every unit time. In this book, flux is represented by F and has units of $cm^{-2} \cdot s^{-1}$. The flux is obviously a vector pointing in the direction of particle flow.

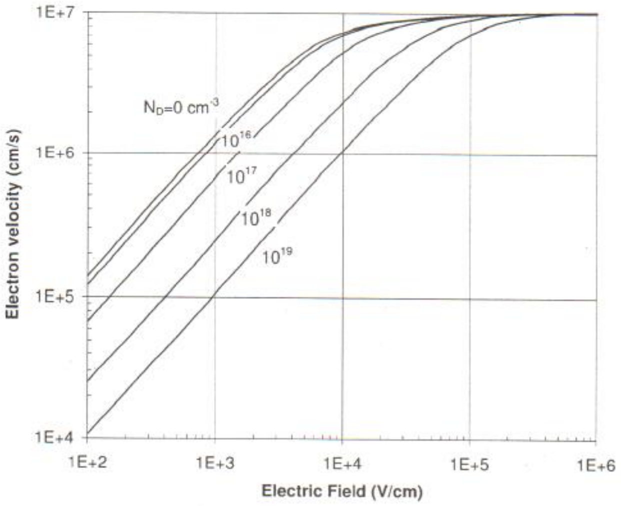


Figure 4.5: Velocity-field characteristics of electrons in n-type Si at 300 K for several doping levels.

In semiconductors, we are particularly interested in charged particles or carriers. Carrier flux hence results in an electric current. The *current density*, J, is defined as the amount of charge that crosses a unit area placed normal to the carrier flow every unit time. The current density is also a vector and its units are $C \cdot cm^{-2} \cdot s^{-1} = A/cm^2$. There is obviously a direct relationship between flux and current density. For electrons and holes, respectively:

$$J_e = -qF_e \tag{4.15}$$

$$J_h = qF_h \tag{4.16}$$

The carrier flux and the current density are directly related to the net velocity of the carriers. Consider, for example, a portion of a semiconductor with n electrons per unit volume where the electrons are moving with an average net velocity v_e. Let us place our fictitious surface of unity area normal to the direction of carrier movement, as sketched in Fig. 4.6. In a time interval dt, there are $nv_e dt$ electrons crossing through this surface. The electron *flux* is then:

$$F_e = nv_e \tag{4.17}$$

and the electron current density is therefore

$$J_e = -qnv_e \tag{4.18}$$

For holes with a concentration p moving at an average net velocity v_h, we analogously get:

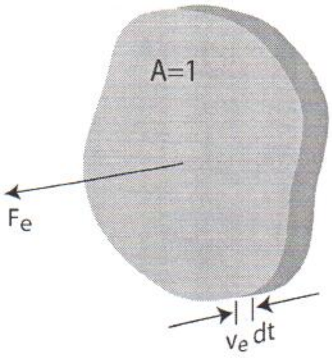


Figure 4.6: Sketch of unity area surface placed normal to electron flow. In a time interval dt, all particles inside the enclosed volume cross the light-shaded surface.

$$F_h = p v_h \tag{4.19}$$

and

$$J_h = q p v_h \tag{4.20}$$

Eqs. 4.18 and 4.20 are rather fundamental expressions. In their derivation no assumptions have been made as to the detailed physical origin of the force that acts on the carriers. These equations state the simple fact that if an ensemble of carriers moves at a certain average velocity, by virtue of the fact that they are charged, an electric current of a certain magnitude is produced. This holds regardless of the physics of the force driving the particles.

We can now go back to the process of drift in which carriers move in response to an electric field. For low electric fields, we can substitute Eqs. 4.11 and 4.12 in 4.18 and 4.20 to get:

$$J_e = q\mu_e n\mathcal{E} \tag{4.21}$$

$$J_h = q\mu_h p\mathcal{E} \tag{4.22}$$

These are the expressions for the drift current densities for electrons and holes under low-field conditions. Since in a semiconductor there are always electrons and holes, the total drift current density is:

$$J_t = q(\mu_e n + \mu_h p)\mathcal{E} \tag{4.23}$$

This is simply *Ohm's law* for a semiconductor. It is important to remember that this equation applies only if the fields are not too high. In Ohm's law, the proportionality constant is called the *conductivity.* The units of conductivity are $\Omega^{-1} \cdot cm^{-1}$. [1] One can define separately a conductivity due to electrons, another one due to holes and a total conductivity as follows:

$$\sigma_e = q\mu_e n \tag{4.24}$$

$$\sigma_h = q\mu_h p \tag{4.25}$$

$$\sigma = q(\mu_e n + \mu_h p) \tag{4.26}$$

The inverse of the conductivity is called the *resistivity* which is represented by the symbol ρ and has units of $\Omega \cdot cm$. Both resistivity and conductivity are frequently used in a variety of contexts in semiconductor engineering. For example, the resistivity is frequently quoted by wafer vendors as a way to specify the doping level of their products. This is because it is easy to measure ρ by contactless techniques and also because it is a very strong function of doping level. In a doped semiconductor in equilibrium, the resistivity is determined by the concentration of majority carriers which at room temperature is nearly equal to the doping level. Expressions for the resistivity for fully ionized extrinsic n-type and p-type semiconductors are respectively:

$$\rho_n \simeq \frac{1}{q\mu_e N_D} \tag{4.27}$$

$$\rho_p \simeq \frac{1}{q\mu_h N_A} \tag{4.28}$$

Fig. 4.7 shows the resistivity of Si as a function of doping level for n- and p-type Si. In both cases, the resistivity drops very quickly as the doping level increases. For an equivalent doping level, the resistivity of p-type Si is always higher than the resistivity of n-type Si. This is a consequence of the lower mobility of holes with respect to electrons that was presented in Fig. 4.3.

As the electric field increases, the linear relationships between current density and electric field, Eqs. 4.21, 4.22 and 4.23, do not hold anymore. In fact, as discussed above, for high enough fields the carrier velocity saturates. At that point, the current density reaches a maximum value. Plugging in v_{esat} and v_{hsat} respectively in Eqs. 4.18 and 4.20, we get:

$$J_{esat} = -qnv_{esat} \tag{4.29}$$

$$J_{hsat} = qpv_{hsat} \tag{4.30}$$

These equations convey the important message that the only way to increase the drift current density in a semiconductor that is under a high electric field is to increase the carrier density. Increasing the magnitude of the field does not help at all.

---

[1] Sometimes Ω^{-1} is called *Mho* or *Siemens* (with symbol S). Hence, conductivity can also be given in Mho/cm or S/cm.

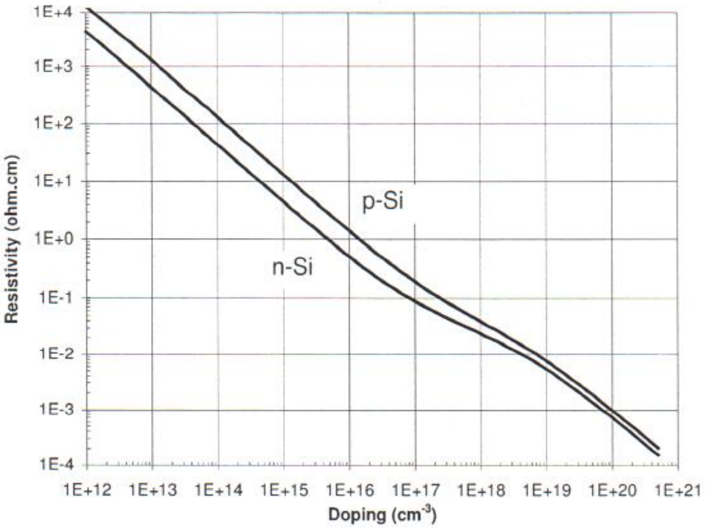


Figure 4.7: Resistivity of n- and p-type Si at 300 K.

Exercise 4.2: *Consider an n-Si sample with a resistivity of* $0.1\ \Omega \cdot cm$. *In a certain region of this sample, there is an electric field of* $1000\ V/cm$. *Estimate the electron and hole drift velocity and the total drift current density in that region.*

From Fig. 4.7, the doping level of the sample is $9\times10^{16}\ cm^{-3}$. The equilibrium hole concentration is about $1.3\times10^{3}\ cm^{-3}$. From Fig. 4.3, the mobility of electrons in this sample is about $800\ cm^2/V\cdot s$ and that of holes is $460\ cm^2/V\cdot s$. Let us assume that the field is small enough so that Eqs. 4.11 and 4.12 apply. Then:

$$v_e^{drift} = -\mu_e \mathcal{E} = -800\ cm^2/V\cdot s \times 1000\ V/cm = -8\times10^5\ cm/s$$

$$v_h^{drift} = \mu_h \mathcal{E} = 460\ cm^2/V\cdot s \times 1000\ V/cm = 4.6\times10^5\ cm/s$$

where the minus sign of v_e^{drift} indicates that electrons are flowing against the electric field. The drift velocities are much smaller than the saturation velocity. Therefore our initial assumption is validated. This region is operating in the mobility regime.

Since we already have the carrier drift velocities, the drift current densities are easy to calculate using Eqs. 4.18 and 4.20:

$$J_e = -qnv_e^{drift} = 1.6\times10^{-19}\ C\times9\times10^{16}\ cm^{-3}\times8\times10^5\ cm/s = 1.2\times10^4\ A/cm^2$$

$$J_h = qpv_h^{drift} = 1.6\times10^{-19}\ C\times1.3\times10^{3}\ cm^{-3}\times4.6\times10^5\ cm/s = 9.6\times10^{-11}\ A/cm^2$$

Clearly, since holes are minority carriers, the hole drift current density is negligible next to the electron drift current density. In consequence, the total drift current density is the value given by electron drift, about $1.2\times10^4\ A/cm^2$.

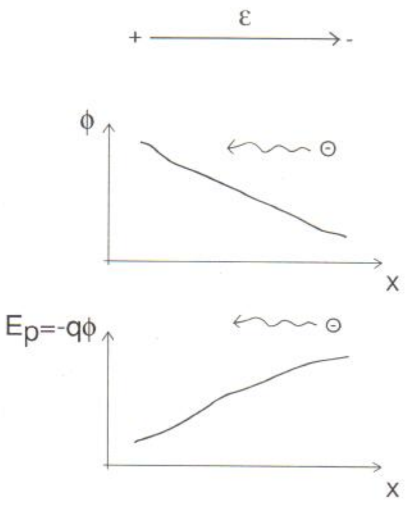


Figure 4.8: Sketch of electrostatic potential and potential energy associated with an electric field in vacuum.

4.2.4 Energy band diagram under electric field

Under the application of an electric field, the energy band diagram of a semiconductor needs to be modified to account for the *potential energy* of the electric field. Before we learn how to do this, let us review first the familiar situation in vacuum, as sketched in Fig. 4.8.

In Fig. 4.8 an electric field points from left to right in vacuum. The electrostatic potential, ϕ, associated with this field increases towards the left. The potential energy of the field, $E_p = -q\phi$, increases towards the right. As a result, an electron inside this field moves from right to left. The electron trades potential energy for kinetic energy as it accelerates towards the left while its total energy is left unchanged.

Consider now a similar situation inside a semiconductor, as in Fig. 4.9. The potential energy due to the electric field needs to be added to the band diagram. When this is done, the bands tilt as indicated in the figure. An electron inside this semiconductor drifts towards the left. In between collisions, the electron exchanges potential energy for kinetic energy as it moves along a constant energy trajectory. As an electron gains kinetic energy from the electric field, the balance between phonon emission and phonon absorption rates is broken: it becomes more likely to emit than to absorb phonons. This is precisely the path to restoration of thermal equilibrium. So, more frequently than not, the electron emits a phonon and it looses some of its kinetic energy which is given off to the lattice in the form of heat. The end result is that the electron "slides" down the conduction band lowering its energy along the way. A hole drifting inside the same semiconductor moves from left to right "bubbling up" the valence band. This intuitive view of carrier movement in the energy domain is useful when dealing with complex energy band diagrams.

When an electric field exists inside a semiconductor, the conduction and valence bands bend to reflect the change in potential energy. This is called *band bending*. The potential energy of an electric field, as we saw above, is related to the electrostatic potential through the electron charge.

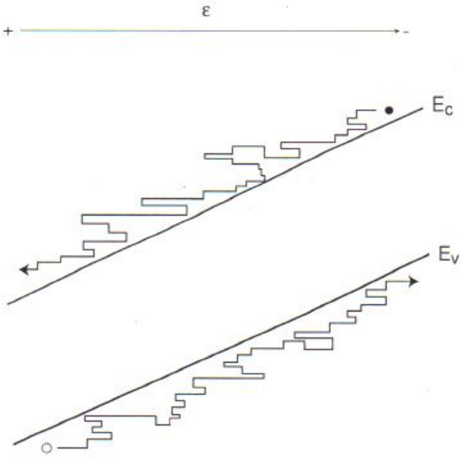


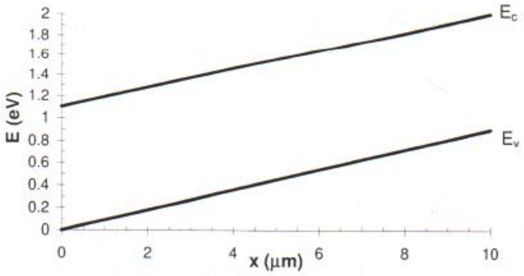
Figure 4.9: Sketch of an energy band diagram in the presence of an electric field. Electrons "slide" down the conduction band while holes "float" up the valence band.

Mathematically we can then write $E_c + E_{ref} = E_p = -q\phi$, where E_{ref} is some arbitrary reference energy. A similar equation applies to the valence band. The shape of the bands therefore reflects the shape of ϕ upside down (with the proper units). In consequence:

$$\mathcal{E} = -\frac{d\phi}{dx} = \frac{1}{q}\frac{dE_c}{dx} = \frac{1}{q}\frac{dE_v}{dx} \tag{4.31}$$

This is a useful equation that permits the computation of electric fields from band diagrams and that allows the drawing of band diagrams in regions where an electric field is present.

Exercise 4.3: *In a certain region of a semiconductor, the bands are bent as sketched in the figure below. Calculate the magnitude and direction of the electric field that exists in that region.*



Eq. 4.31 provides the answer. Note the special handling of the units:

$$\mathcal{E} = \frac{1}{q}\frac{dE_c}{dx} = \frac{1}{1\ e}\frac{0.9\ eV}{10 \times 10^{-4}\ cm} = 900\ V/cm$$

This field is positive since with the axis selected in the figure, the gradient of the bands is positive.

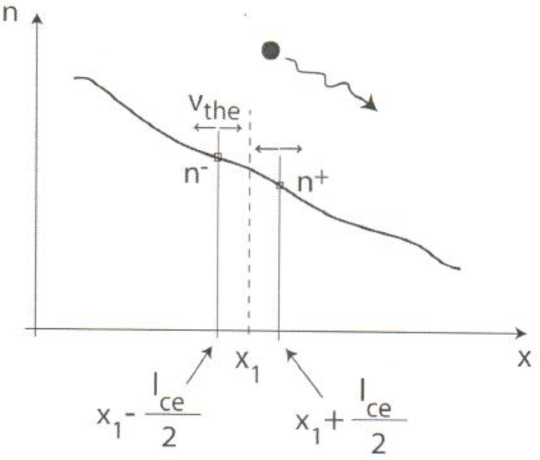


Figure 4.10: Sketch of electron diffusion down a concentration gradient.

4.3 Diffusion

Diffusion is the second physical mechanism that causes electrons and holes to flow inside semiconductors. The process of diffusion is very common in nature. It takes place whenever there is a non-uniform concentration of particles. The spread of a drop of cream in a cup of coffee, or perfume from an opened bottle in a room are familiar examples of diffusion. In diffusion, the particles in question are knocked around by the medium matter. This causes them to flow from regions where their concentration is high to regions where their concentration is low. If there is nothing trying to maintain the concentration gradient, eventually the particles will be spread evenly throughout the whole volume. In the case of a drop of cream in coffee, it is the collisions of the cream molecules with those of water that disperse the cream. For perfume, it is the air molecules that spread the perfume molecules around. In a semiconductor, carrier diffusion takes place when electrons and holes get kicked around by the vibrating semiconductor atoms. Temperature is essential in diffusion. At zero absolute temperature, there is no diffusion.

4.3.1 Fick's first law

Let us look at the process of diffusion in more detail. Consider the one-dimensional situation depicted in Fig. 4.10 where an electron gradient has somehow been established in a certain region of a semiconductor. Let us attempt to compute the electron flow due to diffusion at a certain location, such as x_1. To the left of x_1, there is a higher electron concentration than to its right. If all electrons are randomly knocked around by the lattice atoms, on average, there are more electrons that flow from the left of x_1 towards the right than the other way around. In consequence, there is a net flow of electrons from left to right down the gradient.

We can easily derive a first order expression for the electron flux at x_1. Let us assume that in the vicinity of x_1 all electrons are scattered about by the semiconductor atoms at two precise locations half a mean-free path away from x_1 on either side. Let us denote the electron concentration at $x_1 - l_{ce}/2$, and $x_1 + l_{ce}/2$, as n^- and n^+ respectively. If all n^- electrons are knocked randomly at $x_1 - l_{ce}/2$, then on average $n^-/2$ electrons get scattered towards the right, and $n^-/2$ get scattered towards the left. Since $n^- > n^+$, it is clear that at x_1 there is a net flow of electrons towards the right which is given by:

$$F_e(x_1) = \frac{n^-}{2} v_{the} - \frac{n^+}{2} v_{the} = -\frac{1}{2} v_{the}(n^+ - n^-) \tag{4.32}$$

where v_{the} is the thermal velocity for the electrons. Since the mean-free path is rather short in comparison with the dimensions of typical problems we are concerned with, it is advantageous to write the difference in electron concentrations $n^+ - n^-$ in terms of the electron gradient at x_1:

$$n^+ - n^- = \frac{dn}{dx}|_{x_1} l_{ce} \tag{4.33}$$

Plugging this result into the expression for the flux, we get:

$$F_e(x_1) = -\frac{1}{2} v_{the} l_{ce} \frac{dn}{dx}|_{x_1} \tag{4.34}$$

There are a number of important dependences in this result. First, the flow of electrons is proportional to the gradient of their concentration. It is *the difference* in carrier concentrations that drives diffusion. This is a very fundamental result of general applicability. It applies to electrons and holes in semiconductors as well as to molecules of cream in coffee and to perfume in air. The minus sign that is obtained in Eq. 4.34 makes good physical sense. As Fig. 4.10 shows, electrons diffuse down the gradient.

The second important dependence of Eq. 4.34 is that the diffusion flux is proportional to the thermal velocity. This makes sense. The faster carriers travel between scattering events, the higher the net flux at a given location.

The third dependence involves the mean free path: the longer this is, the higher the diffusion flux becomes. As Eq. 4.33 indicates, what drives diffusion is the difference of carrier concentration in the scale of the mean free path. Given a certain gradient of electron concentration, the longer the mean free path, the bigger the difference between left-going and right-going fluxes becomes.

The combined prefactor multiplying the gradient of electron concentration in Eq. 4.34 is called the *diffusion coefficient*, with symbol D, and units of cm^2/s:

$$D_e = \frac{1}{2} v_{the} l_{ce} \tag{4.35}$$

In terms of the diffusion coefficient, we can write the electron flux as:

$$F_e = -D_e \frac{dn}{dx} \tag{4.36}$$

This is called Fick's first law which is derived more rigorously in many solid-state books. Similarly, for holes, we have:

$$F_h = -D_h \frac{dp}{dx} \tag{4.37}$$

The diffusion coefficient gives a sense of the "ease" by which electrons diffuse in a semiconductor. For a given concentration gradient, the higher D is the higher the diffusion flux. The value of the diffusion coefficient embodies the strength of the scattering mechanisms that carriers suffer as well as the nature of the carriers. It also depends on temperature through the temperature dependence of the thermal velocity and that of the scattering time.

4.3.2 The Einstein relation

Since the diffusion coefficient is intimately tied to scattering, it is reasonable to expect that it bears some relationship to the mobility which itself reflects carrier scattering in the presence of an electric field. This relationship is in fact very fundamental and is known as the **Einstein relation**. Using Eqs. 4.34, 4.3, 4.8, 4.10, and 4.35 we can easily find:

$$\frac{D_e}{\mu_e} = \frac{kT}{q} \tag{4.38}$$

The Einstein relation for holes can be derived in a similar way:

$$\frac{D_h}{\mu_h} = \frac{kT}{q} \tag{4.39}$$

The ratio kT/q is referred to as the *thermal voltage*. This is an important "tickmark" in the voltage scale that is pervasive in the analysis of semiconductor devices. [2]

The Einstein relation is quite general and is in fact valid for all systems that obey Maxwell-Boltzmann statistics. Interestingly, the Einstein relation does not depend on doping, only on temperature. The Einstein relation can also be derived for a degenerate carrier gas. In this case, a doping dependence arises (see Problem 4.9).

The Einstein relation provides a connection between mobility and diffusion coefficient in thermal equilibrium. Outside equilibrium, the Einstein relation is expected to hold if the disturbance from equilibrium is not so high that the scattering rates are modified. A good working criterion

---

[2] In some books and articles, the thermal voltage is denoted as V_t, or V_{th}. This notation can easily lead to confusion with the threshold voltage of a MOSFET and it will not be followed in this book.

is that the electric fields must be low enough so that there is a linear proportionality between drift current and electric field.

Exercise 4.4: *Estimate the electron and hole diffusion coefficients in* $0.1\ \Omega \cdot cm$ *p-Si at room temperature.*

The first step is to find the doping level of this sample. From Fig. 4.7 we find that p-type Si with a resistivity of $0.1\ \Omega \cdot cm$ has a doping level of about $N_A = 3 \times 10^{17}\ cm^{-3}$.

The second step is to obtain the corresponding mobilities for electrons and holes. Holes are majority carriers in p-Si. Electrons are minority carriers. Care has to be exercised to select the proper curve in Fig. 4.3. For holes we find that for $N_A = 3 \times 10^{17}\ cm^{-3}$, the mobility is $\mu_h \simeq 240\ cm^2/V \cdot s$. For electrons, we find $\mu_e \simeq 530\ cm^2/V \cdot s$.

The diffusion coefficients are obtained using Einstein's relation. For holes:

$$D_h = \frac{kT}{q}\mu_h = \frac{0.026\ eV}{1\ e}\ 240\ cm^2/V \cdot s = 6.2\ cm^2/s$$

Similarly, for electrons we find $D_e = 13.8\ cm^2/s$.

4.3.3 Diffusion current

Since electrons and holes are charged particles, their diffusion produces an electric current. Multiplying the expressions for carrier flux by the respective carrier charges, the current densities due to diffusion are obtained:

$$J_e = qD_e\frac{dn}{dx} \tag{4.40}$$

$$J_h = -qD_h\frac{dp}{dx} \tag{4.41}$$

Rather than trying to remember the signs of Eqs. 4.40 and 4.41, it is much easier to derive them in an intuitive way. Consider the sketches of Fig. 4.11. In the presence of a positive gradient of electron concentration, such as the one pictured on the left of Fig. 4.11, electrons flow down the gradient towards the left. The electron flux is then negative, as Eq. 4.36 indicates. Since the electrons are negatively charged, the electron current will flow from left to right, in the same direction as the axis and it is then positive, in agreement with Eq. 4.40. In the presence of a similar gradient, the hole flux is also negative but the hole current is negative too since the hole charge is positive (right sketch in Fig. 4.11).

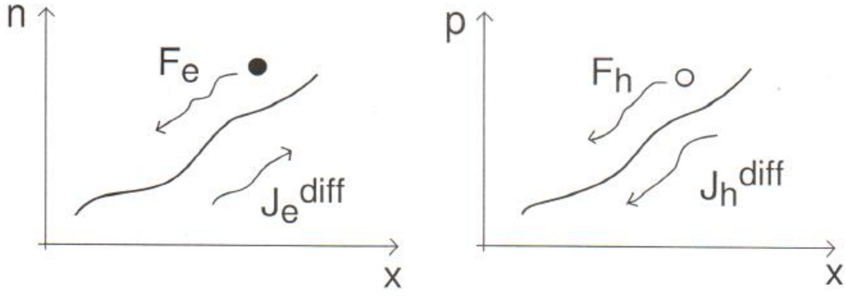


Figure 4.11: Sketch of relationship between fluxes and diffusion current densities for electrons and holes.

Exercise 4.5: *In a certain region of a semiconductor characterized by* $0 \leq x\ (\mu m) \leq 1$, *there is an electron concentration that follows a profile* $n(x) = 10^{17}(1 + 10x)\ cm^{-3}$ *with* x *in* μm. *At* $x = 0$, *calculate the electron diffusion flux, the electron diffusion current density, and the electron diffusion velocity. Assume an electron diffusion coefficient of* $10\ cm^2/s$.

Before doing any calculations, it is convenient to homogenize the units in the equation that describes the electron profile. In this equation, x needs to be given in μm. If instead we want to use x in cm, then we need to change x into $10^4 x$. Hence the electron profile can be described as $n(x) = 10^{17}(1 + 10^5 x)\ cm^{-3}$ with x in cm.

At $x = 0$, the electron diffusion flux is given by Eq. 4.36:

$$F_e|_{x=0} = -D_e \frac{dn}{dx}|_{x=0} = -10 \times 10^{17} \times 10^5 = -10^{23}\ cm^{-2}s^{-1}$$

The minus sign indicates that the electron flux takes place against the axis x.

The electron current density is simply obtained by multiplying the flux by the electron charge (Eq. 4.40):

$$J_e = -qF_e = 1.6 \times 10^{-19} \times 10^{23} = 1.6 \times 10^4\ A/cm^2$$

The electron velocity can be obtained from the fundamental relationship that relates carrier flux and carrier velocity, Eq. 4.17:

$$v_e^{diff} = \frac{F_e}{n} = -\frac{10^{23}}{10^{17}} = -10^6\ cm/s$$

The minus sign indicates that the electrons are traveling in a direction contrary to x.

4.4 Transit time (briefly)

As discussed above, if in a semiconductor region there is a concentration gradient or an electric field, there is net movement of carriers in space. It is often of interest to compute the time that it takes for a carrier, on average, to get from one point in space to another. This is called the *transit time*. Since the carrier velocity in general depends on location, the most general way of calculating the transit time in one dimension is by integrating the differential of time required

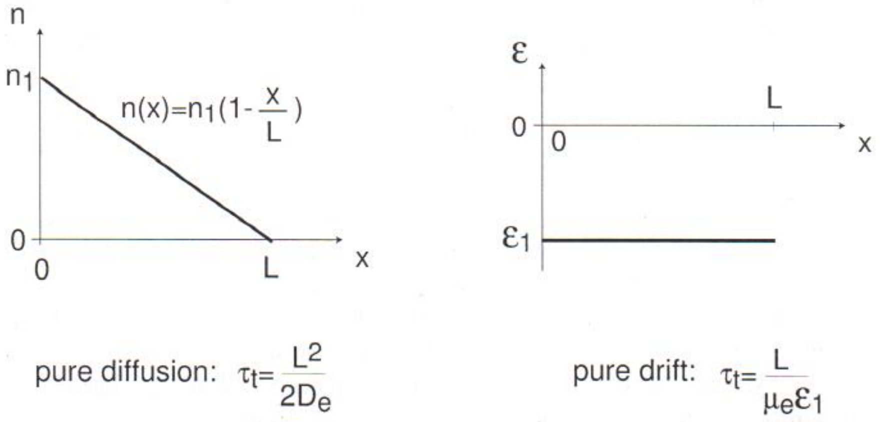


Figure 4.12: Two simple examples of electron flow and the resulting transit times from $x = 0$ to $x = L$: (a) diffusion down a linear gradient, (b) drift in a uniform electric field. In both cases, electrons flow from the left towards the right.

to travel through a differential of space. In integral form, then, the transit time from a point at $x = 0$ to a point $x = L$ is given by :

$$\tau_t = \int_0^{\tau_t} dt = \int_0^L \frac{dx}{v(x)} \tag{4.42}$$

The net carrier velocity $v(x)$ might be due to drift, diffusion or a combination of the two. It is useful to separately examine these two cases.

For a purely diffusive case, the combination of Eqs. 4.18 and 4.40 gives an expression for the electron diffusion velocity:

$$v_e^{diff} = -D_e \frac{1}{n}\frac{dn}{dx} \tag{4.43}$$

Inserting this in Eq. 4.42 yields:

$$\tau_t = -\frac{1}{D_e}\int_0^L \frac{n}{\frac{dn}{dx}} dx \tag{4.44}$$

This is the most general expression of the transit time due to electron diffusion. It depends solely on the electron concentration profile and the electron diffusion coefficient. A simple case helps to see the leading dependences.

Consider a linear electron concentration profile in space such as the one depicted in Fig. 4.12a). At $x = 0$ the electron concentration is n_1, and at $x = L$ the electron concentration goes to zero. This is not as unphysical a situation as it might seem. The boundary condition at $x = L$

ressembles the base edge of the base-collector junction of a bipolar transistor in the forward active regime. An analytical expression for this electron profile is $n(x) = n_1(1-\frac{x}{L})$. Plugging this profile into Eq. 4.44 yields after some simple algebra:

$$\tau_t = \frac{L^2}{2D_e} \tag{4.45}$$

This is an interesting result. The transit time due to electron diffusion goes as the square of the length of the region through which the carriers diffuse and is inversely proportional to the diffusion coefficient. The inverse proportionality on diffusion coefficient is readily expected. The quadratic dependence on L also makes good sense. For a given value of n_1, if L increases, the electron gradient decreases proportionately and the distance that the carriers have to travel increases by the same amount. Hence, a $\tau_t \sim L^2$ dependence results.

Consider now a case of pure electron drift. Eq. 4.11 gives an expression for the electron drift velocity in terms of the electric field for small fields. Plugging this into Eq. 4.42 yields the transit time due to electron drift:

$$\tau_t = -\frac{1}{\mu_e}\int_0^L \frac{dx}{\mathcal{E}(x)} \tag{4.46}$$

This depends on electron mobility and the field distribution in space.

Let us examine the simple example depicted in Fig. 4.12b) in which a uniform electric field of magnitude $\mathcal{E}_1$ is set up in a certain region in space. The electric field is negative in sign so that electrons drift from left to right. For this case, since the electric field is uniform, $\mathcal{E}_1$ can be taken out of the integral in 4.46 to yield:

$$\tau_t = \frac{L}{\mu_e \mathcal{E}_1} \tag{4.47}$$

The transit time due to drift is linear in L and inversely proportional to $\mathcal{E}_1$, as expected from physical arguments.

Identical results can be obtained for the transit time for holes under similar circumstances. In situations with more complex carrier concentration or electric field profiles, the transit time expression might well be more complicated and perhaps take a non-analytical form. The physics, however, is unchanged.

Situations that combine drift and diffusion are also straightforward to handle. In the case of electrons, for example, the total current is the sum of a drift contribution plus a diffusion contribution:

$$J_e = J_e^{drift} + J_e^{diff} \tag{4.48}$$

The electron velocity in this case can be obtained using Eq. 4.18:

$$v_e = -\mu_e \mathcal{E} - D_e \frac{1}{n}\frac{dn}{dx} = v_e^{drift} + v_e^{diff} \tag{4.49}$$

In order to compute the transit time in a situation like this, the sum of the carrier drift and diffusion velocities must be used in Eq. 4.42.

4.5 Non-uniformly doped semiconductor in thermal equilibrium

In thermal equilibrium, there are situations in which it is possible for an electric field to exist inside a semiconductor without applying a voltage from the outside. This is the case, for example, when the doping level inside a semiconductor region changes in space. This is quite common in devices since most doping techniques result in non-uniform impurity profiles. We study these situations here because even in thermal equilibrium, they bring into play both drift and diffusion simultaneously. Our analysis is also the basis for a future treatment of carrier behavior outside equilibrium in non-uniformly doped semiconductors.

This section introduces a number of important new concepts. The starting point is a review of one of Maxwell's equations that plays a key role in semiconductor devices: Gauss' law.

4.5.1 Gauss' law

Gauss' law states that the divergence of the electric field at a certain point is equal to the volume charge density at that same point divided by the permittivity of the material. In one dimension, Gauss' law can be written as:

$$\frac{d\mathcal{E}}{dx} = \frac{\rho}{\epsilon} \tag{4.50}$$

where ρ is the volume charge density [3] and ϵ is the permitivity of the semiconductor.

It is important to understand Gauss' law well. If the volume charge density at a certain point is zero, it does not follow that the electric field is zero too. For example, in between the plates of a capacitor with air as dielectric and a voltage applied, the volume charge is zero but the electric field is not. The charge at the plates of the capacitor generates an electric field. If ρ is zero inside the capacitor, then $d\mathcal{E}/dx$ is zero at every point and $\mathcal{E}$ does not change from point to point. In order to have an electric field at a certain point, there is no need to have charge at that same location.

In semiconductors there are several sources of volume charge. There is the *mobile charge* that arises from electrons and holes. There is also *fixed charge* that results from the ionized dopants

---

[3] It is unfortunate that in the semiconductor literature, the symbol ρ is used for both resistivity and volume charge density. Since this is common practice, this notation will be followed in this book. The context and the units help identify which variable the symbol refers to in any particular equation.

- ionized donors are positively charged and ionized acceptors are negatively charged. A general expression for the net volume charge density in a semiconductor is then:

$$\rho = q(p - n + N_D^+ - N_A^-) \tag{4.51}$$

Throughout this book we will assume that the working temperature is high enough for all dopants to be ionized. Thus $N_D^+ \simeq N_D$ and $N_A^- \simeq N_A$. Combining this with equations 4.50 and 4.51, we get:

$$\frac{d\mathcal{E}}{dx} = \frac{q}{\epsilon}(p - n + N_D - N_A) \tag{4.52}$$

Under thermal equilibrium conditions, in the bulk of a uniformly-doped semiconductor sufficiently far away from any surface, *charge neutrality* prevails. This is because every dopant releases a carrier of the contrary charge sign. In a charge-neutral situation, Eq. 4.52 states that the electric field cannot change in space. Furthermore, since in thermal equilibrium no electric fields are applied from the outside, we can conclude that $\mathcal{E}_o = 0$ everywhere (the subindex o is used to denote thermal equilibrium).

The situation is quite different in the presence of a non-uniformly doped semiconductor in thermal equilibrium. To illustrate the issues involved in a simple way, let us consider a long bar with a gradient of donors along its length as sketched in Fig. 4.13. If we can assume that nothing changes in the other two dimensions, a one dimensional treatment should be adequate. Let us also assume that the semiconductor is sufficiently extrinsic so that we need not account for the tiny charge contribution of the minority holes. Under these conditions, Eq. 4.52 can be simplified to:

in thermal equilibrium.

$$\frac{d\mathcal{E}_o}{dx} = \frac{q}{\epsilon}(N_D - n_o) \tag{4.53}$$

where we have made explicit the thermal equilibrium situation by adding the subindex o to $\mathcal{E}$ and n.

At first sight, it may seem reasonable to expect that the electron concentration replicates exactly the donor concentration since each donor releases one electron, that is, $n_o(x) = N_D(x)$. This assures precise charge neutrality at every point of space. If that were the case, however, the electron distribution would end up being non uniform. The electron concentration gradient would then produce a diffusion of electrons from the right where the initial concentration is high to the left where it is smaller and current would accordingly flow. This is clearly not an equilibrium situation.

We may be tempted to conclude then that as a result of carrier diffusion, electrons will flow until the electron concentration is uniform everywhere. If that were the case, on the right side of the bar there would be a certain amount of positive charge resulting from the unbalanced concentrations of ionized donors and electrons. On the left, there would be net negative charge since the electron concentration is larger there than the donor concentration. Following Eq. 4.53,

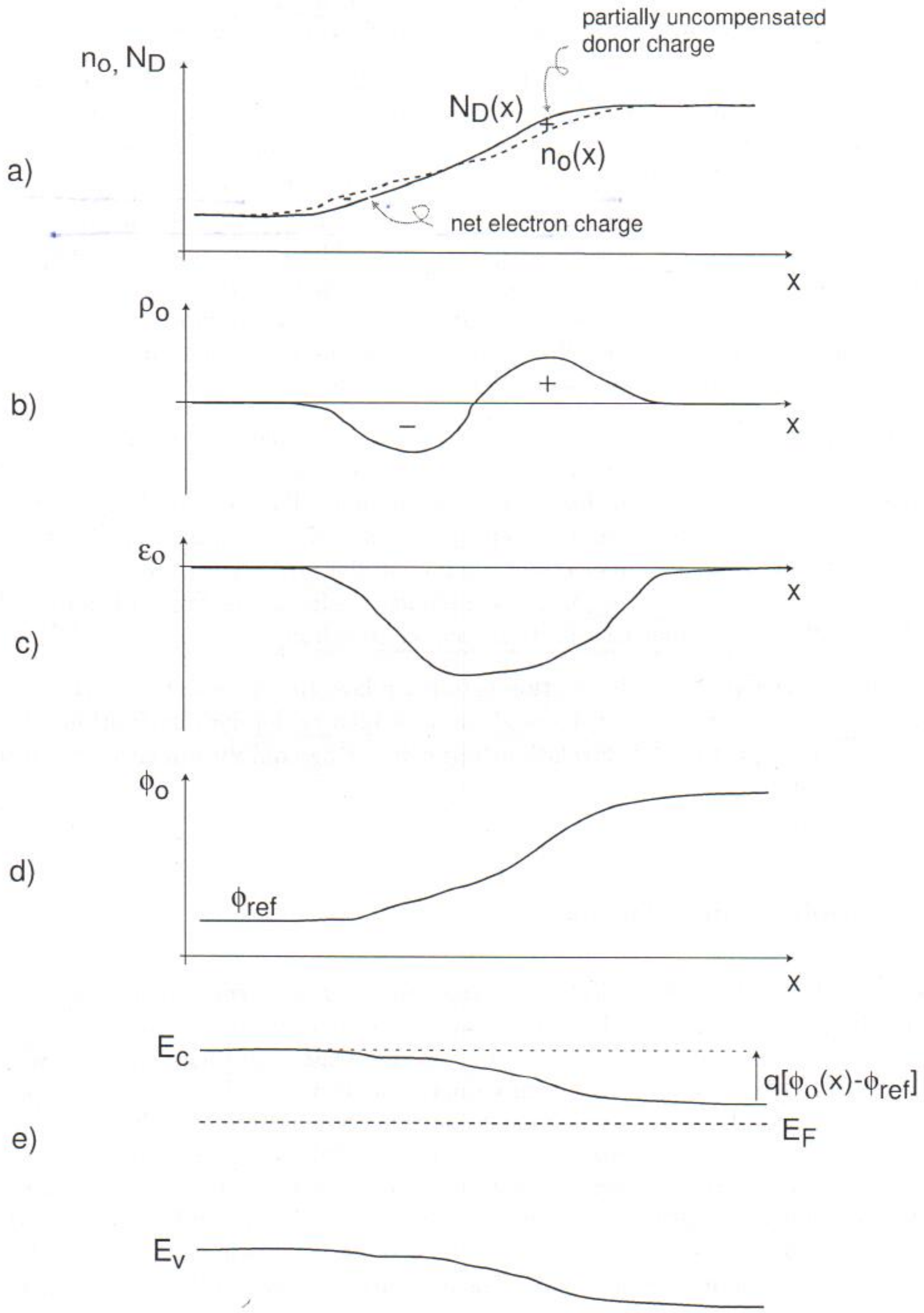


Figure 4.13: Non-uniformly doped semiconductor bar in thermal equilibrium: (a) doping level and resulting carrier concentration, (b) volume charge density, (c) electric field, (d) electrostatic potential, and (e) energy band diagram.

this charge distribution would result in an electric field that would pull electrons back to the right and away from the left region. This again is not an equilibrium situation.

The two extreme electron distributions that we have discussed are not possible in equilibrium. When $n_o(x) = N_D(x)$, charge neutrality at every point results in an electron diffusion current. If n_o is uniform in space, the presence of net volume charge results in a drift current. The one thing that we know must be fulfilled in thermal equilibrium is that the total electron current be zero everywhere. This implies that electron drift and diffusion must precisely balance out at every point. For a drift current to flow, we need an electric field. This can be generated if a non-uniform electron distribution is established that is somehow different from the donor concentration, *i.e.* $n_o(x) \neq N_D(x)$. The electron gradient associated with the non-uniform electron distribution results in a diffusion current. The equilibrium carrier profile is the one for which the drift and diffusion currents cancel out at every point. This is sketched in Fig. 4.13.

The mismatch between electron concentration and donor concentration in the non-uniformly-doped bar leads to a dipole of charge, as sketched in Fig. 4.13. There is net positive charge towards the right and net negative charge towards the left. This dipole of charge produces an electric field that points from right to left. With our choice of axis, $\mathcal{E}_o$ is negative. An equivalent way of expressing this is through an electrostatic potential ϕ_o that increases from left to right, as also sketched. The energy band diagram, in consequence, shows substantial bending. However, the Fermi level is flat throughout to reflect thermal equilibrium.

Our goal for the remainder of this section is to learn how to calculate the carrier profiles, and to solve the complete electrostatic problem given an arbitrary doping distribution. Before that, we introduce a very important set of relationships connecting equilibrium carrier concentrations and electrostatic potential.

4.5.2 The Boltzmann relations

The application of the principle of detailed balance to a non-uniformly doped semiconductor in thermal equilibrium demands that the electron and hole currents be separately zero everywhere. This may appear obvious but it is instructive to consider what would happen if this were not the case (see Fig. 4.14). Assume a non-uniformly doped bar in which the electron current happens to balance out precisely the hole current so as to have net zero current. In this instance since the two carrier types have charges of the contrary sign, electrons and holes must flow in the same direction from one side of the semiconductor to the other. In order to sustain this, electron-hole pairs must continuously be generated at one side of the bar and recombine at the other. This would imply that one side of the semiconductor is cooling down while the other is heating up. Clearly this is not an equilibrium situation. The conclusion is that in thermal equilibrium, both carrier currents must be zero separately.

Coming back to the non-uniformly doped bar in thermal equilibrium depicted in Fig. 4.13, we just concluded that the total electron current must be zero. If the electric field resulting from the non-uniform dopant distribution is not too high, Eq. 4.21 can be used for the drift component of the current density. The total electron current density can then be written as:

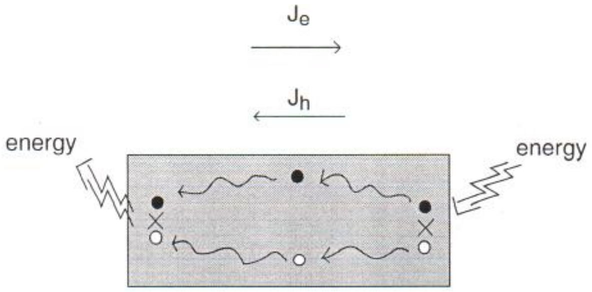


Figure 4.14: Sketch of impossible thermal equilibrium situation with $J_e = -J_h \neq 0$. The net carrier currents result in spontaneous absorption of energy at one end of the bar and emission of energy at the other. This cannot happen in thermal equilibrium.

$$J_e = q\mu_e n_o \mathcal{E}_o + qD_e \frac{dn_o}{dx} = 0 \tag{4.54}$$

where the electron concentration has been explicitly denoted as n_o to indicate that this is an equilibrium situation. In general, n_o depends on position.

Eq. 4.54 suggests that in thermal equilibrium there is a direct connection between the equilibrium carrier concentration and the electric field. The electric field drives drift, the concentration gradient drives diffusion and they both must be in perfect balance in thermal equilibrium. Hence n_o and $\mathcal{E}_o$ should be in some kind of relationship.

If we solve for $\mathcal{E}_o$ in Eq. 4.54, we get:

$$\mathcal{E}_o = -\frac{kT}{q}\frac{1}{n_o}\frac{dn_o}{dx} \tag{4.55}$$

where we have used the Einstein relation (Eq. 4.38).

If we apply an equivalent procedure to the minority carrier, in this case holes, we arrive at a similar equation:

$$\mathcal{E}_o = \frac{kT}{q}\frac{1}{p_o}\frac{dp_o}{dx} \tag{4.56}$$

This equation can also be derived directly from 4.55 by invoking the fact that $n_o p_o = n_i^2$.

Eqs. 4.55 and 4.56 are useful expressions since they allow the calculation of the electric field in a certain region of a semiconductor provided that either n_o or p_o are known. Equations 4.55 and 4.56 can also be used to derive a relationship between the electrostatic potential in thermal equilibrium ϕ_o and the equilibrium carrier concentration. From Eq. 4.55, for example, one gets the differential equation:

$$\frac{d\phi_o}{dx} = \frac{kT}{q}\frac{d(\ln n_o)}{dx} \tag{4.57}$$

which after integration results in:

$$\phi_o(x) - \phi_{ref} = \frac{kT}{q}\ln\frac{n_o(x)}{n_o(\text{ref})} \tag{4.58}$$

Had we started from Eq. 4.56 and proceeded in a parallel way, we would have reached the following result:

$$\phi_o(x) - \phi_{ref} = \frac{kT}{q}\ln\frac{p_o(\text{ref})}{p_o(x)} \tag{4.59}$$

These equations give a relationship between the *ratio* of the equilibrium carrier concentration at two different points and the *difference* in the electrostatic potential between the same two points. If the carrier concentration between two regions changes by a factor of 10, for example, the electrostatic potential differs by $(kT/q)ln10$. At room temperature this is about 60 millivolts per decade or simply mV/dec, a handy number to remember.

ϕ_{ref} is in principle not set and in consequence, $\phi_o(x)$ is not completely specified. This should not be a source of difficulty since in electrostatic problems the physics is always in the *potential difference* and not in its absolute value. In any one problem, one is free to choose the reference for potentials. A common choice consists on assigning the value of $\phi_{ref} = 0$ to the intrinsic carrier concentration $n_o(ref) = n_i$. With this choice, Eqs. 4.58 and 4.58 respectively become:

$$\left\{\begin{array}{lllr} n_o & = & n_i \exp\dfrac{q\phi_o}{kT} & (4.60) \\[2ex] p_o & = & n_i \exp\dfrac{-q\phi_o}{kT} & (4.61) \end{array}\right.$$

Note that these equation satisfy the thermal equilibrium requirement that $n_o p_o = n_i^2$.

Both Eqs. 4.60 and 4.61 give a relationship between the equilibrium carrier concentration and the electrostatic potential in a case in which we have selected $\phi_{ref} = 0$ at the intrinsic point $n_o = p_o = n_i$. The relationship is exponential. These equations are known as the Boltzmann relations. If either one of them is known, the other one is straightforward to derive.

It cannot be stressed enough the fact that these equations are reference sensitive. A different choice of potential reference results in expressions that look different. The physics of the problem is not changed, but its mathematical expression is. The resulting equations, in all cases, are known as the Boltzmann relations. A useful corollary is that clever choice of reference might lead to a simpler mathematical formulation. In this book we will make different choices in different circumstances. You have to be ready to work with a menu of equivalent expressions.

Exercise 4.6: *Derive equivalent Boltzmann relations as Eqs. 4.60 and 4.61 above for a reference choice that assigns $\phi_{ref} = 0$ to a point where the equilibrium hole concentration is N_A.*

We can do this using Eq. 4.58 directly, or deriving an equivalent one from 4.56. Through either path, one gets the same result. If we proceed through the first route, at a point where the equilibrium hole concentration is N_A, the equilibrium electron concentration is n_i^2/N_A. At this point we select $\phi_{ref} = 0$. Plugging this in Eq. 4.58, we get:

$$n_o = \frac{n_i^2}{N_A} \exp \frac{q\phi_o}{kT} \tag{4.62}$$

Using again $n_o p_o = n_i^2$, we get:

$$p_o = N_A \exp \frac{-q\phi_o}{kT} \tag{4.63}$$

As an additional algebraic exercise, show that you get the same equations starting from Eq. 4.56.

The Boltzmann relations, as described here, are nothing but expressions of a fundamental law of statistical mechanics known as Boltzmann's law. In its most generality, Boltzmann's law applies to any system of ideal particles immersed in a field of conservative forces (*i.e.* a field of forces that can be described by a potential) in thermal equilibrium. Boltzmann's law states that the probability of finding particles in a given spatial arrangement varies exponentially with the negative of the potential energy of that arrangement, divided by kT. For particle concentration, Boltzmann's law can be expressed as:

$$n \propto e^{-E_p/kT} \tag{4.64}$$

where E_p is the potential energy.

Interesting

This law applies to many other systems in nature. One particular system, the distribution of molecules in the ideal atmosphere, provides us with an intuitive and visual analogy. Imagine an ideal atmosphere that is characterized by a constant temperature (in our real atmosphere it gets colder as we go up in altitude). All gas molecules of this ideal atmosphere are subject to the gravitational force. It is then clear that the weight of the gas column extending from a given altitude to infinity decreases as we go up in the atmosphere. The atmosphere gets "thinner" the higher we go. The concentration of molecules in height in this ideal atmosphere follows Boltzmann's law, that is, it decreases exponentially.

We can exploit this analogy to visualize the electron distribution in the conduction band of a semiconductor in thermal equilibrium. If the semiconductor is non degenerate, at any location, electrons obey the Maxwell-Boltzmann distribution function, that is, the probability of finding an electron at a certain energy inside the conduction band goes down as we go up in energy inside the conduction band. This follows the law $\exp(-E/kT)$. At the same time, if the doping level is non uniform, the total electron concentration (the integral of a "vertical column" of electrons in the atmosphere analogy) also changes from point to point as $\exp(q\phi/kT) = \exp(-E_p/kT)$. Hence we can think of this non-uniform doped semiconductor in thermal equilibrium as having a "stratified" distribution of electron occupation probability in the conduction band much as the gas concentration would be in an ideal atmosphere if there is a landscape in the terrain underneath. This is pictorially illustrated in Fig. 4.15. One can think of the electrostatic potential landscape

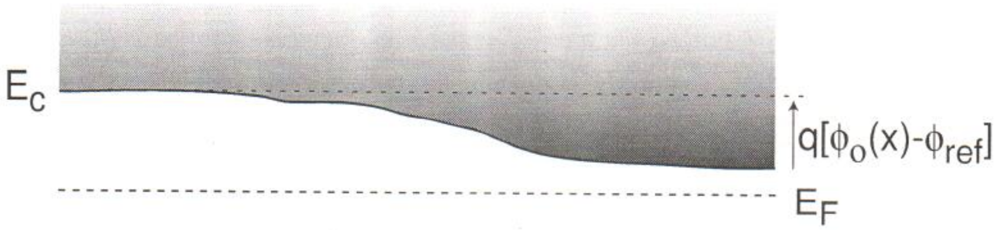


Figure 4.15: Pictorial view of electron concentration in conduction band of non-uniformly doped semiconductor in thermal equilibrium.

as "carving out" the electron probability distribution in the conduction band. The same applies for holes in the valence band, but the picture is upside down, that is, the higher in energy, the denser the hole concentration is.

A final point to be made about the Boltzmann relations is that they have been derived under the assumption of Maxwell-Boltzmann statistics which apply to non-degenerate carrier situations. In the case of carrier degeneracy, the Einstein relation takes on a different form (see Problem 4.9) and as consequence, the Boltzmann relations become somehow more complex. Still the basic fact remains that in thermal equilibrium, there is a one-to-one relationship between changes in electrostatic potential and changes in carrier concentration.

4.5.3 Equilibrium carrier concentration

We are finally in a position to formalize an approach to calculating the equilibrium carrier concentration in a non-uniformly doped semiconductor in thermal equilibrium. If we substitute Eq. 4.55 into Gauss' law 4.53, we get:

$$\frac{\epsilon kT}{q^2}\frac{d^2(\ln n_o)}{dx^2} = n_o - N_D \tag{4.65}$$

Given a certain doping distribution $N_D(x)$, the solution of this second-order non-linear differential equation provides the spatial distribution of electrons in thermal equilibrium $n_o(x)$. Everything else in this equation is known. Once $n_o(x)$ is known, it is straightforward to compute $\rho_o(x)$, $\mathcal{E}_o(x)$, $\phi_o(x)$ and the energy band diagram in space.

As it turns out, 4.65 is a fairly challenging differential equation to solve. Analytical solutions are only available for a small set of simple cases. In most situations, this equation needs to be solved through numerical techniques. We study next a particularly simple example that brings some important issues associated with solutions to this differential equation.

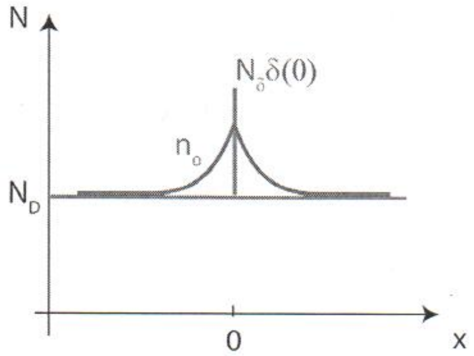


Figure 4.16: Schematic of delta doping. A sheet of n-type dopants with an areal density N_δ is inserted in an otherwise uniformly doped n-type semiconductor. All the surfaces are very far away.

Example 1: delta doping

Delta doping is illustrated in Fig. 4.16. This figures sketches an extrinsic n-type semiconductor doped with a uniform doping concentration N_D. At a certain location that we label $x = 0$, there is an additional *sheet* of n-type dopants with a concentration N_δ (per unit area). The situation depicted in Fig. 4.16 is of a 1D nature and all the surfaces are very far away. We are interested in calculating the equilibrium electron concentration n_o throughout this sample. Technologically, it is possible to fabricate these kind of delta-doped layers using modern epitaxial techniques. They are widely used in GaAs High-Electron Mobility Transistors (HEMTs).

In the absence of the delta-doped layer, charge neutrality dictates that the electron concentration everywhere be exactly equal to N_D. The delta-doped layer introduces more dopants and more electrons into the system. The dopants are positively charged and remain stuck in the lattice in a sheet-like distribution. The electrons are free to roam around. In thermal equilibrium, the balance between drift and diffusion leads to an electron distribution that is a somehow smeared version of the delta function, as indicated in Fig. 4.16. At a sufficient distance away from $x = 0$, we expect n_o to approach N_D.

The picture shown in Fig. 4.16 suggests that while there is overall charge neutrality in the system, there are regions with net charge. At $x = 0$, there is a sheet of positively charged donors with an areal density N_δ. Surrounding this, there is a region with net negative charge that decays the further away we move from the delta-doped layer. The charge imbalance creates an electric field and a potential build-up. A sketch of the complete electrostatic problem, including the energy band diagram, is shown in Fig. 4.17.

Simple as this problem seems, an analytical solution can only be obtained if the delta doping concentration is not too high with respect to the background doping level. We make this assumption below. After solving the problem, we study the limits imposed by this assumption on the value of N_δ.

It is easiest to work with the differential equation expressed in terms of the equilibrium electrostatic potential ϕ_o. If we choose as origin of potentials the value that corresponds to

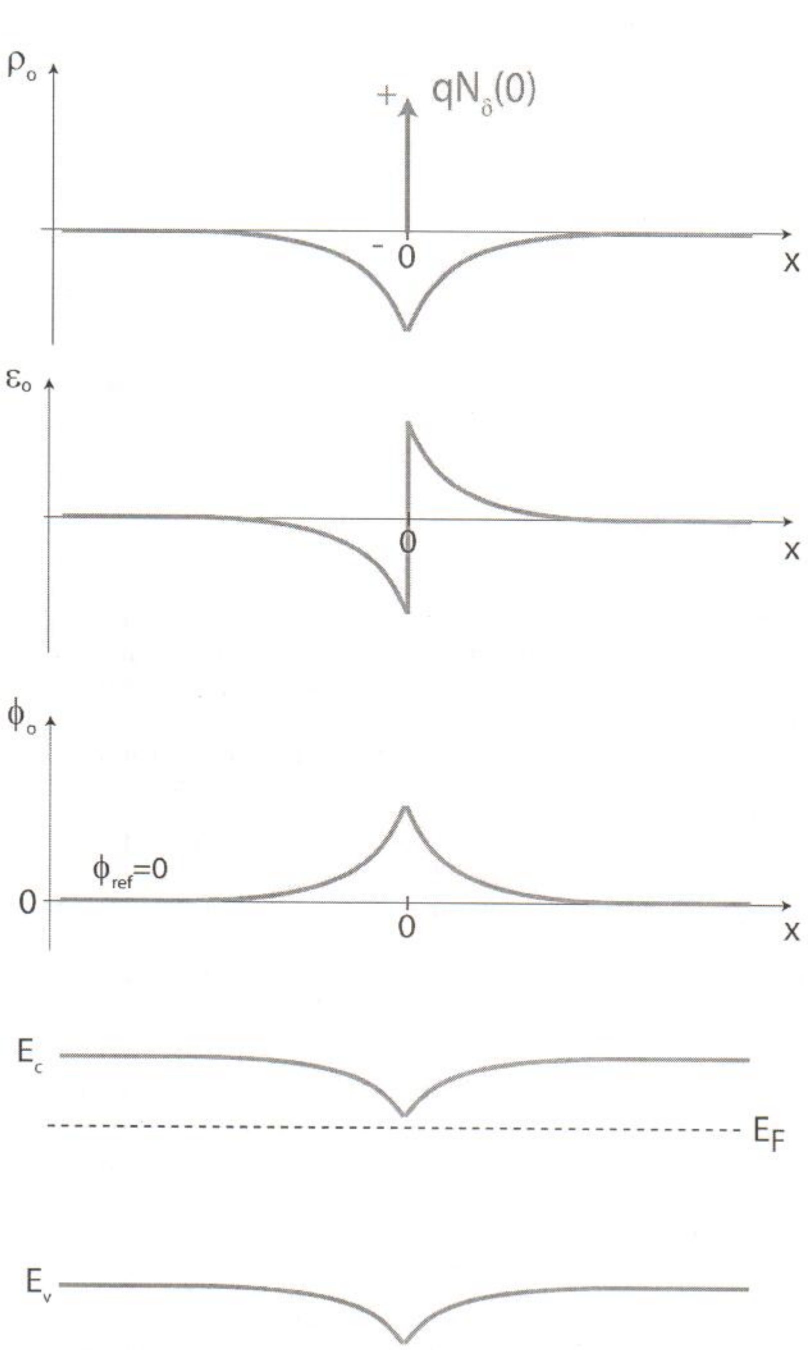


Figure 4.17: Sketch of solution to delta doping problem of Fig. 4.16. From top to bottom: electron concentration, charge density, electric field, electrostatic potential and energy band diagram.

$n_o = N_D$, then following a procedure similar to that of Exercise 4.6, n_o can be written as $n_o(x) = N_D \exp(q\phi_o/kT)$. In terms of ϕ_o, we can rewrite Gauss' law in Eq. 4.53 as:

$$\frac{d^2\phi_o}{dx^2} = \frac{qN_D}{\epsilon}(\exp\frac{q\phi_o}{kT} - 1) \tag{4.66}$$

This equation does not include the donor charge of the delta doped layer and therefore applies only for $x > 0^+$ and $x < 0^-$. We treat the charge associated with the donors in the delta-doped $x = 0$ as a boundary condition to the solution of the differential equation.

This equation cannot be solved analytically in a general case. However, if the doping in the delta doped layer is not too high, the electrostatic potential in the semiconductor will be everywhere small enough in the scale of kT/q for us to be able to expand the exponential in Eq. 4.66 around $\phi_o = 0$ and select its first two terms (see Taylor series expansion in Appendix D). When we do this, the differential equation is simplified to:

$$\frac{d^2\phi_o}{dx^2} \simeq \frac{\phi_o}{L_D^2} \tag{4.67}$$

where L_D has units of length and is defined as:

$$L_D = \sqrt{\frac{\epsilon kT}{q^2 N_D}} \tag{4.68}$$

We discuss the physics of L_D below.

A general solution to Eq. 4.67 can be written as:

$$\phi_o(x) = Ae^{\frac{x}{L_D}} + Be^{-\frac{x}{L_D}} \tag{4.69}$$

If we focus on $x > 0^+$ (the solution for $x < 0^-$ can be easily obtained by symmetry), the coefficient A must be equal to zero since we know that ϕ_o must decay away with x. We obtain the value of B by matching boundary conditions at $x = 0$. At that location, we can create a Gaussian pill box with sides located at $x = 0^-$ and $x = 0^+$. Due to the symmetry of the problem, the field emerging through the right side of the pill box is given by:

$$\mathcal{E}_o(x = 0^+) = \frac{qN_\delta}{2\epsilon} = -\frac{d\phi_o}{dx}|_{x=0^+} = \frac{B}{L_D} \tag{4.70}$$

We solve for B here and insert the result into 4.69, to get:

$$\phi_o(x) = \frac{qN_\delta L_D}{2\epsilon} e^{-\frac{x}{L_D}} \quad \text{for } x > 0^+ \tag{4.71}$$

From this, the electric field can be obtained as:

$$\mathcal{E}_o(x) = -\frac{d\phi_o}{dx} = \frac{qN_\delta}{2\epsilon}\, e^{-\frac{x}{L_D}} \quad \text{for } x > 0^+ \tag{4.72}$$

The volume charge density is given by:

$$\rho_o(x) = \epsilon\frac{d\mathcal{E}_o}{dx} = -\frac{qN_\delta}{2L_D}\, e^{-\frac{x}{L_D}} \quad \text{for } x > 0^+ \tag{4.73}$$

Finally, n_o is given by:

$$n_o(x) = N_D - \frac{\rho_o}{q} = N_D + \frac{N_\delta}{2L_D}\, e^{-\frac{x}{L_D}} \quad \text{for } x > 0^+ \tag{4.74}$$

You can now verify that the integral of $n_o(x)$ from $x = 0^+$ to infinity adds up to $N_D + N_\delta/2$, as it should be. All these results are sketched in Fig. 4.17.

There is a condition that needs to be fulfilled for our solution to be acceptable. The truncated Taylor series expansion of $\exp(q\phi_o/kT)$ is reasonably valid if ϕ_o is not too high in the scale of kT/q. The maximum value of ϕ_o occurs at $x = 0$. If we accept that the maximum tolerable value for ϕ_o is kT/q, then, from Eq. 4.71, we can conclude that $N_\delta < 2N_D L_D$. For higher values than this, our analytical solution is suspect and we need to use numerical techniques to solve this problem.

To summarize, we see that the delta doped layer gives rise to a distribution of volume charge density that has a peculiar shape. At the location of the delta-doped layer, there is a spike of positive charge that arises from the donors. This is surrounded by a cloud of negative charge that is associated with an electron concentration that exceeds the background uniform volume doping level. The region with net charge is limited in space. Sufficiently far away from the delta doped layer, neutrality is reestablished. The solution to the problem exhibits a characteristic length that is given by L_D. The region with substantial volume charge density is confined to a few L_D lengths around $x = 0$.

L_D emerges as the key characteristic length in many electrostatics problems in semiconductors. It is called the **Debye length**[4]. From this problem, we can see that the Debye length has to do with the ability of a semiconductor region to "screen" volume charge. Notice from its expression in Eq. 4.68 that L_D is a property of the semiconductor region surrounding the delta-doped layer and temperature. In this lightly doped limit for N_δ that we are discussing, L_D is not a function of N_δ.

This example illustrates the notion that L_D reflects the balance between drift and diffusion that must exist in a semiconductor in thermal equilibrium. If for some reason, there is net fixed charge in a semiconductor region, carriers redistribute themselves around this region in an effort to screen out the charge as much as possible. The more carriers there are, the more effective they are

---

[4] L_D as defined in Eq. 4.68 is actually called the *extrinsic* Debye length. It is also possible to define an *intrinsic* Debye length in an intrinsic semiconductor. Since we will not use this later parameter in this book, here we will refer to L_D as simply the Debye length.

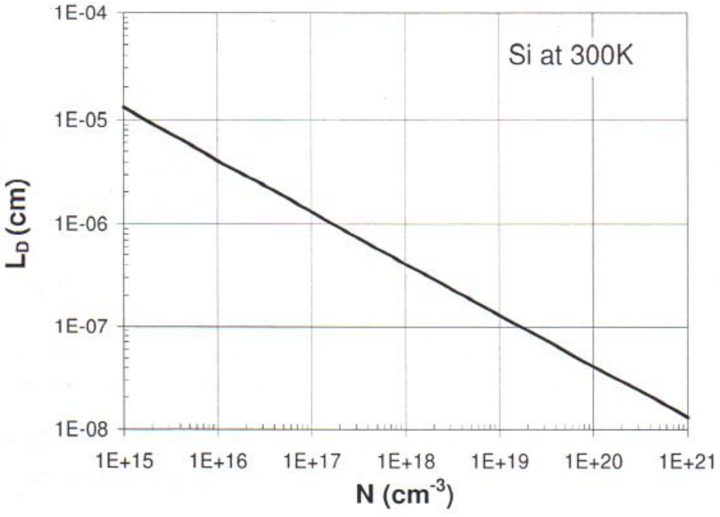


Figure 4.18: Debye length for Si at room temperature as a function of doping level.

at this. This explains the doping dependence of L_D. What works against perfect screening of net fixed charge is diffusion. As carriers pile up around the fixed charge, the concentration gradient that ensues drives them away through diffusion. As the temperature is increased, diffusion gets comparatively stronger (through the Einstein relation) and the Debye length increases. It is the balance between drift and diffusion that sets the value of the Debye length.

Fig. 4.18 graphs the Debye length for Si at room temperature. The Debye length is the same regardless of the doping type. For typical doping levels the Debye length is quite short, of order 10-100 nm. As the doping level increases, the Debye length gets very short. It highly doped regions, any disturbances from neutrality are confined to fairly small dimensions.

Our use of Boltzmann statistics makes the expression for Debye length in Eq. 4.68 and the plot of Fig. 4.18 to not be valid in the degenerate regime. New expressions can be derived that adequately deal with Fermi-Dirac statistics.

Quasi-neutral situations

The example in the previous section leaves clear how unwieldy differential equation 4.65 (or 4.66) is. Analytical solutions are difficult even in relatively simple situations. There is an important class of problems, however, that readily yields analytical solutions. We study them in this section.

Suppose that N_D is a function of space that changes rather slowly with x. In this case, it is reasonable to expect that n_o also displays a slowly changing spatial dependence. In fact, we saw in the delta doped example, that n_o is a somehow "blunter" version of the doping distribution. Under these circumstances, it is possible for the left-hand side term of Eq. 4.65 to be much smaller than either term on the right hand side. If so, it then follows that:

$$n_o(x) \simeq N_D(x) \tag{4.75}$$

Expressed in a different way, if the doping profile changes slowly enough with position, the equilibrium majority carrier concentration closely tracks the doping concentration in space. In such a case, the net volume charge density is very small everywhere. It is because of this that these are called *quasi-neutral situations.*

Once we postulate $n_o(x)$ through Eq. 4.75, the electric field can be obtained from Eq. 4.55, the electrostatic potential by integration of the electric field, and the volume charge density by using Gauss' law in Eq. 4.50. Note that it would be inappropriate to conclude that Eq. 4.75 implies that $\rho_o = 0$. ρ_o is small, that is what the "quasi" in quasi-neutrality means, but not zero.

How can we assess if we are in front of a quasi-neutral situation? Eq. 4.65 allows us to derive a mathematical condition. If Eq. 4.75 holds, then the following inequality has to be fulfilled:

$$|\frac{\epsilon kT}{q^2}\frac{d^2(\ln N_D)}{dx^2}| \ll N_D \tag{4.76}$$

or, equivalently:

$$|\frac{n_o - N_D}{N_D}| \ll 1 \tag{4.77}$$

Eq. 4.76 gives a condition that one can *a priori* check in a given impurity profile. This can be expressed in a more compact way in terms of the Debye length. Using Eq. 4.68 in 4.76, we can write:

$$L_D^2|\frac{d^2(\ln N_D)}{dx^2}| \ll 1 \tag{4.78}$$

We can understand the physical meaning of this condition by inserting Eq. 4.55 in Eq. 4.77 to get:

$$L_D^2|\frac{d\mathcal{E}_o}{dx}| \ll \frac{kT}{q} \tag{4.79}$$

$L_D|d\mathcal{E}_o/dx|$ represents the change in the electric field over a Debye length. In consequence, if $\mathcal{E}_o$ is small, $L_D^2|d\mathcal{E}_o/dx|$ is roughly the change of the electrostatic potential over a Debye length. Eq. 4.79 then leads to a practical rough guideline for quasi-neutrality: *For a non-uniformly doped profile to be quasi-neutral in thermal equilibrium, the change of the electrostatic potential over a Debye length must be much smaller than the thermal voltage.*

This understanding helps us to figure out when is appropriate to consider non-uniformly doped situations as quasi-neutral. What we need is for the doping profile not to change too fast in the scale of L_D. So, the gentler the profile, the less volume charge density is going to appear and the more likely it is that the quasi-neutrality assumption applies. Also, the shorter the Debye length, the sharper the profile can be before the quasi-neutrality approximation becomes inappropriate. From looking at the reduction of L_D with doping level in Fig. 4.18, we can conclude that the higher the doping level, the more likely it is that non-uniformly doped regions can be considered quasi-neutral. This is relevant because in modern microelectronic devices, size reduction and the need to obtain low parasitic resistance has meant that doping levels have been increasing over time. In many devices, moderately and heavily doped regions can be considered quasi-neutral. This greatly facilitates their analysis.

The following example illustrates how the quasi-neutrality assumption is to be used in a non-uniformly doped situation in thermal equilibrium.

Example 2: a hyperbolic doping distribution

Consider an n-type semiconductor with a doping distribution in space given by:

$$N_D(x) = N_{Do} + \Delta N_D \tanh \frac{x}{L_c} \tag{4.80}$$

Depending on the relative values of N_{Do}, ΔN_D and L_c, this profile can be quite steep or rather gradual. An example is shown in Fig. 4.19 for $N_{Do} = 10^{16}\ cm^{-3}$, $\Delta N_D = 5 \times 10^{15}\ cm^{-3}$ and various values of L_c spanning from 10 to 200 nm.

The figure shows also the electron concentration that results from this donor profile. This is obtained by numerically solving differential equation 4.65. As expected, the electron profile is a somehow smeared version of the donor profile. It is interesting to see the role of the characteristic length of this donor distribution, L_c, in the resulting electron concentration. A short value of L_c means that the donor concentration changes rather abruptly in space. In this case, the electron profile ends up being quite different from the donor profile. For long values of L_c, the electron concentration approaches the donor concentration.

The relative discrepancy between $n_o(x)$ and $N_D(x)$ is plotted in the figure below. For the shortest value of L_c, the discrepancy is found to be quite large, as high as 50%. This is clearly a situation with substantial volume charge density in space. For the longest value of L_c, the discrepancy is as small as a few percentage points. This is a quasi-neutral situation.

The Debye length corresponding to the mean doping level of this profile of $N_{Do} = 10^{16}\ cm^{-3}$ is about 40 nm. It is clear, then, that when $L_c \gg L_D$, the profile can be considered quasi-neutral. Otherwise, it cannot.

Under the assumption of quasi-neutrality, it is rather straightforward to solve this problem analytically. The electron concentration is approximately equal to the doping level everywhere:

$$n_o(x) \simeq N_D(x) = N_{Do} + \Delta N_D \tanh \frac{x}{L_c} \tag{4.81}$$

The electric field can be easily calculated using Eq. 4.55:

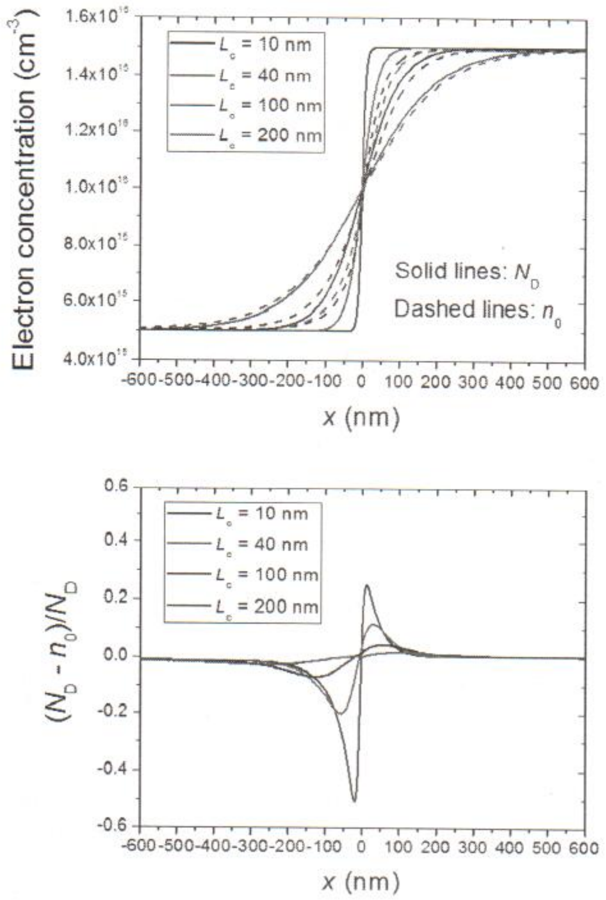


Figure 4.19: Example of non-uniform n-type doping situation. The doping profile is given by Eq. 4.80 with values of $N_D = 10^{16}\ cm^{-3}$, $\Delta N_D = 5 \times 10^{15}\ cm^{-3}$ and various values of L_c spanning from 10 to 200 nm. Top: doping profile and resulting electron concentration. Bottom: relative imbalance between electron concentration and donor concentration *(calculations and graphics courtesy of Ling Xia).*

$$\mathcal{E}_o \simeq -\frac{kT}{q}\frac{1}{L_c}\frac{\Delta N_D}{N_{Do}}\frac{1}{\cosh^2\frac{x}{L_c}+\frac{\Delta N_D}{2N_{Do}}\sinh\frac{2x}{Lc}} \tag{4.82}$$

If we select as origin of potentials $\phi(n_i) = 0$, the electrostatic potential distribution is easily obtained from Eq. 4.60:

$$\phi_o(x) \simeq \frac{kT}{q}\ln\frac{1}{n_i}(N_{Do} + \Delta N_D \tanh\frac{x}{L_c}) \tag{4.83}$$

Finally, the volume charge distribution can be obtained from Gauss' law, Eq. 4.50:

$$\rho_o \simeq \epsilon \frac{kT}{q} \frac{\Delta N_D}{L_c^2} \frac{N_{Do} \sinh \frac{2x}{L_c} + \Delta N_D \cosh \frac{2x}{L_c}}{(N_{Do} \cosh^2 \frac{x}{L_c} + \frac{\Delta N_D}{2} \sinh \frac{2x}{Lc})^2} \tag{4.84}$$

For $x \geq 0$, ρ_o is positive since there are a few more donors than electrons. For $x \leq 0$, there are more electrons than donors and ρ_o is negative. All these results are graphed in Fig. 4.20.

We can now check the quasi-neutrality condition. From Eq. 4.77, quasi-neutrality prevails when:

$$|\frac{n_o - N_D}{N_D}| = |\frac{1}{N_D} \frac{\rho_o}{q}| = |\frac{L_D^2}{L_c^2} \frac{\Delta N_D}{N_{Do}} \frac{\sinh \frac{2x}{L_c} + \frac{\Delta N_D}{N_{Do}} \cosh \frac{2x}{L_c}}{(\cosh^2 \frac{x}{L_c} + \frac{\Delta N_D}{2N_{Do}} \sinh \frac{2x}{L_c})^2}| \ll 1 \tag{4.85}$$

The fraction with the hyperbolic functions is at most unity. We can then rewrite the quasi-neutrality condition as:

$$|\frac{n_o - N_D}{N_D}| \leq (\frac{L_D}{L_c})^2 \frac{\Delta N_D}{N_{Do}} \ll 1 \tag{4.86}$$

In a case in which ΔN_D is of the same order as N_{Do}, quasi-neutrality is satisfied when the characteristic length L_c over which the doping profile changes is somehow larger than the Debye length L_D. This is precisely what we found in the numerical results above.

This example illustrates how in spite of assuming $n_o(x) \simeq N_D(x)$ in the quasi-neutral approximation, we can still derive a first-order expression for the electric field and charge distribution in space that result from the small imbalance that exists between $n_o(x)$ and $N_D(x)$.

4.6 Quasi-Fermi levels and quasi-equilibrium

In Ch. 2 we learned that the Fermi level in a semiconductor uniquely characterizes the electron and hole densities in equilibrium as well as their energy distributions in their respective bands. Eqs. 2.10 and 2.20 related the equilibrium carrier densities to the effective density of states in the corresponding bands. A single value of E_F established this connection for both electrons and holes simultaneously.

In general, outside equilibrium the carrier densities are different from their equilibrium values. It is then not possible to find a single value of E_F that simultaneously relates n with N_c and p with N_v in the manner that Eqs. 2.27 and 2.35 do. We can however *define* two "quasi-Fermi levels," one for electrons and another one for holes such that the functional relationships between carrier densities and effective density of states given in Eqs. 2.27 and 2.35 are preserved, *i.e.*:

$$n = N_c \mathcal{F}_{1/2}(\frac{E_{fe} - E_c}{kT}) \tag{4.87}$$

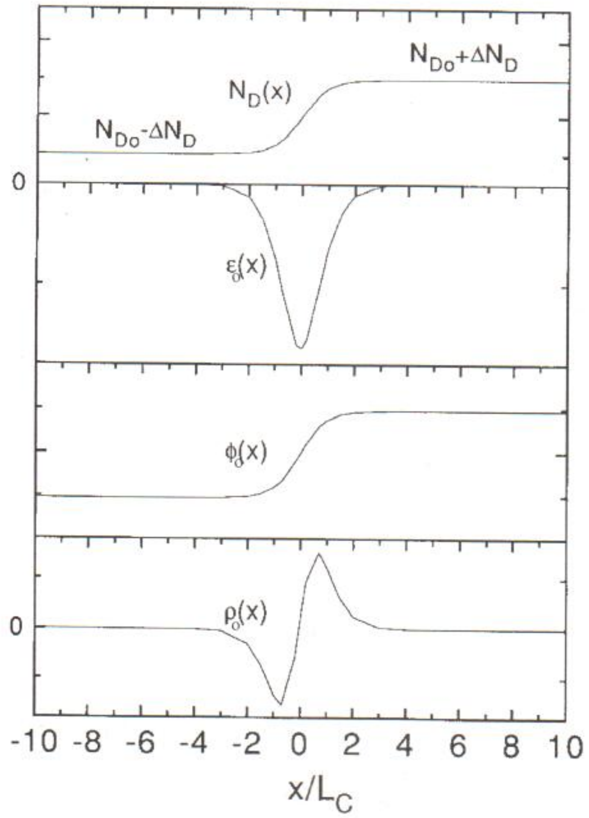


Figure 4.20: Example of quasi-neutral doping profile given by $N_D(x) = N_{Do} + \Delta N_D \tanh \frac{x}{L_c}$. From top to bottom: N_D, $\mathcal{E}_o$, ϕ_o, and ρ_o as a function of the normalized space coordinate x/L_c.

$$p = N_v \mathcal{F}_{1/2}(\frac{E_v - E_{fh}}{kT}) \tag{4.88}$$

For clarity, we first simplify Eqs. 4.87 and 4.88 for non-degenerate situations:

$$n = N_c \exp \frac{E_{fe} - E_c}{kT} \tag{4.89}$$

$$p = N_v \exp \frac{E_v - E_{fh}}{kT} \tag{4.90}$$

The quasi-Fermi levels defined in this way serve as excellent visualization tools for carrier concentrations and generation and recombination in semiconductors out of equilibrium. By their very definition, carrier concentrations are directly related to the distance of each quasi-Fermi level to its corresponding band edge through Eqs. 4.89 and 4.90. The usefulness of these two equations goes beyond this. Note that we can write the np product as:

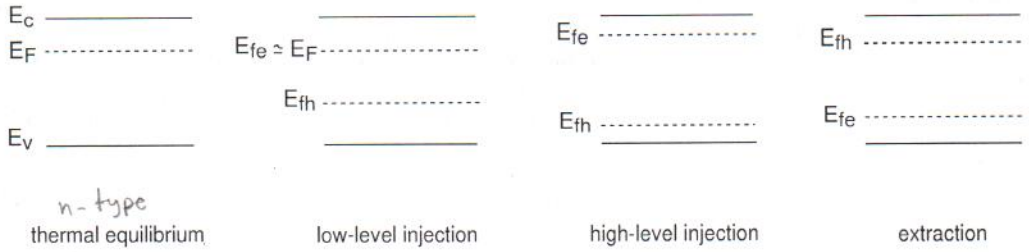


Figure 4.21: Illustration of the location of the Fermi level and quasi-Fermi levels in an n-type sample under four different situations: thermal equilibrium, low-level injection, high-level injection, and extraction.

$$np = n_i^2 \exp \frac{E_{fe} - E_{fh}}{kT} \tag{4.91}$$

This implies that if E_{fe} is drawn above E_{fh}, then $np > n_i^2$. This is a situation outside equilibrium with excess carrier concentrations where net recombination is taking place. On the other hand, if $E_{fe} < E_{fh}$, then $np < n_i^2$ and the semiconductor is below equilibrium. Net generation is taking place. If the two quasi-Fermi levels coincide, then $np = n_i^2$ and the carrier concentrations are in thermal equilibrium. All these facts can be quickly verified in a band diagram by checking the relative position of the two quasi-Fermi levels.

As an illustration, Fig. 4.21 sketches four different energy band diagrams depicting different situations. The diagram on the left corresponds to an n-type sample in thermal equilibrium. In this case, there is only one Fermi level. The next diagram depicts a situation in which the same sample is under low-level injection resulting, for example, from external light illumination. Since under low-level injection $n \simeq N_D$, the quasi-Fermi level for electrons has not moved from its equilibrium position. The quasi-Fermi level for holes, on the other hand, moves towards the valence band because $p \gg p_o$. Note also that E_{fe} is above E_{fh}. This means that net recombination is taking place in this region of the semiconductor.

The third diagram is under high-level injection conditions, in which both carrier concentrations are increased from their equilibrium values. In consequence, both quasi-Fermi levels move towards their respective bands. If the sample remains quasi-neutral (as is usually the case), the distance between each quasi-Fermi level and its corresponding band is about the same (we are assuming here that the effective density of states of the conduction and valence bands are not too different from each other).

The diagram on the far right represents the same sample in extraction, with the carrier concentrations below their equilibrium values. In the energy band diagram, this is depicted by the quasi-Fermi levels becoming further away from their respective bands. In extraction, E_{fh} is above E_{fe} and net generation prevails.

Exercise 4.7: *Calculate the relative location with respect to the band edges of the Fermi level or quasi-Fermi levels, as appropriate, for a uniformly-doped n-type Si sample with* $N_D = 10^{16}\ cm^{-3}$ *at room temperature in the following situations:* i) *in equilibrium;* ii) *in low-level injection such that there are* $10^{14}\ cm^{-3}$ *holes;* iii) *in high-level injection such that there are* $10^{18}\ cm^{-3}$ *electrons; and* iv) *in extraction where the equilibrium carrier concentrations have been reduced by 10 orders of magnitude.*

i) In thermal equilibrium, $n_o \simeq N_D = 10^{16}\ cm^{-3}$ and $p_o = n_i^2/n_o \simeq 10^4\ cm^{-3}$. Using Eq. 2.30, the Fermi level is found to be:

$$E_c - E_F = kT \ln \frac{N_c}{n_o} = 0.026\, ln \frac{2.9 \times 10^{19}}{10^{16}} = 0.21\ eV$$

ii) In low-level injection with $p = 10^{14}\ cm^{-3}$, the electron concentration does not change from its equilibrium value. In consequence, E_{fe} is located with respect to E_c just where E_F was in equilibrium. Hence, $E_c - E_{fe} = 0.21\ eV$. The location of E_{fh} can be obtained from Eq. 4.90:

$$E_{fh} - E_v = kT \ln \frac{N_v}{p} = 0.026\, ln \frac{3.1 \times 10^{19}}{10^{14}} = 0.33\ eV$$

iii) Under high-level injection conditions with $n = 10^{18}\ cm^{-3}$, p takes the same value. In consequence,

$$E_c - E_{fe} = kT \ln \frac{N_c}{n} = 0.026\, ln \frac{2.9 \times 10^{19}}{10^{18}} = 0.088\ eV$$

$$E_{fh} - E_v = kT \ln \frac{N_v}{p} = 0.026\, ln \frac{3.1 \times 10^{19}}{10^{18}} = 0.089\ eV$$

Both quasi-Fermi levels approach their respective bands.

iv) Under extraction conditions with ten orders of magnitude fewer carriers than in equilibrium, the location of the quasi-Fermi levels is:

$$E_c - E_{fe} = kT \ln \frac{N_c}{n} = 0.026\, ln \frac{2.9 \times 10^{19}}{10^{6}} = 0.81\ eV$$

$$E_{fh} - E_v = kT \ln \frac{N_v}{p} = 0.026\, ln \frac{3.1 \times 10^{19}}{10^{-6}} = 1.53\ eV$$

Notice that E_{fe} has gotten rather close to the valence band edge while E_{fh} has penetrated into the conduction band.

Quasi-Fermi levels are not only great tools to visualize carrier concentrations and generation and recombination, but also they help us picture carrier flow. To realize this, let us examine the expressions of the carrier currents outside thermal equilibrium. Let us do it first for electrons. If the electric field is not very high, from Eqs. 4.21 and 4.40, the electron current is given by:

$$J_e = q\mu_e n\mathcal{E} + qD_e \frac{dn}{dx} \tag{4.92}$$

This equation can be rewritten in a very interesting way. Taking the derivative with respect to x in Eq. 4.89, we find:

$$\frac{dn}{dx} = \frac{n}{kT}\frac{dE_{fe}}{dx} - \frac{q}{kT}n\mathcal{E} \tag{4.93}$$

Substituting this into Eq. 4.92 and using the Einstein relation Eq. 4.38, we finally obtain:

$$J_e = \mu_e n \frac{dE_{fe}}{dx} \tag{4.94}$$

Had we carried out an identical exercise for holes, we would have obtained an equivalent expression:

$$J_h = \mu_h p \frac{dE_{fh}}{dx} \tag{4.95}$$

Eqs. 4.94 and 4.95 are extremely important and useful relationships. They state that the gradient of the quasi-Fermi levels can be viewed as a sort of unified driving force for carrier flow that combines drift and diffusion. The reason behind this is made obvious if we equate Eqs. 4.94 and 4.95, respectively, to the fundamental current relationships 4.18 and 4.20. Solving for the gradient of the quasi-Fermi level we find:

$$\frac{dE_{fe}}{dx} = -\frac{q}{\mu_e}v_e \tag{4.96}$$

$$\frac{dE_{fh}}{dx} = \frac{q}{\mu_h}v_h \tag{4.97}$$

For both electrons and holes, we find that the gradient of the quasi-Fermi level is proportional to the carrier velocity, indeed, a very fundamental property of a carrier ensemble. It should not be surprising, then, that even though equations 4.94 and 4.95 were derived under fairly restrictive conditions (low-field and Maxwell-Boltzmann statistics), they are in fact very general. The same expressions can be obtained in a degenerate semiconductor and also in the presence of high electric fields.

The quasi-Fermi levels provide great insight into carrier flow in a variety of situations. In the presence of a complex situation with drift and diffusion simultaneously taking place, the behavior of the quasi-Fermi levels quickly indicates such things as the net direction of carrier flow and the relative magnitude of carrier velocity. This can be understood by examining Eqs. 4.96 and 4.97 which state that the gradient of the quasi-Fermi level is proportional to the corresponding carrier velocity. This suggests that if in a certain region a quasi-Fermi level is depicted as flat, then the corresponding carrier velocity is zero or very small. If a quasi-Fermi level exhibits a slope, then carriers are flowing. The carrier velocity is proportional to the slope of the quasi-Fermi level in the band diagram.

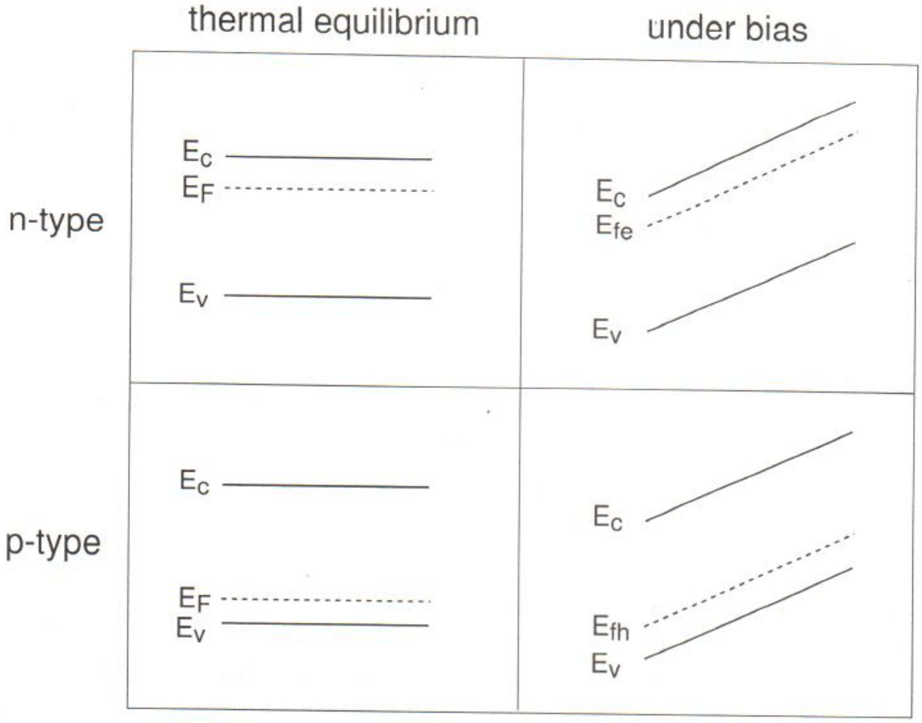
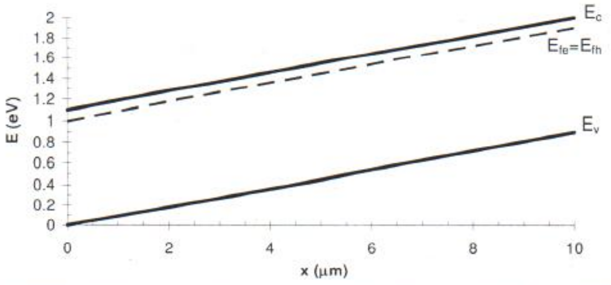


Figure 4.22: Illustration of current visualization in energy band diagrams through quasi-Fermi levels. Top two diagrams depict a uniformly-doped n-type semiconductor in thermal equilibrium (left) and under bias (right). Bottom two diagrams show a uniformly-doped p-type semiconductor in thermal equilibrium (left) and under bias (right).

Fig. 4.22 shows several examples. The top left diagram shows a uniformly-doped n-type semiconductor in thermal equilibrium. We can judge that this is the case because the Fermi level is labeled E_F and it is sketched flat. To the right, is the same semiconductor under bias. We understand this because the conduction and valence bands are shown inclined in space. We also see the quasi-Fermi level for electrons running parallel to them. With an imaginary x-axis running from left to right in this diagram, the gradient of E_{fe} would be positive, and therefore Eq. 4.96 would suggest that the electron velocity is negative, that is, electrons flow from right to left down the conduction band.

The two figures underneath depict similar situations for a uniformly-doped p-type semiconductor under thermal equilibrium (on the left) and under bias (on the right). In this last case, hole current flows also from left to right which corresponds to a hole flow from left to right.

Exercise 4.8: *Estimate the current density flowing through a Si sample with the energy band diagram sketched below at room temperature.*



This energy band diagram reveals a sample with a uniform electric field applied to it. We could compute the current density from the drift expression Eq. 4.23. As we have just learned in this section, we can also use Eqs. 4.94 and 4.95. Let us proceed this way.

In the diagram, the two quasi-Fermi levels overlap. They are also fairly close to the conduction band but quite far away from the valence band. This implies that $n \gg p$ and we only need to account for electron current. The electron concentration can be estimated from the 0.1 eV distance between E_c and E_{fe}. As we learned in Ch. 2, this yields:

$$n = N_c \exp\frac{E_{fe} - E_c}{kT} = 2.9 \times 10^{19} \exp\frac{-0.1}{0.0259} = 6.1 \times 10^{17}\ cm^{-3}$$

This is then an n-type sample with $N_D \simeq 6.1 \times 10^{17}\ cm^{-3}$. For this doping level, Fig. 4.3 shows that the electron mobility is about 300 $cm^2/V.s$.

The gradient of E_{fe} is:

$$\frac{dE_{fe}}{dx} = \frac{0.9\ eV}{10 \times 10^{-4}\ cm} = 900\ eV/cm$$

Putting it all together into Eq. 4.94, we finally get:

$$J_e = 300\ cm^2/V.s \times 6.1 \times 10^{17}\ cm^{-3} \times 900\ eV/cm = 1.7 \times 10^{23}\ e/cm^2 \cdot s = 2.6 \times 10^4\ A/cm^2$$

The concept of quasi-Fermi level hinges on the notion of *quasi-equilibrium*. Basically, even though we are dealing with situations outside equilibrium, the carrier distribution in energy never departs too far from thermal equilibrium. The fundamental reason for this is the fact that the scattering time, or average time between collisions, is several orders of magnitude smaller than the time scales that characterize dynamic carrier behavior in semiconductor devices (depending on the particular device, this would be the carrier lifetime or the transit time due to drift or diffusion, as we will see in the next chapter). Thus, on time scales of interest to us, carriers undergo many scattering events that keep their energy distribution in close thermal equilibrium with the lattice. It is for this reason that it is possible to define a quasi-Fermi level that plays a role equivalent to the Fermi level in strict thermal equilibrium. Often, this near equilibrium carrier distribution in energy is called a *Maxwellian distribution*.

Figure 4.23 helps to clarify the concept of quasi-equilibrium in a situation with excess carriers. The figure shows the sequence of events that follows the sudden generation of electron-hole pairs by a pulse of energetic photons. Let us assume that the energy of the photon beam is high

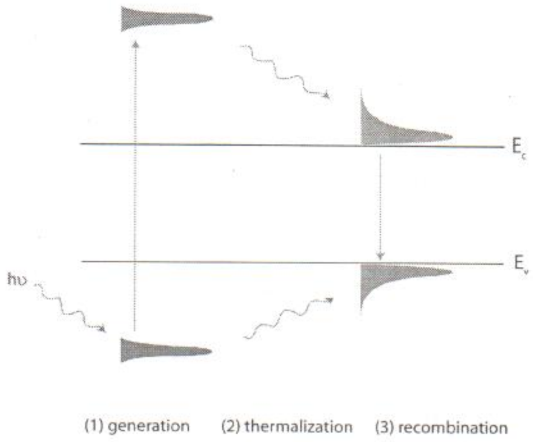


Figure 4.23: Sketch illustrating the applicability of the notion of quasi-equilibrium to describe the recombination process that follows a burst of hot carrier generation by a pulse of energetic photons. Hot electrons and holes lose their extra energy on a time scale much shorter than the recombination lifetime. As a result, recombination takes place among carriers that have reached a quasi-equilibrium state with the lattice and are well described by their respective quasi-Fermi levels.

enough so that right after the generation event hot electrons and hot holes with kinetic energies substantially higher than the thermal energy have been produced (left diagram). Right after this burst of carrier generation, two things are going to happen. First, electrons and holes will start losing excess energy by phonon emission, a process that is known as "thermalization." In addition, excess electrons and holes will start recombining with each other. For common semiconductors under typical situations, the time it takes to "thermalize" the hot carriers (a few scattering times, or $< 1\ ps$) is much shorter than the recombination lifetimes ($\mu s - ms$). Therefore, thermalization takes place first (middle diagram) and recombination occurs after that (right diagram). In this scenario, recombination events mostly involve excess electrons and holes that have reached a quasi-equilibrium state with the lattice. The recombination dynamics can therefore be studied with the electron and hole distributions described through their respective quasi-Fermi levels.

In the situation described here, while recombination is taking place, the electrons are in near equilibrium among themselves and with the lattice (meaning, they suffer many collisions with the lattice in the time scale of the carrier lifetime). Similarly, holes are also near equilibrium among themselves and with the lattice. However, the electrons are not in equilibrium with the holes. That is what the process of recombination is trying to correct and it will take several lifetimes for equilibrium to be established. It is for this reason that we need to define two different quasi-Fermi levels that characterize the Maxwellian distributions of electrons and holes separately.

It is important to also think about the implications and constraints of quasi-equilibrium in a different situation, one in which drift is the dominant mechanism. Consider a semiconductor bar with a certain electron concentration and an electric field applied (we study how to do this in the next chapter). In steady state, as electrons drift, they pick up kinetic energy from the electric field and release it to the lattice through phonon emission. Quasi-equilibrium means that the electron distribution does not significantly depart from the equilibrium one. We can ensure this if in the time scale in which carriers would pick up a thermal energy worth of kinetic energy from

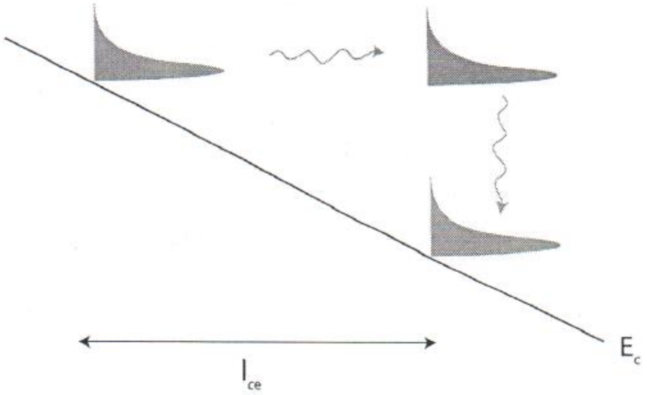


Figure 4.24: Sketch illustrating the applicability of the notion of quasi-equilibrium to describe electron drift in an electric field. For quasi-equilibrium to apply, electrons must not gain more than about a kT worth of kinetic energy from the electric field in a mean free path.

the electric field, they suffer many collisions that dissipate this energy into the lattice through phonon emission. This applies if the distance over which the potential energy changes by an amount kT is much longer than a mean free path. This is the relevant distance because if carriers suffered no energy dissipating collisions, as the potential energy associated with the electric field drops by kT, their kinetic energy increases by the same amount (Fig. 4.24).

Let's define this distance as λ_{kT}. We need that $\lambda_{kT} \gg l_{ce}$. Since, to the first order, $\lambda_{kT} = kT/\mathcal{E}$, we must then require that $\mathcal{E} \ll kT/l_{ce}$. l_{ce} depends on doping level. For the value determined in Exercise 4.1, $l_{ce} = 17\ nm$, the constraint on the electric field is $\mathcal{E} \ll 1.5 \times 10^4\ V/cm$. Higher electric field values are tolerable if the mean-free path is shorter. The order of magnitude of this maximum electric field is quite small for electrons in Si at room temperature. This means that in many Si devices today, carriers in an electric field can get quite "hot" and quasi-equilibrium is not an appropriate concept. When this happens, the use of the quasi-Fermi level is suspect. We will still use this concept to sketch energy band diagrams in these situations mostly for convenience. It is important to realize that we are really operating beyond the usability of this concept.

4.7 Summary

- Electrons and holes are charged particles that can move in a semiconductor ("carriers"). If they do, an electric current flows. The current density that is produced is directly proportional to the carrier concentration and the carrier velocity:

$$J_e = -qnv_e \qquad\qquad J_h = qpv_h$$

- Carriers drift in response to the application of an electric field. If the electric field is small enough, carriers drift with a velocity that is proportional to the field. The current density is then:

$$J_t = q(\mu_e n + \mu_h p)\mathcal{E}$$

- For high enough fields, a saturation velocity is reached. The current density becomes:

$$J_{t,sat} = q(n v_{esat} + p v_{hsat})$$

- Carriers diffuse as a result of concentration gradients. The diffusion currents for electrons and holes are, respectively:

$$J_e = qD_e \frac{dn}{dx} \qquad\qquad J_h = -qD_h \frac{dp}{dx}$$

- Detailed balance in thermal equilibrium demands that $J_e = J_h = 0$ everywhere.
- In a non-uniformly doped semiconductor in thermal equilibrium, a balance between carrier drift and diffusion results in a majority carrier profile in space that in general is different from the doping concentration distribution. This results in a spatial charge and an electric field inside the semiconductor. The electrostatic potential distribution is given by:

$$\phi_o(x) - \phi_{ref} = \frac{kT}{q} \ln \frac{n_o(x)}{n_o(\text{ref})} = \frac{kT}{q} \ln \frac{p_o(\text{ref})}{p_o(x)}$$

- For a reasonably extrinsic semiconductor, if the doping distribution does not change too abruptly in space, the semiconductor remains "quasi-neutral" and the majority carrier concentration closely tracks the doping level.
- The time scale over which carriers can have energy distributions that are very different from equilibrium is much shorter than the time scales of interest in typical device operation, such as carrier lifetimes or transit times. For practical purposes, the carrier energy distributions can be considered close to equilibrium.
- Outside equilibrium, it is useful to define electron and hole quasi-Fermi levels as follows:

$$n = N_c \mathcal{F}_{1/2}(\frac{E_{fe} - E_c}{kT}) \qquad\qquad p = N_v \mathcal{F}_{1/2}(\frac{E_v - E_{fh}}{kT})$$

- Gradients of quasi-Fermi levels can be considered as unified driving forces for carrier flow, regardless of whether the carriers move by drift, diffusion, or a combination of both processes:

$$J_e = \mu_e n \frac{dE_{fe}}{dx} \qquad\qquad J_h = \mu_h p \frac{dE_{fh}}{dx}$$

The gradient of the quasi-Fermi level is linearly proportional to the net carrier velocity.

4.8 Further reading

There are several books with material relevant to this chapter.

The Feynman Lectures on Physics, Volume I by R. P. Feynman, R. B. Leighton, and M. Sands, Addison-Wesley 1963 (ISBN 0-201-02116-1-P, QC23.F435). Ch. 14 discusses forces and potential energy. Ch. 40 introduces the Boltzmann law and the velocity distribution of ideal particles. Ch. 43 contains an intuitive and easy to understand discussion about thermal motion, drift, and diffusion of gas ions and molecules. These chapters make easy but extremely illuminating reading. Highly recommended.

Introduction to Semiconductor Physics by R. B. Adler, A. C. Smith, and R. L. Longini, Wiley, 1964 (ISBN 0-471-00887-7, QC612.S4.S471 v.1). The first volume of the pioneering SEEC (Semiconductor Electronics Education Committee) series is nothing but a masterpiece. To this date, the clarity of the explanations and the physical insight that this little book provides has remained unmatched by any other textbook. In spite of its age, the selection of topics is still relevant and of high priority reading for serious students of semiconductor device physics. Material that is significant to this chapter is presented in sections 3.6 (non-uniformly doped situations) and 4.2 (Gauss' law and charge neutrality).

Fundamentals of Solid-State Electronics by C.-T. Sah, World Scientific, 1991 (ISBN 9810206372, TK7871.85.S23). In addition to other topics relevant to other chapters, Ch. 3 of Sah's book describes in detail drift, diffusion and the quasi-Fermi levels. In particular, there is a fairly thorough and quantitative presentation of the various scattering mechanisms. The discussion surrounding the introduction of the quasi-Fermi levels is insightful.

Fundamentals of Semiconductor Theory and Device Physics by S. Wang, Prentice Hall, 1989 (ISBN 0-13-344409-0, QC611.W32). Always at a substantially higher level, Wang's book has a lot of material that merits reading, even if ignoring the math. The scattering mechanisms are carefully and rigorously described in §6.7. §§10.2 and 10.8 have excellent discussions on velocity saturation. §§10.3 and 10.9 describe the electron transfer effect and the peculiar velocity-field characteristics of III-V semiconductors.

Fundamentals of Carrier Transport by M. Lundstrom, 2nd Edition, Cambridge University Press, 2000 (ISBN 0-521-63134-3). This is an excellent treatment of carrier transport in semiconductors far more rigorous and at a substantially higher level than the present text. Particularly relevant here are Ch. 2 on carrier scattering, Ch. 3 on the Boltzmann transport equation, and Ch. 4 on low-field transport.

AT4.1 Selected properties of the Gamma function

The Gamma function is defined as:

$$\Gamma(p) = \int_0^\infty \eta^{p-1} e^{-\eta} d\eta \tag{4.98}$$

In situations involving semiconductors, the Gamma function typically makes an appearance for a few positive integer or fractional values of p.

For real values of p (other than zero or negative integers), the gamma function has the following property:

$$\Gamma(p+1) = p\Gamma(p) \tag{4.99}$$

For positive integer values of p, let us refer to them as n, the Gamma function has the following property:

$$\Gamma(n) = (n-1)! \tag{4.100}$$

Commonly used values of the Gamma function are:

$$\Gamma(1) = 1 \tag{4.101}$$

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \tag{4.102}$$

Using these and the properties in Eqs. 4.99 and 4.100, other common values of the Gamma function can be easily derived.

AT4.2 Hot carrier effects

A charged particle immersed in an electric field drifts due to the electrostatic force that acts upon it. Between two scattering events, the carrier acquires a velocity component in the direction of the electric field. At the same time, the particle extracts kinetic energy from the potential energy of the field. If the additional kinetic energy is much smaller than the thermal energy, then the average energy of the particle is not much affected and the scattering rates are insignificantly upset from their equilibrium values. This is a quasi-equilibrium situation. If the particle acquires a kinetic energy from the field that is comparable or exceeds the thermal energy, the physics of the situation change substantially. This is known as the "hot"-carrier regime.

When carriers become hot, interesting effects occur. Earlier in this chapter we mentioned velocity saturation. In Ch. 3, we briefly described impact ionization, avalanche multiplication, and avalanche breakdown. These are all hot-carrier effects. The study of the physics of hot carriers is a rich field with a plethora of activity. As microelectronic devices shrink in size, the magnitude of the electric fields inside tends to increase and hot-carrier effects become more prominent and worrysome. Device designers ignore them at their own peril!

The aim of this Advanced Topics Section is to outline some of the underlying physics of hot-carrier effects and some of the most important consequences for microelectronic device designers. In order to do this, we must first look in more detail at the physics of phonon scattering. This leads into a discussion about the criteria for the onset of hot-carrier effects. We then study in more detail two important effects: hot-carrier transport and impact ionization. Two related physical phenomena, avalanche multiplication and avalanche breakdown are discussed in Ch. 5.

AT4.2.1 Energy relaxation vs. momentum relaxation

In a solid, not all phonons are created equal. A complex lattice, such as the one of a semiconductor, allows many vibrational modes. All of these modes are very effective in exchanging *momentum* with carriers, that is, in randomizing a carrier's velocity. Some of these modes, however, are more effective than others in exchanging *energy* with carriers. In typical semiconductors, such as Si and GaAs, the most frequent collisions between phonons and carriers succeed in randomizing the velocity of the carrier but exchange little energy (a few meV). Only collisions with special kinds of phonons, the so-called *optical phonons*, exchange a sizable amount of energy, the optical phonon energy, E_{opt}, a few tens of eV.

This important distinction is captured in two definitions: the *momentum relaxation time*, τ_M, and the *energy relaxation time*, τ_E. τ_M is the average time required to randomize the momentum of a carrier. τ_E is the average time required for a carrier to lose any excess energy it might have over the thermal energy. In most semiconductors $\tau_M < \tau_E$. In lowly-doped Si, for example, under quasi-equilibrium conditions (for low fields), $\tau_M \simeq 0.2\ ps$, while $\tau_E \simeq 0.5\ ps$.

When a carrier is immersed in an electric field, this distinction between momentum and energy relaxation matters a lot. As we discussed above, between two collisions, a carrier picks up not only velocity (or momentum) in the direction of the electric field, but also kinetic energy. A momentum-exchanging collision randomizes the velocity, and therefore affects the net forward velocity of the carrier, but only absorbs a small amount of energy. In consequence, the average energy of the carriers can build up. When this energy is high enough, optical phonon emission becomes more likely and the release of energy to the lattice is enhanced. Eventually a steady state situation is reached in which on average the carrier ensemble does not change its kinetic energy. The time that it takes for this steady state situation to be established is on the order of the energy relaxation time. If the electric field is suddenly turned off, it will take a time of a few times τ_E for the carriers to thermalize by releasing their excess kinetic energy to the lattice.

The steady-state kinetic energy acquired by the carriers from an electric field is easy to estimate. Let us consider a chunk of uniformly-doped n-type semiconductor of cross-sectional area A and length L (similar arguments apply for a p-type semiconductor). With a voltage V

applied to it, a current I flows. The power dissipated in this piece of semiconductor is:

$$W = |IV| = qnv_e^{drift}AL\mathcal{E} \tag{4.103}$$

where we have used Eq. 4.18 and we have assumed that the electric field is uniform along the length of the sample.

Since nAL is the total number of electrons in the sample, the power dissipated per electron is:

$$w = qv_e^{drift}\mathcal{E} \tag{4.104}$$

For an electron it takes a mean time τ_E to release its extra kinetic energy to the latice. On average, then, the kinetic energy picked up from the field by one electron is:

$$\Delta E_K = qv_e^{drift}\mathcal{E}\tau_E \tag{4.105}$$

This result indicates that the excess kinetic energy is linearly proportional to the electric field. This is actually what is found if detailed calculations are performed, as shown in Fig. 4.25. This figure shows the average kinetic energy of electrons, E_K, in Si at room temperature in a uniform electric field, as a function of the electric field, as calculated using a Monte Carlo technique by Fischetti. Below a field of about 8 kV/cm, the average energy of the electrons is equal to the thermal energy. Above this field, the average energy increases rapidly. For a field of 200 kV/cm, for example, the average energy is about 1 eV above the thermal energy.

When the excess kinetic energy is comparable to the thermal energy, hot-electron effects become important. For n-type Si, for example, using a value of $\tau_E \simeq 0.5\ ps$, a value of $\mathcal{E} = 7.8\ kV/cm$ is required for ΔE_K to equal the thermal energy $\frac{3}{2}kT = 39\ meV$ at room temperature.

The energy relaxation time depends slightly on the average energy. Fig. 4.26 shows its dependence on the magnitude of a uniform electric field in Si, also calculated by Fischetti. A consequence of the weak dependence of τ_E on $\mathcal{E}$, is that the average energy increases nearly linearly with electric field, as shown in Fig. 4.25.

AT4.2.2 Hot-electron transport

An important consequence of carriers acquiring additional kinetic energy in an electric field is that the likelihood of velocity randomizing collisions increases. This is because the higher the energy, the larger the density of states of the bands and the more states the carriers can scatter into. In consequence, the momentum relaxation time decreases as the average carrier energy increases. At high enough energy, $\tau_M \sim 1/E_K$ (this is a consequence of $g(E) \sim \sqrt{E}$). In a uniform electric field, since $E_K \sim \mathcal{E}$, then $\tau_M \sim 1/\mathcal{E}$, as seen in Fig. 4.26.

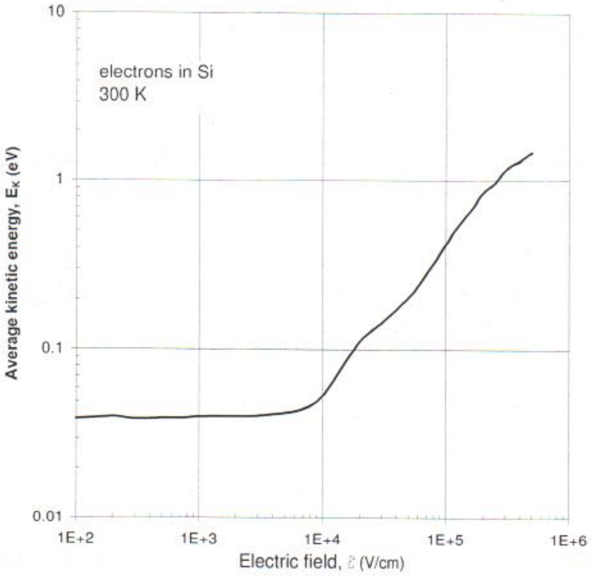


Figure 4.25: Average kinetic energy of electrons in uniform electric field in Si at room temperature as a function of the electric field [data courtesy of M. Fischetti, IBM].

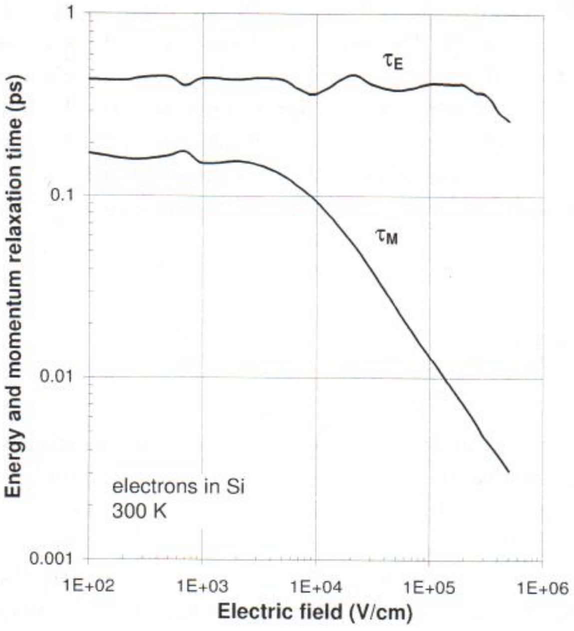


Figure 4.26: Energy and momentum relaxation times for electrons immersed in a uniform electric field in Si at room temperature as a function of electric field [data courtesy of M. Fischetti, IBM].

A decrease in τ_M brings with it, according with Eq. 4.10, a reduction in the mobility. Eventually, at high enough fields, since $\tau_M \sim 1/\mathcal{E}$, the mobility drops also as $\sim 1/\mathcal{E}$, and the drift velocity saturates.

A rough estimate of the saturation velocity can be obtained from the following simple argument. At high electric fields under steady-state conditions, the kinetic energy acquired from the electric field by a carrier in between two collisions must, on average, be entirely released at the next collision. If that were not the case, the average carrier energy would still be increasing and a steady state situation would not be reached. The maximum energy that a collision with the lattice can release, as discussed in Ch. 1, is the optical phonon energy, E_{opt}. In consequence, equating the kinetic energy acquired between collisions to E_{opt}, yields (within a small numerical factor):

$$v_{sat} \simeq \sqrt{\frac{8}{3\pi}\frac{E_{opt}}{m_c^*}} \tag{4.106}$$

The factor of $8/3\pi$ takes care of the statistical distribution of the carrier velocities. This equation is derived under the assumption of an ideal parabolic band. At high energies, the bands of Si and GaAs are more complicated than this simple picture. In spite of that, Eq. 4.106 predicts values of v_{sat} for electrons and holes in Si and GaAs that are fairly close to the measured ones.

Eq 4.106 suggests that the saturation velocity is temperature independent. Experimentally, however, it is found that there is a weak but noticeable negative temperature dependence to it. The higher the temperature, the lower the saturation velocity. The reason for this temperature dependence is not difficult to understand. The probability that an electron with energy higher than E_{opt} emits an optical phonon is not unity but it increases the higher the energy in excess of E_{opt}. At higher temperatures the electron energy distibution is more spread out and there are proportionally more electrons at high energies. As a result, at high temperatures, the probability of phonon emission increases as E increases over E_{opt}, bringing down with it the saturation velocity.

AT4.2.3 Impact ionization

If a carrier that is sufficiently hot collides with the lattice, it can donate the kinetic energy that it has acquired from the electric field in the generation of an electron-hole pair. This is the process of impact ionization briefly discussed in Ch. 3. Impact ionization can be viewed in two ways, as a generation mechanism as in Ch. 3, or as a scattering mechanism that modifies the momentum and the energy of the impact ionizing carrier. For most purposes, we are interested in the first picture since the impact generated carriers contribute to the total current which gets in consequence multiplied. The scattering picture is less relevant because phonon emission dominates the scattering rates of hot electrons at all energies but the highest ones.

Impact ionization is quantified by defining an *impact ionization coefficient*, α, as the average number of ionizations per unit length per carrier. It has the units of cm^{-1}. The inverse of the impact ionization coefficient can be thought of as the average distance traveled by a carrier in the

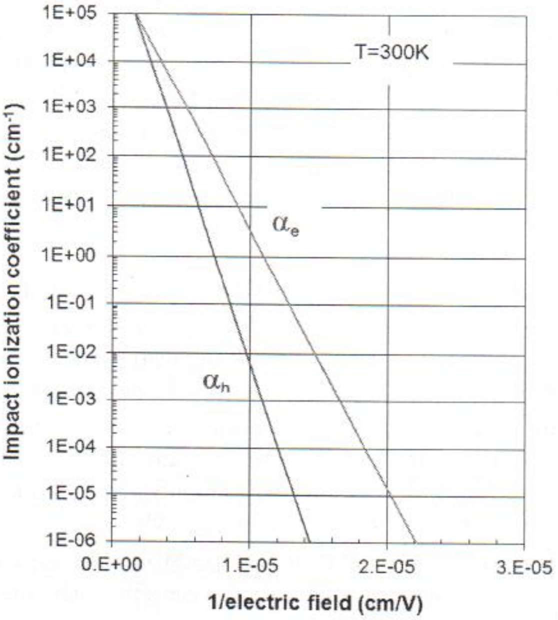


Figure 4.27: Impact ionization coefficient of electrons and holes as a function of electric field for Si at room temperature.

field direction between two impact ionizing collisions. α is a strong function of the electric field. Fig. 4.27 shows the room temperature impact ionization coefficients of electrons and holes in Si under a uniform electric field. This is a semilogarithmic plot as a function of the inverse electric field. This plot reveals that both α_e and α_h follow a law of the form:

$$\alpha \simeq A \exp(-\frac{B}{|\mathcal{E}|}) \tag{4.107}$$

with A and B two constants that depend slightly on the electric field. The values of A and B are given in Appendix E.

This functional relationship makes reasonable physical sense. In order for a carrier immersed in an electric field to produce an impact ionization event, it must first acquire sufficient energy from the electric field. This minimum value is known as the *impact ionization threshold energy*, E_{ii}. For a carrier to acquire this energy, it must at least travel a distance $l_{ii} = E_{ii}/e\mathcal{E}$ inside the electric field. The likelihood of a carrier traveling this distance without undergoing a single collision is $P(l_{ii}) = \exp(-l_{ii}/l_c)$, where l_c is the mean free path. In consequence, the ionization rate should go as:

$$\alpha \simeq \frac{1}{l_{ii}} \exp(-\frac{l_{ii}}{l_c}) = \frac{q|\mathcal{E}|}{E_{ii}} \exp(-\frac{E_{ii}}{q|\mathcal{E}|l_c}) \tag{4.108}$$

This equation, in spite of the simplifying assumptions that were made in its derivation, captures several key dependencies that are observed in experiments. First, it reveals the inverse exponential dependence on the electric field. Second, it shows an exponential dependence, with a minus sign, on the threshold energy E_{ii}. This is consistent with the observation that semiconductors with a wide bandgap and a higher value of E_{ii} exhibit lower values of α. Finally, the dependence on the mean free path in the expression is consistent with the observation that the impact ionization coefficient decreases with temperature. The higher the temperature, the shorter the mean free path and the harder it is for a carrier to acquire the required threshold energy.

The proper value of E_{ii} and its relationship with the bandgap of the semiconductor has been a matter of discussion among researchers in the field. This is not merely an academic issue. For a long time it was believed that $E_{ii} \sim 1.5E_g$. This suggested that using voltages below about 1.5 V or so in scaled-down Si MOSFETs would completely eliminate impact ionization which is a major reliability concern. Recently, impact ionization has been observed all the way down to sub-bandgap voltages. The reason is two-fold. On the one hand, the carrier distribution has a high energy tail. On the other hand, it is possible for the impact ionization event to coincide with a phonon collision. In this way, E_{ii} can be as low as E_g. The way to view E_{ii} is not so much as a hard threshold below which impact ionization is impossible, but a pragmatic threshold energy above which impact ionization becomes significant. Consistent with this approach, a value of E_{ii} for Si equal to the bandgap is sensible choice.

Impact ionization is a generation mechanism. When carriers drift in an electric field, some of the energy that they acquire from the electric field can be spent in generating electron-hole pairs. Naturally, the generation rate must depend on the carrier drift flux. In fact, it should be linear in the flux, that is, the higher the flux of carriers that drift in an electric field, the higher the generation rate due to impact ionization.

The relationship between the generation rate (in units of $cm^{-3} \cdot s^{-1}$) and drift flux (with units of $cm^{-2} \cdot s^{-1}$), is easy to derive. Let us do it for electrons. Looking back at Fig. 4.6, in an elemental time dt, electrons drift a distance $v_e^{drift}dt$, where v_e^{drift} is the electron drift velocity. The probability that an electron undergoes an impact ionization event in this distance is $\alpha_e v_e^{drift}dt$. If the concentration of electrons is n, the total number of generation events in time dt is $n\alpha_e v_e^{drift}dt$. Per unit time, then, the generation rate is simply $n\alpha_e v_e^{drift} = \alpha_e F_e(drift)$. Using similar arguments, it is easy to derive an expression for the generation rate due to impact ionization of holes. In both cases, the generation rate must be a positive number, regardless of the sign of the drift flux. The total generation rate due to impact ionization is then:

$$G_{ii} = \alpha_e|F_e(drift)| + \alpha_h|F_h(drift)| \tag{4.109}$$

where the absolute symbols are used to insure that G_{ii} is always positive independently of the choice that is made for the axis.

Eq. 4.109 is the basis for the treatment of carrier multiplication and avalanche breakdown that is presented in Ch. 5.