Chapter 5

Carrier Flow

Outline

In the preceeding chapters we have learned about a variety of physical mechanisms that take place in semiconductors: generation, recombination, drift and diffusion. In a semiconductor device under operation, several of these mechanisms can be in action simultaneously. For example, in the base of a bipolar transistor, minority carriers that are injected from the emitter diffuse towards the collector. Some of them may recombine in the base. If there is an electric field, minority carrier drift can also be significant through the base. A computation of the collector current of the transistor demands a correct description of all these processes and their interactions.

This chapter starts by formulating the set of equations that describe carrier behavior in semiconductors and the various boundary conditions that can be encountered. The governing system of equations is rather complex and analytical solutions are generally not obtainable. Computer-aided design (CAD) tools have been developed to solve these equations in semiconductor devices.

For the purpose of understanding device operation and for developing simple models that can be used in design, it is essential to simplify these equations as much as possible. Fortunately, this is feasible in many practical situations. We will distinguish three broad classes of situations: majority-carrier-type, minority-carrier-type and space-charge problems. In the first kind, majority-carrier drift and diffusion constitute the dominant phenomena. This is what happens, for example, in the source of a MOSFET. In the second kind, minority carrier behavior is the bottleneck. This is the situation in the base of a bipolar transistor. These two families of problems occur in quasi-neutral regions. In space-charge regions, such as in the depletion region of a p-n junction, the physics of carrier transport is somehow different and deserves special attention.

Minority-carrier-type problems can be mathematically difficult, particularly in dynamic situations. This is because of the subtle interplay of carrier diffusion and drift, bulk and surface generation or recombination. In many circumstances, device engineers do not need a complete solution to the problem. Identifying the limiting phenomena suffices to diagnose the situation and articulate a course of action. In this chapter, we solve a number of classic problems and use them to identify the key length scales and time scales for carrier behavior. Learning to do this correctly is a very valuable skill that is helpful in the design and analysis of microelectronic devices.

This chapter is, in some ways, about the management of complexity. In front of a given situation, the entire set of equations that completely describes carrier phenomena is usually a blunt instrument. It is often far more effective to analyze the physics of the problem, identify the bottlenecks, extract the essential terms from the equations, and discard all the rest. The simplified set is more likely to lead to a simple, analytical and physically intuitive solution of great practical value. The final step in this process is to verify the assumptions that were initially made to simplify the problem. This approach is fairly common in engineering. It is crucial in effective microelectronics device engineering. It will be discussed and illustrated in detail in this chapter in a few important situations.

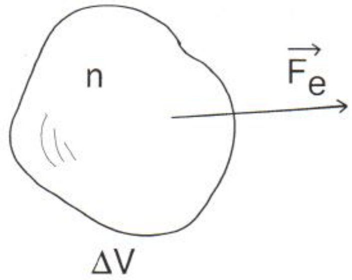


Figure 5.1: Sketch of small semiconductor volume ΔV with a uniform electron concentration n inside. An electron flux $\vec{F_e}$ crosses the surface of ΔV at a certain point.

5.1 Continuity Equations

In Ch. 3 we learned that in a certain region of a semiconductor, the carrier concentrations outside thermal equilibrium are determined by the imbalance between the rates of carrier generation and recombination in that region. Based on this understanding, we derived differential equations that allowed us to compute the time evolution of carrier concentrations in situations outside equilibrium (see Eq. 3.61). This picture is actually incomplete. It only applies to situations that are uniform, i.e., nothing changes in space. In more general situations, in addition to generation and recombination, there is a need to account for carrier flow in and out of the region of interest. The continuity equations that we derive in this section mathematically capture this simple fact.

A continuity equation is in essence a "book-keeping relationship" that keeps track of the number of particles in a region of a semiconductor as a function of time. Consider an elemental volume of a semiconductor ΔV, such as sketched in Fig. 5.1. The number of electrons, say, in this little volume increases with time if there is generation taking place inside it. Analogously, recombination reduces the electron count. Additionally, if there is a net flow of electrons out of that volume, whether by drift or diffusion, this also reduces the number of electrons in ΔV.

We can mathematically capture this by focusing on the *rates* at which these processes take place. Specifically, the rate of increase of the number of electrons in ΔV is equal to the rate of electron generation in ΔV, minus the rate of recombination in ΔV, minus the net flow of electrons leaving ΔV per unit time. If ΔV is small enough, this can be written mathematically as:

$$\frac{\partial(n\Delta V)}{\partial t} = G\Delta V - R\Delta V - \int \vec{F_e}.\vec{dS} \tag{5.1}$$

where the integral on the right-hand side sums all outgoing flux through the surface of ΔV. Dividing all terms by ΔV and taking the limit for ΔV very small, we get:

$$\frac{\partial n}{\partial t} = G - R - \vec{\nabla}.\vec{F_e} \tag{5.2}$$

This is the continuity equation for electrons. Following similar arguments, an identical equation is obtained for holes.

Since there is a direct relationship between carrier flux and current density (Eqs. 4.15 and 4.16), Eq. 5.2 can easily be rewriten as:

$$\frac{\partial n}{\partial t} = G - R + \frac{1}{q}\vec{\nabla}.\vec{J_e} \tag{5.3}$$

for electrons. For holes, we get:

$$\frac{\partial p}{\partial t} = G - R - \frac{1}{q}\vec{\nabla}.\vec{J_h} \tag{5.4}$$

These are the most fundamental expressions of the continuity equations for electrons and holes.

It is also of interest to focus on *continuity of charge*. We saw in Ch. 4 that in a semiconductor, net volume charge density results if there is an imbalance between the total concentrations of positive species (holes and donors) and negative species (electrons and acceptors) (Eq. 4.51). Since the impurity concentrations cannot change in time, the rate of change of volume charge density is given by the difference of the rate of change of hole and electron concentrations multiplied by the elemental charge:

$$\frac{\partial \rho}{\partial t} = q\frac{\partial (p-n)}{\partial t} \tag{5.5}$$

This observation allows us to derive a continuity equation for charge density. We multiply Eqs. 5.3 and 5.4 by q, we subtract one from the other and we substitute the result into Eq. 5.5 to get:

$$\frac{\partial \rho}{\partial t} = -\vec{\nabla}.\vec{J_t} \tag{5.6}$$

where J_t is the total current (sum of electron and hole current).

This equation states that if the volume charge density is changing with time in a region of a semiconductor, it is because there is net current flowing in or out of that region. For example, if the charge density at a certain location is increasing, that means that current is flowing into that region (negative divergence).

The continuity equation for charge is very useful to analyze a variety of important situations in semiconductors. A particularly relevant one that merits being discussed here is a fully *static situation* (though not necessarily in thermal equilibrium). This situation can result, for example, if there is net carrier generation in a region of a semiconductor as a result of external illumination that is steady in time. In a case like this in which nothing changes in time, there are no sinks or sources of net charge and at any location the divergence of the total current must be zero everywhere.

The continuity equations derived here are particularly intuitive when written in integral form. This is shown in Appendix AT5.1.

5.2 Surface continuity equations

A problem is not completely defined until its boundary conditions are specified. In semiconductor devices, boundary conditions often play a critical role in device operation. Many times "boundary condition engineering" is the most effective way to achieve certain device design goals. For the purpose of understanding carrier flow in semiconductor devices, we need to specify the appropriate boundary conditions that apply.

In describing a semiconductor device, the condition of all its surfaces must be specified. We are not equipped at this time to treat semiconductor surfaces rigorously. We will study the electrostatics of surfaces in detail later on in this book. Chapter 7, for example, deals with the metal-semiconductor interface and Chapter 8 discusses the insulator-semiconductor interface. For situations with excess carriers, we can make substantial progress in a number of important problems if we treat the surfaces in an empirical way. This is the approach followed in this section in which surface continuity equations are derived.

We discuss here the two simplest kinds of surfaces. First, we study a "free" surface, one in which the semiconductor is exposed or covered by a non-conductive material. Then we deal with "ohmic contacts" in which a metal is deposited on the surface. The fundamental difference between these two surfaces is that in the case of the free surface, carriers cannot escape from the semiconductor, while in the case of the ohmic contact, conduction is allowed between the metal and the semiconductor. The detailed physics of these two types of surfaces is fairly complex and a detailed analysis is left for later chapters.

5.2.1 Free surface

We denote a free surface as the boundary region of a semiconductor that is not electrically connected to anything else. As we will study in Ch. 8, at a surface there is an energy barrier that prevents carriers from "spilling out" of the semiconductor. A consequence of this is that the current normal to a free surface must be zero:

$$J_{ts} = 0 \tag{5.7}$$

This is illustrated in Fig. 5.2.

While the total electrical current is zero at a free surface, the electron and hole currents need not be zero separately. As discussed in Ch. 4, if the surface is not properly passivated, there can be significant recombination and generation associated with surface states and traps. We can characterize this by defining a surface recombination rate R_s and a surface generation rate G_s.

If there is net generation or recombination at a surface, there must be a flow of carriers into or out of it that come from or go to the rest of the semiconductor. The total current at the surface is zero but the electron and hole components, separately, are not. Similar to the continuity equations derived for the bulk of a semiconductor in Section 5.2, we can write a continuity equation for a surface for each type of carrier. A fundamental difference with the bulk case is that a surface

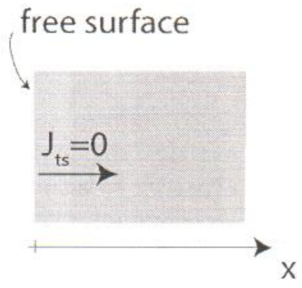


Figure 5.2: At a free surface, carriers cannot spill out and the net electrical current must be zero.

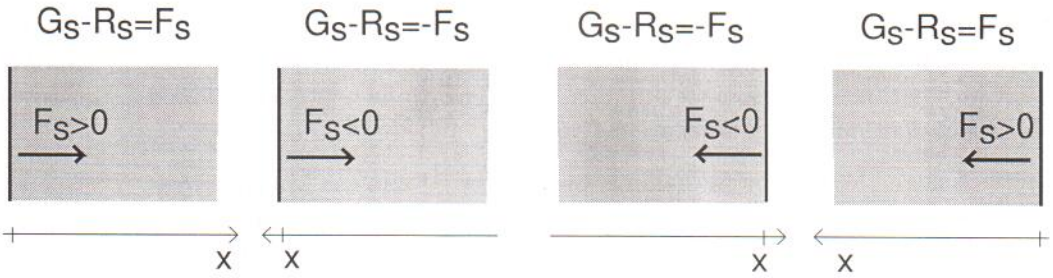


Figure 5.3: Depending on the choice of axis with respect with the location of the surface of a semiconductor, the surface continuity equation takes on different signs.

cannot store carriers since it has no volume. Particle conservation then demands that the rate of surface generation minus the rate of surface recombination, or *net surface generation rate*, equals the carrier flux out of the surface. Mathematically:

$$|G_s - R_s| = |F_s| \tag{5.8}$$

The units of all terms in this equation are $cm^{-2} \cdot s^{-1}$. The sign of F_s is sensitive to the choice of axis and it applies to both electrons and holes. This is illustrated in Fig. 5.3. The absolute symbols in this equation are to ensure that it applies regardless of the choice of axis.

We can rewrite Eq. 5.8 in a slightly more useful way. First, we can express the right-hand side in terms of current density. For this, we use the results obtained in the previous chapter in Eqs. 4.15-4.16. Second, the left-hand side is equal to the absolute of the net recombination rate at the surface, U_s, that was introduced in Ch. 3. With these two changes, Eq. 5.8 becomes:

$$|U_s| = \frac{1}{q}|J_{es}| = \frac{1}{q}|J_{hs}| \tag{5.9}$$

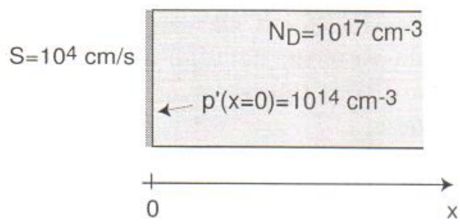
J_{es} and J_{hs} are the current densities normal to the surface. They are identical in magnitude because electrons and holes are generated and recombine in pairs. The sign for each current density in these two equations depends on the choice of axis. To keep the total current zero at the free surface, their signs must always be the opposite of each other.

In some problems, it is useful to introduce an *external* surface generation. This is of course a rather unphysical occurrence. However, depending on the length scale of the problem, it can be a handy simplifying approximation for some situations. Very short wavelength light, for example, is absorbed very effectively in a semiconductor. If all length scales of the problem of interest are much longer than the absorption length of the light, for all practical purposes the generation function can be assumed to be confined to a thin sheet at the surface. Including an external surface generation rate, Eq. 5.9 becomes:

$$|G_s(ext) - U_s| = \frac{1}{q}|J_{es}| = \frac{1}{q}|J_{hs}| \tag{5.10}$$

The units of $G_s(ext)$ are also $cm^{-2} \cdot s^{-1}$. Once again, the sign that must be selected for the current density terms depends on the choice of axis.

Exercise 5.1: *Consider the surface of an n-type Si sample with $N_D = 10^{17}\ cm^{-3}$ at 300K, as sketched below. This surface is characterized by a surface recombination velocity $S = 10^4\ cm/s$. At the surface, there is an excess hole concentration $p'(x=0) = 10^{14}\ cm^{-3}$.*



Under these conditions, calculate: a) the net recombination rate at the surface, b) the hole flux at the surface, c) the electron flux at the surface, d) the hole current density at the surface, e) the electron current density at the surface, and f) the total current density at the surface.

a) Since the semiconductor at the surface is under low-level injection conditions, the net recombination rate at the surface can be calculated using Eq. 3.72:

$$U_s = p'S = 10^{14} \times 10^4 = 10^{18}\ cm^{-2} \cdot s^{-1}$$

b) The magnitude of the hole flux at the surface is equal to the net recombination rate. Since the holes are flowing into the surface to recombine (against the choice of x in the figure above), the hole flux is negative. Then:

$$F_h(x=0) = -U_s = -10^{18}\ cm^{-2} \cdot s^{-1}$$

c) The electron flux at the surface is identical to the hole flux (every hole recombines with one electron):

$$F_e(x=0) = -10^{18}\ cm^{-2} \cdot s^{-1}$$

It is also negative because the electrons are also flowing into the surface.

d) The hole current density at the surface is simply:

$$J_h(x=0) = qF_h(x=0) = -1.6 \times 10^{-19} \times 10^{18} = -0.16\ A/cm^2$$

e) Similarly, the electron current density at the surface is:

$$J_e(x=0) = -qF_e(x=0) = 1.6 \times 10^{-19} \times 10^{18} = 0.16\ A/cm^2$$

f) The total current density at the surface is the sum of the electron and hole current densities:

$$J_t(x=0) = J_e(x=0) + J_h(x=0) = 0$$

This result makes sense since there cannot be net current into a free surface.

5.2.2 Ohmic contact

Implicit in the derivation of the surface continuity equations above is the constraint that carriers cannot "fall out" of the surface, nor be added somehow to the surface from the outside. This

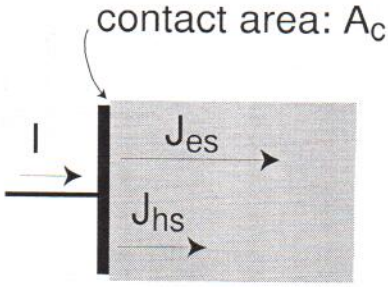


Figure 5.4: Sketch of ohmic contact between a metal and semiconductor with current flowing through. The area of the contact between the metal and the semiconductor is A_c.

assumption holds for a free surface. However, in just about all semiconductor devices, there are some surfaces that are contacted with metal so as to provide an electrical connection to the semiconductor from the outside world. These are called *ohmic contacts.*

The physics of the metal-semiconductor junction will be discussed in detail in Ch. 7, so fundamental aspects of the operation of ohmic contacts have to be left out for later on. At this time, we will have to accept some of its properties on faith until a proper justification is provided. Our goal here is to be able to write suitable continuity equations for ohmic contacts.

Consider the sketch of Fig. 5.4. It depicts an n-type semiconductor surface covered by a metal forming an ohmic contact (we will see in Ch. 7 that not all metal/semiconductor interfaces result in ohmic contacts). The metal surface is connected to a wire through which a current I flows. On the semiconductor side, right at the interface between the metal and the semiconductor, there can in general be an electron current J_{es}, and a hole current, J_{hs}. The ohmic contact has an area A_c.

Kirchoff's law applies at an ohmic contact. That is, current continuity demands that the current through the wire be identical to the sum of the electron and hole currents right at the surface on the semiconductor side. Mathematically,

$$|I| = A_c|J_{es} + J_{hs}| = qA_c|F_{es} - F_{hs}| \tag{5.11}$$

F_{es} and F_{hs} are the carrier fluxes at the semiconductor side of the metal-semiconductor interface. As before, the absolute symbols are required because the current signs on the semiconductor side depend on the choice of axis on the semiconductor. Note that there is an established sign convention for the current at a device contact. The current is considered positive if entering into the device and negative if coming out of the device. This is the convention followed in this book.

In addition to Eq. 5.11, there are a number of statements that can be made about ohmic contacts. The reason for these will be understood when we study the metal-semiconductor junction in Ch. 7.

- First, in ohmic contacts in the absence of net generation or recombination, the metal only communicates with the *majority carriers* in the semiconductor. This means that if there is

Figure 5.5: In the absence of net generation or recombination at an ohmic contact, terminal current is entirely supported by the majority carriers in the semiconductor (n-type on left, p-type on right).

current through the wire, this current is supported entirely by *majority carrier current* in the semiconductor. This is illustrated in Fig. 5.5. Mathematically:

$$I = A_c|J_{es}| = qA_cF_{es} \tag{5.12}$$

An equivalent equation applies for p-type material.

- If there is net generation or recombination at an ohmic contact (that is, at the metal-semiconductor interface), this results in a *minority carrier current.* Additionally, since electrons and holes are generated and recombine in pairs, this also produces an additional component to the majority carrier current of equal magnitude and contrary sign. This current component adds up to the one imposed from the outside, as mentioned above. Mathematically, for n-type material:

$$|U_s| = \frac{1}{q}|J_{hs}| \tag{5.13}$$

An equivalent equation applies for p-type material.

- Finally, because of the intimate contact between the metal and the semiconductor, an ohmic contact produces a surface with *infinite surface recombination velocity.* As a result, at an ohmic contact the minority carrier concentration is equal to its equilibrium value. The detailed reason for this will be better understood in Ch. 7. This is an important boundary condition that we will frequently encounter in this book. Mathematically,

$$n'_s = p'_s = 0 \tag{5.14}$$

To clarify the implications of these properties of ohmic contacts, we study three representative cases sketched in Fig. 5.6. In all three, there is current entering into an n-type semiconductor through the ohmic contact.

The case on the left represents an ohmic contact without any generation or recombination taking place at the interface. This is rather common in devices. It is for example the case of the contacts to the source and drain in MOSFETs, and the contact to the collector of a bipolar transistor in the forward active regime. In this case, there is no minority carrier current, and all the wire current is supported by electrons in the semiconductor. Hence:

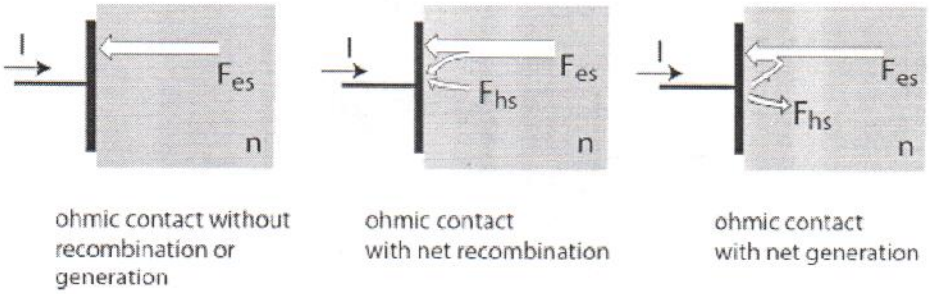


Figure 5.6: Three representative cases of ohmic contacts on n-type semiconductor in which there is current coming out of the semiconductor through the metal. Depending on whether there is net generation or recombination at the metal-semiconductor interface, the carrier fluxes have different signs and magnitudes.

$$U_s = 0 \tag{5.15}$$
$$I = qA_c|F_{es}| \tag{5.16}$$

The case in the center represents one in which there is an ohmic contact at which net recombination prevails. This happens for example in the contact to the emitter of a BJT in the forward active regime. For this case, the net recombination rate and Kirchoff's equation become:

$$U_s = |F_{hs}| \tag{5.17}$$
$$I = qA_c|F_{es} - F_{hs}| < qA_c|F_{es}| \tag{5.18}$$

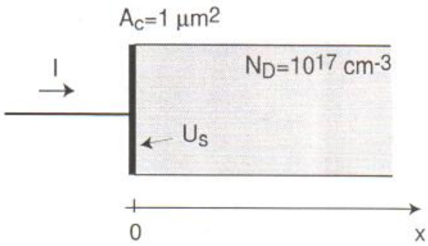
As the relative size of the arrows in Fig. 5.6 attempt to indicate, this is a case in which the majority carrier flux into the surface is increased with respect to the case on the left, because a few majority carriers flow in order to recombine with holes at the surface. The total current is continuous through the structure.

The case on the right represents a contact with net generation. This happens for example in the emitter of a BJT in the reverse regime. In this case, the net recombination at the surface is negative, and the majority carrier flux into the surface into the semiconductor is smaller than in the case on the left. The proper equations are:

$$U_s = -|F_{hs}| \tag{5.19}$$
$$I = qA_c|F_{es} - F_{hs}| > qA_c|F_{es}| \tag{5.20}$$

This understanding should allow us to formulate proper boundary conditions for ohmic contacts in all kinds of device settings.

Exercise 5.2: *Consider an ohmic contact to an n-type Si sample with $N_D = 10^{17}\ cm^{-3}$ at 300K, as sketched below. The area of the contact is $A_c = 1\ \mu m^2$.*



Current, I, can flow through the wire into the semiconductor and there is also the possibility of net recombination with a rate U_s at the metal/semiconductor interface. Calculate the electron and hole current densities at the semiconductor surface right below the contact under the following conditions: a) $I = 0$, $U_s = 10^{22}\ cm^{-2} \cdot s^{-1}$, b) $I = 10\ \mu A$, $U_s = 0$, and c) $I = 10\ \mu A$, $U_s = 10^{22}\ cm^{-2} \cdot s^{-1}$.

a) $I = 0$ is a case equivalent to a free surface. We know in this case that the magnitude of the electron and hole fluxes at the surface are equal to the net recombination rate. However, since there is net recombination at the surface, carriers flow *towards* it, which is against the choice of axis in the figure. Therefore, both F_e and F_h are negative:

$$F_e = F_h = -U_s = -10^{22}\ cm^{-2} \cdot s^{-1}$$

The current densities are, then:

$$J_e = -qF_e = 1.6\ kA/cm^2$$

$$J_h = qF_h = -1.6\ kA/cm^2$$

b) $U_s = 0$ implies that there is no minority carrier flux at the surface of the semiconductor. Hence, the hole current density is:

$$J_h = 0$$

The current coming into the contact from the wire is entirely supported by majority carriers on the semiconductor side. The electron current density is then:

$$J_e = \frac{I}{A_c} = \frac{10\ \mu A}{1\ \mu m^2} = 1\ kA/cm^2$$

This current is positive because it flows along x.

c) With U_s identical to case *a)* above, the hole current density is just as above:

$$J_h = -qU_s = -1.6\ kA/cm^2$$

The electron current density must be such that there is current continuity across the metal-semiconductor interface:

$$J_e = J_t - J_h = \frac{I}{A_c} - J_h = 1 + 1.6 = 2.6\ kA/cm^2$$

Gauss' law:	$\vec{\nabla} \cdot \vec{\mathcal{E}} = \frac{\rho}{\epsilon} = \frac{q}{\epsilon}(p - n + N_D^+ - N_A^-)$
Electron current equation:	$\vec{J_e} = -qn\vec{v_e}^{drift} + qD_e\vec{\nabla}n$
Hole current equation:	$\vec{J_h} = qp\vec{v_h}^{drift} - qD_h\vec{\nabla}p$
Electron continuity equation:	$\frac{\partial n}{\partial t} = G_{ext} - U + \frac{1}{q}\vec{\nabla} \cdot \vec{J_e}$
or	
Hole continuity equation:	$\frac{\partial p}{\partial t} = G_{ext} - U - \frac{1}{q}\vec{\nabla} \cdot \vec{J_h}$
Charge continuity equation:	$\frac{\partial \rho}{\partial t} = -\vec{\nabla} \cdot \vec{J_t}$
Total current equation:	$\vec{J_t} = \vec{J_e} + \vec{J_h}$

Table 5.1: Shockley equations. The continuity equation for charge is redundant with the electron and hole continuity equations. Only two of these three equations are independent.

5.3 Shockley equations

We now have a complete set of equations to study carrier flow in semiconductors under a great variety of circumstances. For convenience, Table 5.1 summarizes all these relationships in their most general form. They are commonly called Shockley equations, in honor of William Shockley, one of the inventors of the bipolar transistor.

Shockley equations are all very fundamental relations. Gauss' law states that net electrical charge of any kind produces an electric field. Next, the electron and hole current equations state that carriers in semiconductors can flow by means of drift and diffusion producing a current. Finally, the electron and hole continuity equations are particle book-keeping relationships. In Table 5.1 we have expressed the continuity equations in a slightly different way. The term $G - R$ in Eqs. 5.3 and 5.4 is rewritten in terms of an external generation rate G_{ext} and the net recombination rate U, as discussed in Ch. 3. We have also added to this set the charge continuity equation, even though it is redundant with the electron and hole continuity equations. Only two out of these three are independent. The last equation in the table states that the total current density at any point of a semiconductor is the sum of the electron and hole contributions.

This is a system of non-linear coupled partial differential equations which generally cannot be solved in a closed form. Commercial computer programs have been developed to solve this set of equations even in three dimensions. It is of great interest to obtain, wherever possible, analytical solutions that provide physical insight and that can be used in device design. A few important cases are amenable to compact solutions. We study them in the following sections.

5.4 Simplifications of Shockley equations to one-dimensional quasi-neutral situations

There are a variety of circumstances for which the Shockley equations can be substantially simplified and analytical results can be derived. Some of these situations are very important because they appear frequently even in fairly dissimilar devices. Recognizing these special cases allows understanding developed in the context of one device to port over to other devices.

The three special cases that we are going to discuss here are referred to as *minority-carrier situations*, *majority-carrier situations* and *space-charge situations*. These three different regimes of carrier transport appear very frequently in the operation of microelectronic devices. Even in a given device under a specific mode of operation, these three situations can be present simultaneously. This is illustrated in Fig. 5.7 which depicts an npn bipolar transistor in the forward-active regime. If we focus on electron transport across the intrinsic portion of the device, we recognize a minority-carrier situation, a space-charge region situation, and a majority-carrier situation in the flow of electrons through the base, the base-collector junction, and the collector, respectively. This will become clear after studying this chapter and the basic operation of the bipolar transistor in Ch. 11.

The rest of the chapter is devoted to discussing in some depth these three important situations. Our goal is to develop an intuitive understanding of carrier behavior in these three cases and to build a handy tool-set with which we will analyze several microelectronic devices in later chapters.

In order to accomplish this, we will perform a number of simplifications. A judgment of when a particular simplification is viable can only be acquired with experience. There will be multiple opportunities in this book to exercise such judgment. A chart summarizing the various simplifications that we are going to perform and their interrelations is shown in Fig. 5.8. We will refer to this chart in the sections that follow.

Although microelectronic devices are three-dimensional structures, in many situations we can simplify the problem to two and even one dimension. This happens because many times the geometry is very different in the three spatial directions, there are axis or planes of symmetry, or carrier phenomena are dominated by the behavior along one particular dimension. For example, in an integrated PN diode, as sketched in Fig. 5.9, the junction is parallel to the wafer surface and the lateral dimensions are typically much larger than the vertical ones. In this case, for a reasonably well designed diode under typical operating conditions, the dominant carrier behavior under the active area takes place along the vertical dimension (x in the diagram) with the physics not changing very much for the different positions on the plane. In these types of situations, a one-dimensional treatment of carrier flow can be highly accurate under most common circumstances [1].

Making the *one-dimensional approximation* drastically simplifies the Shockley equations. All $\vec{\nabla}$ operators become simple spatial derivatives along the selected dimension. These equations are

[1]Even in a geometry like the one sketched in Fig. 5.9, certain phenomena are intrinsically of a two- or three-dimensional nature. A good example is reverse-bias breakdown that is almost always associated with the corners of the PN junction, as we will study later on in this book.

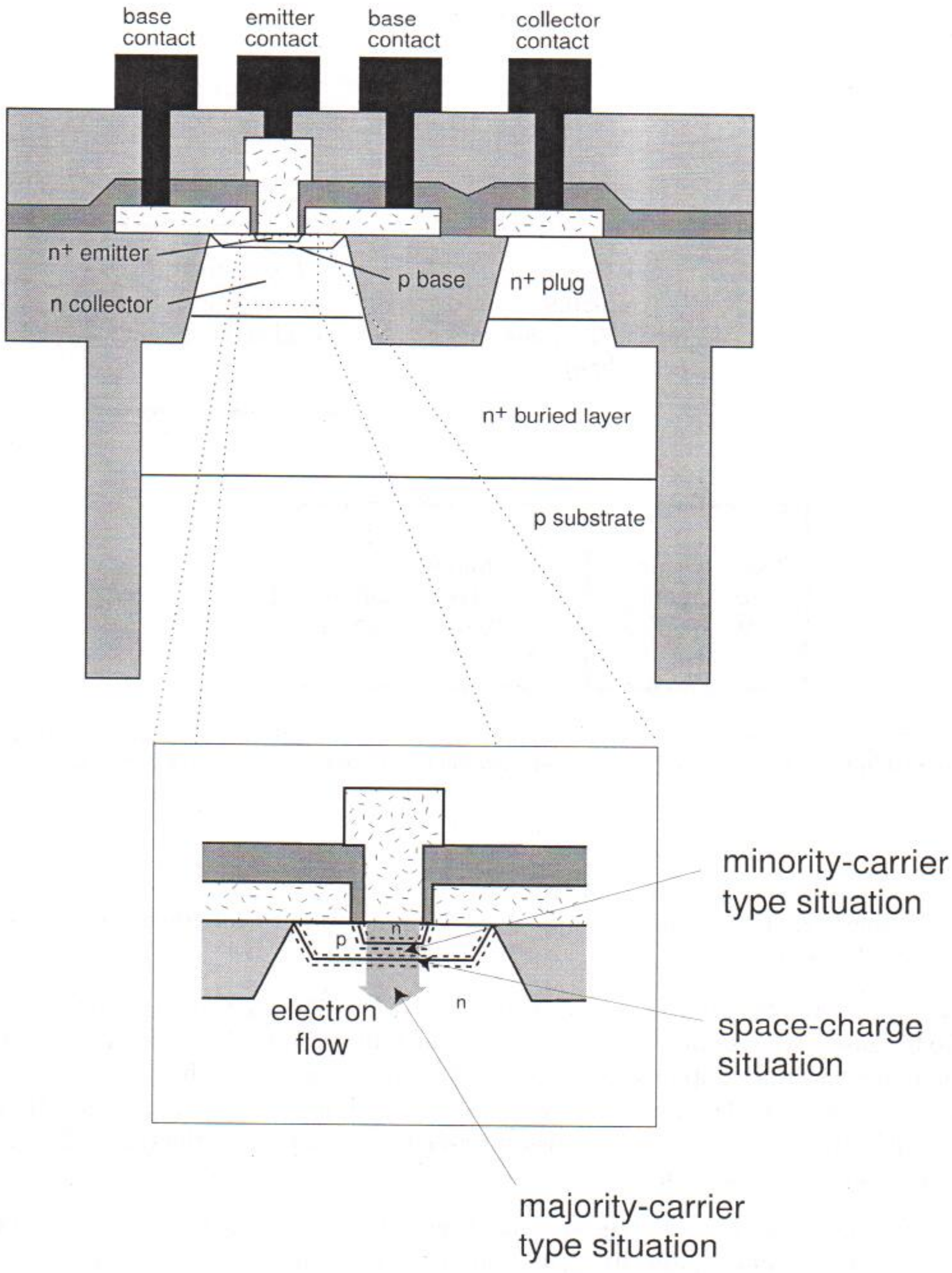


Figure 5.7: Cross section of modern bipolar junction transistor. The inset shows electron transport across the intrinsic portion of the device in the forward-active regime illustrating the different modes of transport that can be encountered.

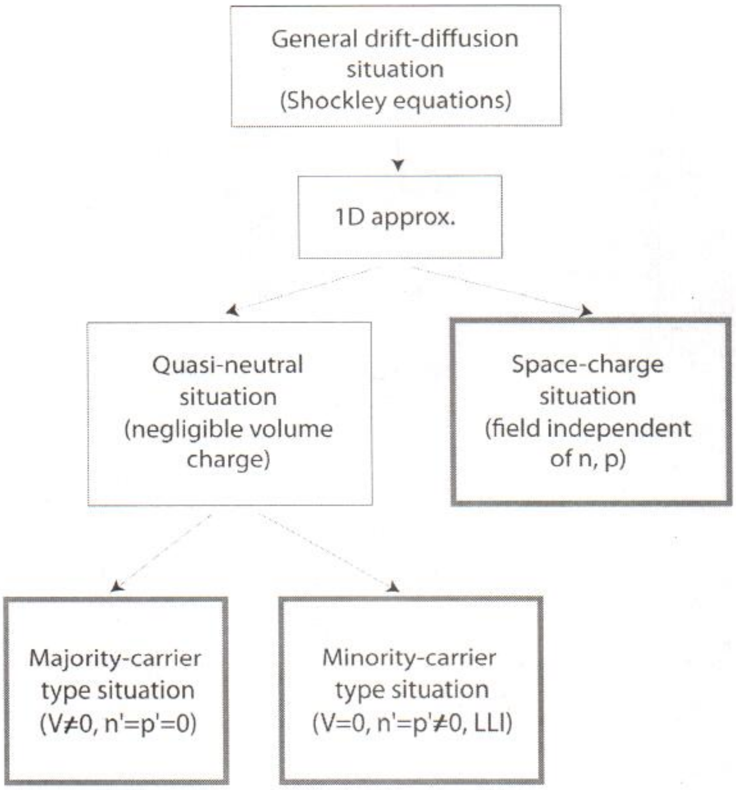


Figure 5.8: Chart summarizing the various approximations of the drift-diffusion formulation that are going to be performed in this Chapter. V refers to the application of an external voltage to a region in a semiconductor.

listed in Table 5.2. This is the first important simplification to the Shockley equations that we perform, as shown in Fig. 5.8.

The second simplification addresses Gauss' law. The Shockley equation set is a difficult one to solve because of the coupling that Gauss' law introduces. Fortunately, there are two important situations for which a simplification of this equation becomes possible, drastically untying the Shockley equation set. These two are quasi-neutral situations and space-charge situations. They are both indicated in Fig. 5.8. We discuss quasi-neutral situations in the remainder of this section and space-charge situations in Sect. 5.9.

We introduced the concept of quasi-neutrality in Ch. 4 when we discussed non-uniform doping distributions in thermal equilibrium. At that time, we defined a quasi-neutral situation as one in which the majority carrier concentration closely tracks the doping level with the consequence that the net volume charge density is negligible. We can generalize this concept to denote as quasi-neutral any situation in which *at every location the net volume charge density that arises from a discrepancy of the concentration of positive and negative species is negligible in the scale of the charge density present.* Mathematically, this implies that:

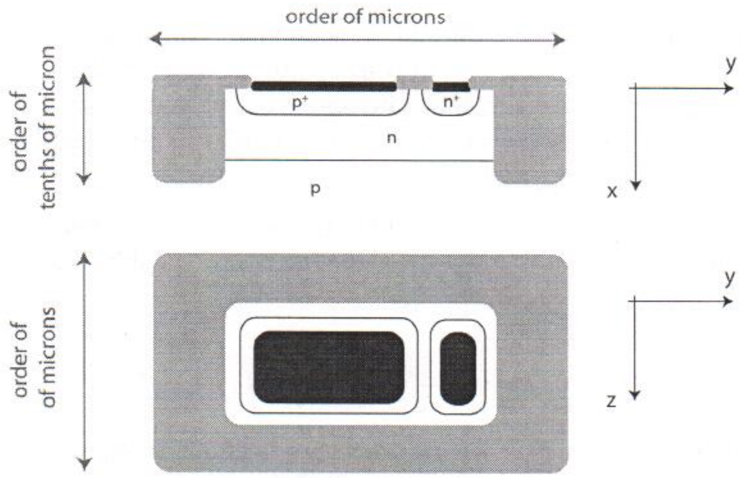


Figure 5.9: Cross-sectional sketch of a simple integrated p-n junction diode. The p-n junction lies parallel to the wafer surface. The extent of the device on the wafer is much larger than into the wafer. This justifies making one-dimensional approximation to study most aspects of device physics.

Gauss' law:	$\frac{\partial \mathcal{E}}{\partial x} = \frac{\rho}{\epsilon} = \frac{q}{\epsilon}(p - n + N_D - N_A)$
Electron current equation:	$J_e = -qnv_e^{drift}(\mathcal{E}) + qD_e\frac{\partial n}{\partial x}$
Hole current equation:	$J_h = qpv_h^{drift}(\mathcal{E}) - qD_h\frac{\partial p}{\partial x}$
Electron continuity equation:	$\frac{\partial n}{\partial t} = G_{ext} - U(n,p) + \frac{1}{q}\frac{\partial J_e}{\partial x}$
or	
Hole continuity equation:	$\frac{\partial p}{\partial t} = G_{ext} - U(n,p) - \frac{1}{q}\frac{\partial J_h}{\partial x}$
Charge continuity equation:	$\frac{\partial \rho}{\partial t} = -\frac{\partial J_t}{\partial x}$
Total current equation:	$J_t = J_e + J_h$

Table 5.2: Shockley equations in one dimension. The continuity equation for charge is redundant with the electron and hole continuity equations. Only two of these three equations are independent.

$$\rho \simeq 0 \tag{5.21}$$

The quasi-neutral approximation, a very reasonable one in many circumstances, has its origin in the fact that electrons and holes are mobile. In response to an electric field, carriers always move in the direction of trying to erase the field. In consequence, if there are large numbers of carriers, as is often the case in semiconductors, it is simply difficult to sustain a substantial volume charge for any appreciable time duration.

The quasi-neutrality assumption drastically simplifies the solution to the Shockley equations because it breaks the coupling of $\mathcal{E}$ with n and p through Gauss' law. Looking back to Eq. 4.51, ρ can be written as:

$$\rho = q(p - n + N_D^+ - N_A^-) = q(p_o - n_o + N_D^+ - N_A^-) + q(p' - n') \tag{5.22}$$

where the equilibrium and excess carriers concentrations have been made explicit.

In equilibrium, when $p' = n' = 0$, the quasi-neutrality condition can be expressed as:

$$|\frac{p_o - n_o + N_D^+ - N_A^-}{N_D^+ - N_A^-}| \ll 1 \tag{5.23}$$

which implies that

$$p_o - n_o \simeq -(N_D^+ - N_A^-) \tag{5.24}$$

Since $n_o p_o = n_i^2$, this equation yields right away the carrier concentrations in equilibrium. Eq. 5.24 is a more general way of expressing the result obtained in Eq. 4.75 in the previous chapter. We also studied in Ch. 4 how to assess when this condition applies.

If quasi-neutrality also holds outside equilibrium, we can then write that:

$$|\frac{p' - n'}{n'}| \simeq |\frac{p' - n'}{p'}| \ll 1 \tag{5.25}$$

which implies that:

$$p' \simeq n' \tag{5.26}$$

This equation represents a substantial simplification of non-equilibrium problems since in fact it all together eliminates one of the carrier concentrations as an unknown. The best way to check when quasi-neutrality applies outside equilibrium is to perform the approximation (Eq. 5.26), completely solve the problem, and then come back and check the condition (Eq. 5.25). We will illustrate this through an example later on in this chapter. From it, we will extract a simple handy guideline that states that quasi-neutrality holds when the characteristic length of the carrier flow problem is much longer than the Debye length.

A direct consequence of the quasi-neutrality approximation comes from its use in Gauss' law. This yields:

$$\frac{\partial \mathcal{E}}{\partial x} \simeq 0 \tag{5.27}$$

This result does not mean that the electric field is zero but rather that it cannot change too abruptly in space. In fact, in general, $\mathcal{E} \neq 0$. Actually, Eq. 5.22 in combination with Gauss' law, suggests that in general, the electric field can be expressed as:

$$\mathcal{E} = \mathcal{E}_o + \mathcal{E}' \tag{5.28}$$

Here $\mathcal{E}_o$ is the electric field in equilibrium which obeys the following equation:

$$\frac{\partial \mathcal{E}_o}{\partial x} = \frac{q}{\epsilon}(p_o - n_o + N_D^+ - N_A^-) \tag{5.29}$$

and $\mathcal{E}'$ is the *excess* electric field outside equilibrium which satisfies:

$$\frac{\partial \mathcal{E}'}{\partial x} = \frac{q}{\epsilon}(p' - n') \tag{5.30}$$

In a given problem, $\mathcal{E}_o$ is obtained as outlined in Ch. 4. In this chapter we will show how to solve for $\mathcal{E}'$ in quasi-neutral situations. Gauss' law then can be used to verify the validity of the quasi-neutrality assumption.

Under the quasi-neutrality assumption, if the volume charge density is negligible, its change in time can also be disregarded. This simplifies the charge continuity equation to:

$$\frac{\partial J_t}{\partial x} \simeq 0 \tag{5.31}$$

This is a simple statement of current continuity. Since the volume charge density cannot change in time, then the total current must be continuous everywhere. This will be very useful in device analysis because if we can figure out the total current at one location in the device, we know it everywhere.

With these changes, the quasi-neutral approximation simplifies the one-dimensional Shockley equations to the set of Table 5.3. To clean up the notation, it has further been assumed that all dopants are ionized.

We now split the discussion into two broad ways in which a semiconductor region can be brought out of equilibrium. These are both indicated in Fig. 5.8. One way is the application

$$p - n + N_D - N_A \simeq 0$$

$$J_e = -qnv_e^{drift} + qD_e\frac{\partial n}{\partial x}$$

$$J_h = qpv_h^{drift} - qD_h\frac{\partial p}{\partial x}$$

$$\frac{\partial n}{\partial t} = G_{ext} - U + \frac{1}{q}\frac{\partial J_e}{\partial x} \quad \text{or} \quad \frac{\partial p}{\partial t} = G_{ext} - U - \frac{1}{q}\frac{\partial J_h}{\partial x}$$

$$\frac{\partial J_t}{\partial x} \simeq 0$$

$$J_t = J_e + J_h$$

Table 5.3: Equation set for one-dimensional quasi-neutral situations.

of a voltage from the outside. The second one is the injection or generation of excess carriers. As we will see below, the relative role played by the majority and minority carriers in these situations is very different. Because of this, we refer to the first kind, application of voltage, as "majority-carrier situations," and to the second, introduction of excess carriers, as "minority-carrier situations." We study them separately. These are, of course, extreme cases of a rich continuum. In real devices, mixed situations often occur. However, to deal with them effectively, good understanding of the simple cases discussed here is essential.

5.5 Majority-carrier situations

This class of problems describes situations faced very frequently in microelectronic devices in which a voltage is applied to a semiconductor region from the outside without upsetting the carrier concentrations from their equilibrium values.

Before attacking a description of these problems, it is useful here to briefly remember what a battery does. If one connects a resistor across the two terminals of a battery, current flows. The standard notation is such that current flows out of the "long" terminal of the battery, through the resistor, into the "short" terminal of the battery, as sketched in Fig. 5.10. Current in metals and conventional resistors consists of electron flow. Since electrons are negatively charged, electrons actually flow in the contrary sense to the current, as also sketched in Fig. 5.10.

What drives the electrons to flow that way? The energy view of the situation, also sketched in Fig. 5.10, helps to understand this. The battery grabs electrons from its positive terminal and raises their energy as they cross through it. Electrons in the negative terminal have a higher energy than those in the positive terminal. If provided with a path, they will flow in an effort to lower their potential energy. The resistor is one such path (the battery itself is not, it blocks the "backward" flow of electrons). How much energy does the battery provide to the electrons? If the battery is rated with a voltage V, the energy of the negative terminal with respect to the

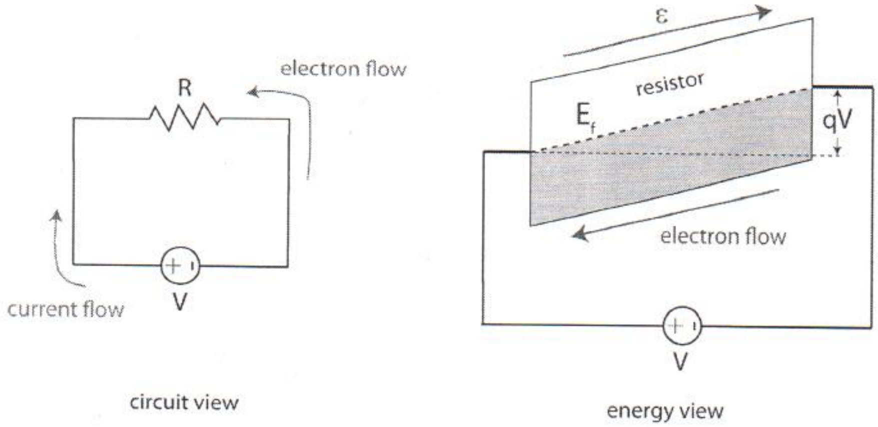


Figure 5.10: Left: simple circuit with resistor and battery. Right: energy view of circuit. The negative terminal of the battery raises the electron energy by qV with respect to the positive terminal. In consequence, electrons flow across the resistor from right to left.

positive terminal is qV. If there are no ohmic losses in the wires and the contacts, the entire voltage of the battery shows up across the resistor creating an electric field and tilting the energy band structure, as shown. This presents electrons in the resistor with empty states to their left and they preferentially flow from right to left. This is what is required by the sign of the battery.

Let us now discuss majority-carrier situations in which a voltage is applied to a semiconductor region. If the electric field that is introduced in the semiconductor is not too high, there is no reason for the dynamic balance that existed in thermal equilibrium between generation and recombination to be upset. This means that the carrier concentrations are not disturbed from their equilibrium values.[2] Under these circumstances, Shockley's equations simplify drastically. First of all, since the carrier concentrations are not upset from equilibrium, their time derivatives are zero. With $G_{ext} - U = 0$, this implies that $\frac{dJ_e}{dx} = \frac{dJ_h}{dx} = 0$.

We can also simplify the current equations in the following way. For electrons, for example, after substituting $n \simeq n_o$ in the equation listed in Table 5.3, we get:

$$J_e = -qn_o v_e^{drift}(\mathcal{E}) + qD_e \frac{dn_o}{dx} \tag{5.32}$$

where we have made explicit the electric field dependence of the electron drift velocity.

In thermal equilibrium, that is, when $\mathcal{E} = \mathcal{E}_o$, the electron current has to be zero. Eq. 5.32 becomes:

$$-qn_o v_e^{drift}(\mathcal{E}_o) + qD_e \frac{dn_o}{dx} = 0 \tag{5.33}$$

[2]If the electric field gets very high, carrier generation arising from impact ionization might invalidate this assumption. This is dealt with in Section 5.10.

n-type	p-type
$n \simeq n_o \simeq N_D$	$p \simeq p_o \simeq N_A$
$J_e = -qn_o[v_e^{drift}(\mathcal{E}) - v_e^{drift}(\mathcal{E}_o)]$	$J_h = qp_o[v_h^{drift}(\mathcal{E}) - v_h^{drift}(\mathcal{E}_o)]$
$\frac{dJ_e}{dx} \simeq 0,\ \frac{dJ_h}{dx} \simeq 0,\ \frac{dJ_t}{dx} \simeq 0$	
$J_t \simeq J_e$	$J_t \simeq J_h$

Table 5.4: Equation set for 1D majority-carrier type situations.

We can solve in this equation for the second term and insert it into Eq. 5.32 to get:

$$J_e = -qn_o[v_e^{drift}(\mathcal{E}) - v_e^{drift}(\mathcal{E}_o)] \tag{5.34}$$

A similar equation can be similarly derived for holes.

The low-field limit of this equation is of particular interest. For small fields, the drift velocity is directly proportional to the electric field. Using Eq. 5.28, we can easily rewrite Eq. 5.34 as:

$$J_e = qn_o\mu_e\mathcal{E}' \tag{5.35}$$

This equation clearly shows that outside equilibrium the electron current is driven by the excess field created by the application of a voltage on the sample. That is also the case for large fields. However, the non-linear velocity-field characteristics of carriers in semiconductors does not allow Eq. 5.34 to be written explicitly in terms of $\mathcal{E}'$.

An additional simplification that can readily be made is that, for reasonably extrinsic material, one of the carrier concentrations dominates over the other. Since, as Eq. 5.34 shows, the carrier currents are directly proportional to the carrier concentrations, in a typical majority carrier situation we can then confidently neglect the current contribution arising from the minority carriers.

This simplifies Shockley's equations to the set of Table 5.4. This equation set does not have time derivatives any more. Effectively, we are dealing with *quasi-static* situations in which the time evolution of the semiconductor is completely determined by the time dependence of the driving term (typically the applied voltage). The dynamics of majority carrier situations are discussed in more detail in Sec. 5.7.

5.5.1 Integrated resistor

The formulation that we have developed to describe majority-carrier situations is particularly useful in the analysis of the integrated resistor, a real device in its own right and an interesting

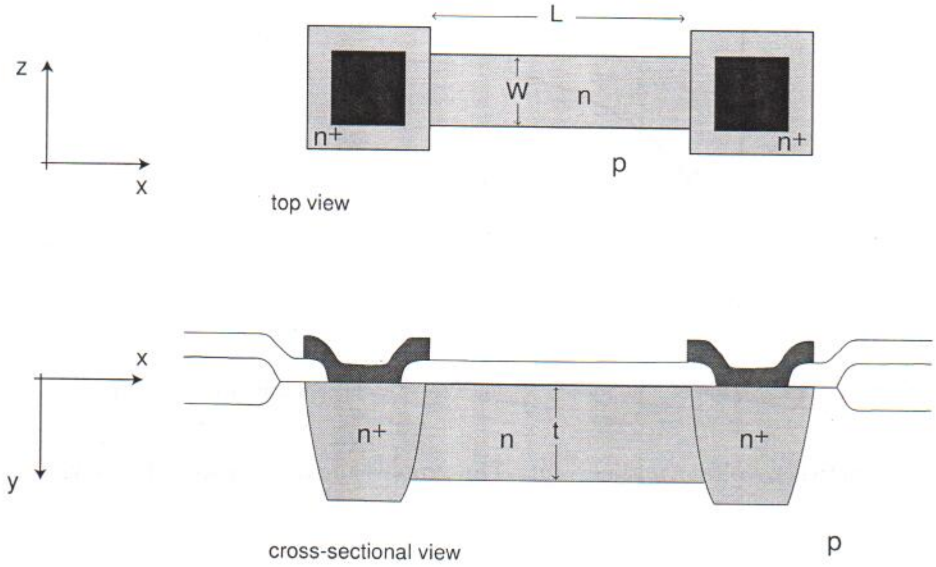


Figure 5.11: Sketch of integrated resistor.

case study for several important issues. A sketch of an integrated resistor is shown in Fig. 5.11. This device consists of an n-type region created inside a p-type wafer. The n-region is accessed from the outside world through two n^+ regions that are provided with ohmic contacts. Upon the application of a voltage across the contacts, current flows through the n-type region. As we will see in Ch. 6, the PN region that is formed between the n-type region and the substrate confines the current flow to the n-type region.

Let us consider a uniformly doped n-type semiconductor. In this case, $\mathcal{E}_o = 0$ and the current equation simplifies to:

$$J_t = -qN_D v_e^{drift}(\mathcal{E}) \tag{5.36}$$

If the voltage applied is not too high, the electron drift velocity is proportional to the electric field and:

$$J_t \simeq qN_D\mu_e\mathcal{E} \tag{5.37}$$

Solving for $\mathcal{E}$ and integrating, allows us to obtain the current-voltage relationship for this region:

$$I = \frac{qA\mu_e N_D}{L}V \tag{5.38}$$

where L is the length of the sample and A is its cross-sectional area. This equation shows a simple

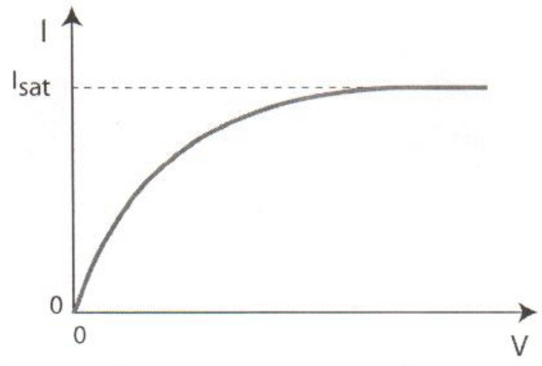


Figure 5.12: Sketch of non-linear I-V characteristics of semiconductor resistor.

linear dependence between voltage and current. The proportionality constant is the resistance of the sample:

$$R = \frac{L}{qA\mu_e N_D} \tag{5.39}$$

If the field is high enough, the linear relationship between electron velocity and electric field does not hold. Rather, a more complex relationship such as the one shown in Ch. 4 applies. Using Eq. 4.13 for the $v - \mathcal{E}$ relationship, solving for $\mathcal{E}$ and proceeding as above yields:

$$I = \frac{qAN_D v_{sat}}{1 + \frac{v_{sat}L}{\mu_e V}} \tag{5.40}$$

For low V this equation converges to Eq. 5.38 above. For moderate V, however, as V increases in Eq. 5.40, the current grows sublinearly with the voltage, or in other words, the resistance of the region increases. In fact, for high enough voltages, the electrons attain saturation velocity and I saturates to:

$$I_{sat} = qAN_D v_{sat} \tag{5.41}$$

and the resistance increases without bounds. The I-V characteristics captured in Eq. 5.40 are sketched in Fig. 5.12. Similar results apply for p-type regions.

In IC's, it is common to fabricate integrated resistors in which the doping level is not uniform in space but rather has a peak close to the surface and decreases in depth, as sketched in Fig. 5.13. In this case, we can easily compute the resistance by viewing this resistor as the parallel of many differential resistors. The resistance is then (for small voltages):

$$R = \frac{L}{qW \int_0^t \mu_e N_D(y)dy} \tag{5.42}$$

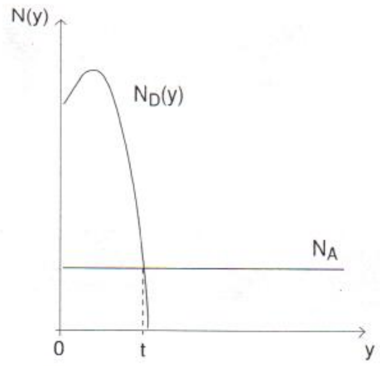


Figure 5.13: Typical doping profile of semiconductor resistor.The p-type background produces a p-n junction at the bottom of the resistor that effectively confines current flow to the n-type region.

This equation yields the result of Eq. 5.38 above if N_D is uniform in space. In general, if the doping level changes with position, the mobility will change too. That is why it has to remain inside the integral in Eq. 5.42.

When designing integrated resistors, device and circuit engineers utilize standard doped layers that are already available in the fabrication process, such as the n^+ or p^+ source implants in a CMOS process, or the n^+-emitter or p-base in an npn bipolar process. For the purpose of designing lateral resistors, the complexity of the doping distribution in depth can be hidden under a single parameter that is called the *sheet resistance*, defined as:

$$R_{sh} = \frac{1}{q\int_0^t \mu_e N_D(y)dy} \tag{5.43}$$

With this definition, the lateral resistance becomes

$$R = R_{sh}\frac{L}{W} \tag{5.44}$$

This is a useful equation that provides simple design rules for integrated resistors. The resistance of a planar resistor is obtained by multiplying the sheet resistance by the ratio of the length over the width of the resistor. In other words, the resistance depends only on the "number of squares" that the integrated resistor has in the direction of the current flow. This is illustrated in Fig. 5.14 where we show two resistors fabricated in the same process with identical resistance since both are made of five squares. Because of this interesting property, the units of the R_{sh} are frequently given in "ohms per square," written as $\Omega/\Box$.

For uniform doping distribution, the sheet resistance follows an expression:

$$R_{sh} = \frac{1}{q\mu_e N_D t} \tag{5.45}$$

The sheet resistance decreases as the doping level of the layer or its thickness increases.

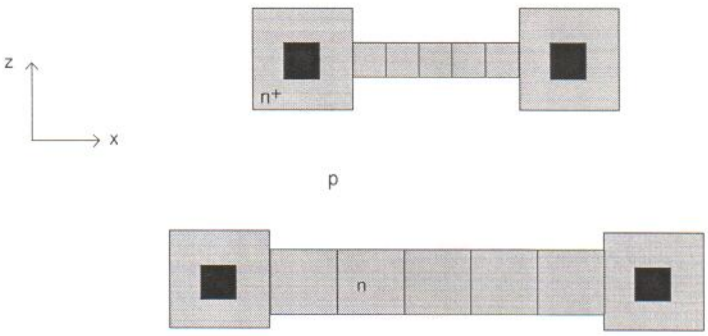


Figure 5.14: Two integrated resistors fabricated on the same process. The geometry is different but the resistance is the same (assuming that the parasitic resistance associated with the contacts is negligible). Coordinate x is in the direction of current flow. Coordinate y is into the semiconductor.

Exercise 5.3: *Consider an integrated resistor, such as the one shown in Fig. 5.11, composed of an n-type Si layer fabricated on a p-type substrate. The doping level of the body of the resistor is $N_D = 10^{19}\ cm^{-3}$. The dimensions of its active region are $L = 5\ \mu m$, $W = 2\ \mu m$, and $t = 0.1\ \mu m$. At room temperature, estimate the sheet resistance of the n-type semiconductor layer and the resistance of this resistor.*

For this doping level, the electron mobility is about $\mu_e \simeq 110\ cm^2/V.s$. The sheet resistance is then:

$$R_{sh} = \frac{1}{q\mu_e N_D t} = \frac{1}{1.6\times 10^{-19}\ C \times 110\ cm^2/V.s \times 10^{19}\ cm^{-3} \times 0.1\times 10^{-4}\ cm} \simeq 570\ \Omega/\square$$

The resistance of the resistor is then obtained from:

$$R = R_{sh}\frac{L}{W} = 570\ \Omega/\square \times \frac{5\ \mu m}{2\ \mu m} \simeq 1.5\ k\Omega$$

Note how in this last expression, there is no need to convert the units of L and W to cm since they cancel out.

5.6 Minority-carrier situations

We will now deal with a completely different class of problems that also frequently appear in microelectronic devices: minority-carrier situations. These are characterized by: *i)* quasi-neutrality, *ii)* the presence of excess carriers as a result of external generation or injection from an adjacent region, and *iii)* the absence of a significant external electric field. In our chart depicting the successive approximations to the Schockley equations (Fig. 5.8), these situations are shown in the box at the bottom center. Situations of this kind arise, for example, in the base of a bipolar transistor under regular bias conditions. In general, a description of these situations requires the complete set of equations of Table 5.3. Fortunately, two important approximations hold under a wide range of circumstances.

The first approximation accounts for the fact that in the absence of external electric fields, the internal electric field that might be present in equilibrium is rarely so high that velocity saturation is an issue. This allows us to assume that the drift velocity is linearly proportional to the electric field.

The second approximation is *low-level injection*. This was already discussed in Ch. 3 in the context of generation and recombination. As a reminder, low-level injection refers to a situation in which the excess carrier concentration is much higher than the equilibrium minority carrier concentration but much smaller than the equilibrium majority carrier concentration. This situation is fairly common because minority-carrier devices exhibit degraded performance if the carrier concentrations reach high-injection levels. In many devices, typical operating conditions avoid high-level injection.

The restriction to low injection levels results in several simplifications:

- First, the majority carrier concentrations out of equilibrium are basically unchanged from their equilibrium values. Thus, for n-type material, $n \simeq n_o$.
- Second, the minority carrier concentrations are overwhelmed. In an n-type semiconductor, this means $p \simeq p'$.
- Third, the recombination rate is proportional to the excess carrier concentration over the carrier lifetime (see Ch. 3). For n-type, we can write $U \simeq p'/\tau$.
- Even though no external fields are applied, internal fields can be generated as a result of carrier injection. However, and this is our fourth simplification, internally generated electric fields are small enough so that minority carrier drift currents produced by them are insignificant (we will later prove this point in one of the examples). The only fields that might significantly act on the minority carriers are those that are present internally under equilibrium conditions as a result of doping gradients. We cannot extend this approximation to the majority carriers because, having so many of them around, small changes in the electric field can produce substantial changes in the current.

These simplifications allow us to rewrite the carrier current equations and the continuity equations in the following manner. For n-type material, for example, the majority carrier current can be written in the following way:

$$J_e \simeq q(n_o + n')\mu_e(\mathcal{E}_o + \mathcal{E}') + qD_e(\frac{\partial n_o}{\partial x} + \frac{\partial n'}{\partial x}) \tag{5.46}$$

where we have made explicit the excess electron concentration and the excess electric field produced outside equilibrium.

In equilibrium, that is, with $n' = 0$ and $\mathcal{E}' = 0$, the electron current has to be zero. This allows us to simplify Eq. 5.46 to:

$$J_e \simeq qn_o\mu_e\mathcal{E}' + qn'\mu_e\mathcal{E}_o + qD_e\frac{\partial n'}{\partial x} \tag{5.47}$$

n-type	p-type
$p_o - n_o + N_D - N_A \simeq 0$	
$p' \simeq n'$	
$J_e = qn_o\mu_e\mathcal{E}' + qn'\mu_e\mathcal{E}_o + qD_e\frac{\partial n'}{\partial x}$	$J_e = qn'\mu_e\mathcal{E}_o + qD_e\frac{\partial n'}{\partial x}$
$J_h = qp'\mu_h\mathcal{E}_o - qD_h\frac{\partial p'}{\partial x}$	$J_h = qp_o\mu_h\mathcal{E}' + qp'\mu_h\mathcal{E}_o - qD_h\frac{\partial p'}{\partial x}$
$D_h\frac{\partial^2 p'}{\partial x^2} - \mu_h\mathcal{E}_o\frac{\partial p'}{\partial x} - \frac{p'}{\tau} + G_{ext} = \frac{\partial p'}{\partial t}$	$D_e\frac{\partial^2 n'}{\partial x^2} + \mu_e\mathcal{E}_o\frac{\partial n'}{\partial x} - \frac{n'}{\tau} + G_{ext} = \frac{\partial n'}{\partial t}$
$\frac{\partial J_t}{\partial x} \simeq 0$	
$J_t = J_e + J_h$	

Table 5.5: Equation set for one-dimensional minority-carrier situations.

There are two drift terms involving electrons (as majority carriers) out of equilibrium. One is due to the excess electric field acting on the equilibrium electron concentration. The second one is the equilibrium electric field acting on the excess electron concentration. Both terms can be important. There is also one diffusion term that accounts for diffusion of excess electrons.

The expression of the minority carrier current can also be substantially simplified:

$$J_h \simeq q(p_o + p')\mu_h(\mathcal{E}_o + \mathcal{E}') - qD_h(\frac{\partial p_o}{\partial x} + \frac{\partial p'}{\partial x}) \tag{5.48}$$

Once again, we can simplify this expression by noting that in equilibrium $J_h = 0$. Eq. 5.48 becomes:

$$J_h \simeq qp'\mu_h\mathcal{E}_o + qp'\mu_h\mathcal{E}' - qD_h\frac{\partial p'}{\partial x} \simeq qp'\mu_h\mathcal{E}_o - qD_h\frac{\partial p'}{\partial x} \tag{5.49}$$

The term on $qp'\mu_h\mathcal{E}'$ has been neglected as discussed above. This expression of the minority carrier current now depends exclusively on the excess minority carrier concentration.

The minority carrier continuity equation is also modified to:

$$\frac{\partial p'}{\partial t} = G_{ext} - \frac{p'}{\tau} - \frac{1}{q}\frac{\partial J_h}{\partial x} \tag{5.50}$$

We can further transform this equation into a particularly useful form by substituting Eq. 5.49 into Eq. 5.50. Rearranging terms, we easily get:

$$D_h \frac{\partial^2 p'}{\partial x^2} - \mu_h \mathcal{E}_o \frac{\partial p'}{\partial x} - \frac{p'}{\tau} + G_{ext} = \frac{\partial p'}{\partial t} \tag{5.51}$$

where we have used the fact that $\partial \mathcal{E}_o / \partial x \simeq 0$ in a quasi-neutral region (see Section 4.5.3).

Eq. 5.51 is now a differential equation with a single unknown, the excess minority carrier concentration. Once the generation function G_{ext} and the boundary conditions are specified, this equation can (in principle) be solved and p' can be obtained throughout the region of interest. From here, the complete solution quickly emerges, as we will see below.

It is interesting to see that if we follow a similar procedure with the majority carriers, that is, we introduce the majority carrier current Eq. 5.47 into its continuity equation, an equation solely on the majority carrier concentration similar to 5.51 does not emerge. There is a term on $\mathcal{E}'$ that prevents this equation from being solved all by itself. Eq. 5.51 then shows that in minority-carrier situations, the dynamics of the minority carriers dominate the solution and everything else revolves around this.

To summarize, under the assumptions of low-level injection and no external field applied, the Shockley equations simplify as listed in Table 5.5. The best way to understand the nature of the approximations made in this section is to study some specific examples. We will consider two of them in some detail. Both are static. We will examine a dynamic case later on in Section 5.8.

5.6.1 Example 1: Diffusion and bulk recombination in a "long" bar

Let us consider the situation sketched in Fig. 5.15. We have a very long uniformly-doped n-type semiconductor bar that is illuminated over a very narrow region with deep penetrating radiation. At this cross section, g_l electron-hole pairs per second per unit area are generated. We wish to compute the electron and hole concentrations and the respective currents throughout under static conditions. Since the doping level is uniform, the electric field inside this sample in thermal equilibrium is zero, that is, $\mathcal{E}_o = 0$.

Before attempting a complete mathematical solution to the problem, it is valuable to discuss qualitatively what is going on. Let us place the origin of the axis at the point of generation. Electron-hole pair generation at $x = 0$ causes a build up of carrier concentration above the equilibrium background. In consequence, a gradient in the carrier concentration develops which results in carrier diffusion away from the generation point. As excess carriers spill into the neighboring regions of the semiconductor the recombination rate increases above the thermal equilibrium rate. Eventually, a steady state situation is attained in which the total generation rate matches the total recombination rate. The same principle that we have encountered a number of times before is at play here. When thermal equilibrium is perturbed, the semiconductor reacts trying to reestablish it. In this case, as the carrier concentration builds up, the recombination rate increases in an effort to wipe out the excess carriers.

A photogenerated hole has a lifetime that is set by the material, as we discussed in Ch. 3. This means that it will only have a chance to diffuse a certain distance before it recombines. This makes us expect a steady-state carrier profile with a shape as sketched in Fig. 5.15 in which

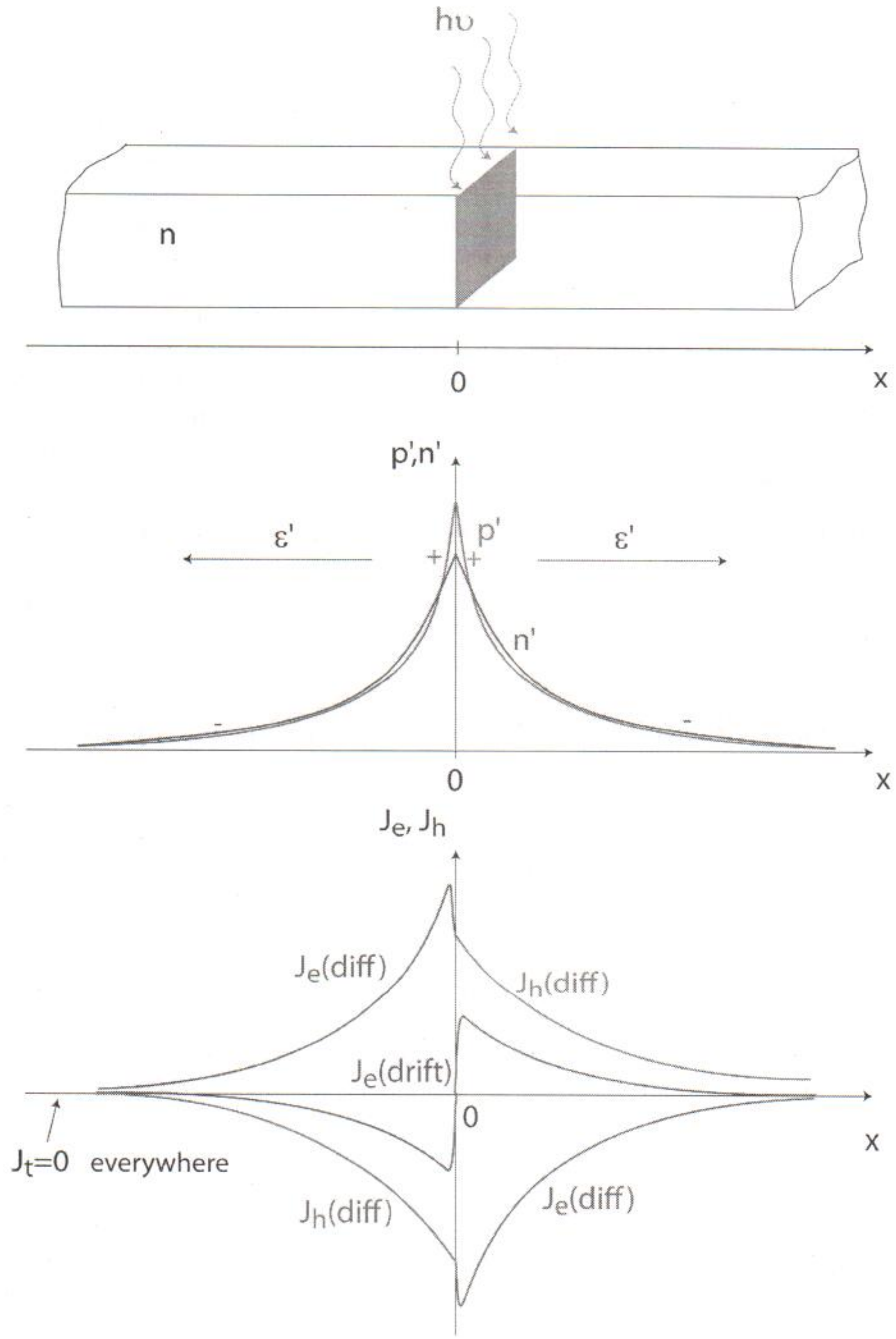


Figure 5.15: Top: an infinitely long bar illuminated at $x = 0$ by a sheet of light; middle: excess electron and hole concentrations; bottom: electron and hole current components.

the carrier concentration peaks at the generation point and decays to the equilibrium value at a certain distance. Far enough away from the point of generation, the semiconductor will not be aware that the region around $x = 0$ is being perturbed from equilibrium.

It is interesting to think about what happens to the photogenerated electrons. It is tempting to say that the photogenerated electrons diffuse away from the generation point, just as the photogenerated holes do. This is, however, not correct. The moment an electron is generated, it cannot be distinguished from any other one of the many majority electrons that were in the bar to begin with. The best way to think about what happens to the electrons is to realize that, by itself, the hole buildup close to the generation point would drive the bar outside charge neutrality. The electric field that would result would get the electrons moving immediately in the direction of attempting to erase the drift field. This can be accomplished if the excess electron concentration precisely matches the hole electron concentration at every point. As we see next, however, this is not a sustainable situation.

If the semiconductor bar is reasonably extrinsic, we know that quasi-neutrality has to prevail. We concluded earlier in this chapter that this makes the total current constant at any cross section in the structure. Since far enough away from the generation point the semiconductor is not upset from equilibrium, the total current there is precisely zero. Hence $J_t = 0$ everywhere. Closer to $x = 0$, in the region invaded by excess carriers, the sum of the electron and hole currents has to cancel out. Summing Eqs. 5.47 and 5.49 and noting that $\mathcal{E}_o = 0$, we have:

$$0 = J_t \simeq q n_o \mu_e \mathcal{E}' + q(D_e - D_h)\frac{dp'}{dx} \tag{5.52}$$

This equation suggests that if $D_e = D_h$ precisely, the electric field is exactly zero and there is no drift current of any kind. But, since in most semiconductors $D_e \neq D_h$, the diffusion currents do not cancel out. Hence J_t cannot vanish. We therefore need a drift current to cancel the imbalance of the diffusion currents and provide $J_t = 0$ everywhere.

A null total current everywhere demands that p' be close but not exactly equal to n'. A small deviation from perfect charge neutrality is all that is needed to provide the necessary electric field that produces the required drift current. For $x > 0$, since $D_e > D_h$ and $\frac{dp'}{dx} < 0$, the diffusion term in Eq. 5.52 is negative. Hence, the drift term must be positive, which demands a positive electric field. This is established by having $n' < p'$ in the vicinity of $x = 0$, and $n' > p'$ far away from $x = 0$, as sketched in Fig. 5.15.

We can also conclude from this qualitative analysis that both $\mathcal{E}'$ as well as $|n' - p'|$, must be proportional to $D_e - D_h$. This is because it is the difference between D_e and D_h that drives the need for an electric field.

Having discussed the physics of this example in an intuitive way, let us now solve it completely. The symmetry of the situation simplifies our study to $x \geq 0$. The place to start is the minority carrier continuity equation of Table 5.5. In this uniformly doped bar, the initial electric field is zero and there is no applied field from the outside. As already argued above, the field term in the minority carrier continuity equation is negligible. Also in steady state, we can neglect the time derivative. Furthermore, with the exception of $x = 0$ (which we will treat below as a boundary condition), the bar is in the dark. The minority carrier continuity equation everywhere but at the origin becomes:

$$\frac{d^2p'}{dx^2} - \frac{p'}{L_h^2} = 0 \tag{5.53}$$

where we have defined L_h as:

$$L_h = \sqrt{D_h \tau} \tag{5.54}$$

L_h is called the hole diffusion length. Its significance will become clear very soon.

The solution of the differential equation 5.53 can take several mathematical forms. A suitable one in this example is the following:

$$p' = A \exp \frac{x}{L_h} + B \exp \frac{-x}{L_h} \tag{5.55}$$

This solution consists of a rising exponential and a decaying exponential in x. Since we know that eventually far away from the generation point the excess hole concentration disappears, A must be equal to zero.

We get B from considering the boundary condition at $x = 0$. A particularly easy way to think about it is to exploit the symmetry of the problem. At $x = 0$ there is a sheet of generation of holes at a rate g_l. Half of the generated holes diffuse to the right, and the other half diffuse to the left. Hence, at $x = 0^+$,

$$\frac{g_l}{2} = \frac{1}{q} J_h(0^+) = -D_h \frac{dp'}{dx}|_{x=0^+} \tag{5.56}$$

Combining this equation with Eq. 5.55 (with $A = 0$), we easily get $B = g_l L_h/(2D_h)$ (this expression of B has units of cm^{-3}).

The excess hole concentration is then:

$$p' = \frac{g_l L_h}{2D_h} \exp \frac{-x}{L_h} \tag{5.57}$$

This has the shape that was qualitatively predicted above. Eq. 5.57 applies everywhere to the right of $x = 0$ and it also applies to $x = 0$ because p' cannot be discontinuous.

The hole current everywhere can now be easily calculated:

$$J_h = -qD_h \frac{dp'}{dx} = \frac{qg_l}{2} \exp \frac{-x}{L_h} \tag{5.58}$$

where we have assumed that $J_h(drift) \ll J_h(diff)$.

The electron current can be found by first thinking of the total current J_t. As argued above, $J_t = 0$ everywhere. This means that the electron current is:

$$J_e = -J_h = -\frac{qg_l}{2} \exp \frac{-x}{L_h} \tag{5.59}$$

In the electron current, we must consider both the drift and the diffusion components. The diffusion component is easy to obtain since the quasi-neutrality requirement implies that:

$$n' \simeq p' = \frac{g_l L_h}{2D_h} \exp \frac{-x}{L_h} \tag{5.60}$$

and

$$J_e(diff) = qD_e \frac{dn'}{dx} = -\frac{qg_l}{2} \frac{D_e}{D_h} \exp \frac{-x}{L_h} \tag{5.61}$$

The drift component is simply the balance of the electron current minus the diffusion component:

$$J_e(drift) = J_e - J_e(diff) = \frac{qg_l}{2} \frac{D_e - D_h}{D_h} \exp \frac{-x}{L_h} \tag{5.62}$$

This is the result that was expected, that is, an electron drift current flows to the extent that D_e is different from D_h.

At this point, we have obtained expressions for the carrier and current distributions everywhere and the problem is completely solved. Still, we need to discuss the accuracy of the assumptions that were made and understand the constraints under which they work.

Let us first check the quasi-neutrality assumption. This requires obtaining the electric field distribution. From the electron drift current expression, we can easily get:

$$\mathcal{E}' = \frac{J_e(drift)}{q\mu_e n_o} = \frac{kT}{q} \frac{g_l}{2n_o} \frac{D_e - D_h}{D_e D_h} \exp \frac{-x}{L_h} \tag{5.63}$$

Using this expression of the electric field in Gauss' law, we can obtain the difference between n' and p':

$$p' - n' = -\frac{\epsilon kT}{q^2 n_o} \frac{g_l}{2L_h} \frac{D_e - D_h}{D_e D_h} \exp \frac{-x}{L_h} \tag{5.64}$$

which together with Eq. 5.60 can now be used to verify the quasi-neutrality condition:

$$|\frac{p' - n'}{p'}| = (\frac{L_D}{L_h})^2 \frac{D_e - D_h}{D_e} \tag{5.65}$$

where L_D is the extrinsic Debye length defined in Ch. 4. Let us look at the order of magnitude of the various terms in this equation. The term involving the diffusion coefficients is slightly smaller than unity. The term involving the two length scales is many times smaller than one. For example, in Si with $N_D = 10^{16}\ cm^{-3}$ at room temperature the hole diffusion length is 400 μm, while L_D is about 0.04 μm. Clearly, the difference between n' and p' is about 8 orders of magnitude smaller than the sum of their values! The quasi-neutrality assumption is indeed extremely good.

We can now check our assumption of neglecting the drift contribution to the minority carrier current. Let's compute the relative magnitude of the hole drift current to the hole diffusion current:

$$|\frac{J_h(drift)}{J_h(diff)}| = |\frac{q\mu_h p' \mathcal{E}'}{-qD_h \frac{dp'}{dx}}| = \frac{1}{2}\frac{p'}{n_o}\frac{D_e - D_h}{D_e} \tag{5.66}$$

This equation says that this approximation is as good as the low-level injection condition. If the low-level injection condition is satisfied with a margin of ten ($p' \simeq 0.1n_o$), then the assumption that the drift contribution to the minority carrier current can be neglected is good to about one part in ten.

It is easy to see why these two conditions come hand in hand. As already argued, an electric field is required to the extent that D_e is different from D_h. In consequence, the resulting majority carrier drift current has an order of magnitude similar to the minority carrier diffusion current. All together, this implies that the relative magnitude of the minority carrier drift current with respect to the minority carrier diffusion current is about the same as the relative magnitude of p' with respect to n_o.

It is easy to calculate the maximum generation rate that satisfies the low-level injection condition. The worst location in the semiconductor is $x = 0$. For low-level injection to apply there, we need to demand that $p'(0) \ll n_o$. This implies that

$$g_l \ll \frac{2D_h n_o}{L_h} \tag{5.67}$$

A margin of safety of 10 in this inequality gives an error of about 10% in the computation of the current. This is sufficient for many applications.

We must also check the assumption that the drift field that is set is not so large that we need to be concerned with velocity saturation effects in the computation of the majority carrier current. The worst point is $x = 0$ where $\mathcal{E}'$ in Eq. 5.63 is maximum. At this point, using Eq. 5.63:

$$\mu_e \mathcal{E}' = \frac{g_l}{2n_o}\frac{D_e - D_h}{D_h} \ll \frac{D_e - D_h}{L_h} \tag{5.68}$$

where we have used the low-level injection condition Eq. 5.67. Typical values for the right hand side of Eq. 5.68 are of the order of 500 cm/s, much smaller than the electron saturation velocity of 1×10^7 cm/s.

Two further comments before closing this example. First, a word about the physical meaning of the minority carrier diffusion length. The diffusion length is the average distance that a carrier travels by diffusion before recombining. In three diffusion lengths, for example, 95% of the excess carriers have recombined. The diffusion length is a very important length scale in problems in which minority carrier behavior plays a key role. More about this in Section 5.6.3.

Second is a simple calculation of the average velocity at which holes diffuse in the semiconductor bar. This is easy to obtain. From Eqs. 5.57 and 5.58, we have for $x > 0$:

$$v_h^{diff} = \frac{J_h^{diff}(x)}{qp(x)} \simeq \frac{J_h(x)}{qp'(x)} = \frac{D_h}{L_h} \tag{5.69}$$

which is independent of position. This is an important result. We will use it in the computation of the forward bias current in a pn diode.

5.6.2 Example 2: Diffusion and surface recombination in a "short" bar

Let us now consider another uniformly-doped n-type bar illuminated by deep penetrating radiation over a narrow region at its center so that g_l electron-hole pairs are produced per unit second per unit area. In this case, the bar is of finite length L and it has end surfaces with infinite surface recombination velocity (Fig. 5.16). The length of the bar is much shorter than the diffusion length so that bulk recombination is very small. This situations is oftentimes referred to as "transparent." As in the above example, we seek to solve for the carrier concentrations and current densities everywhere in steady state.

Qualitatively we can see that this problem is similar to the previous one. The photogenerated holes diffuse away from the generation point at $x = 0$, half to each side. The important difference with the previous case is that before the holes have a chance to recombine with any electrons in the body of the bar, they reach the surfaces at $x = \pm L/2$. Since the surfaces display an infinite surface recombination velocity, all holes recombine there. In consequence, the excess hole concentration goes to zero at the surface. In other words, the surface is maintained in thermal equilibrium. The shape of the carrier concentration in this example falls again from a maximum at $x = 0$ to 0 at $x = \pm L/2$. Since there is no recombination in the bulk of the bar, the hole current is constant. This calls for a constant slope in the excess carrier concentration or a linear profile, as sketched in Fig. 5.16.

Let us now solve the problem quantitatively. The differential equation that governs hole flow in this case is simply:

$$\frac{d^2p'}{dx^2} = 0 \tag{5.70}$$

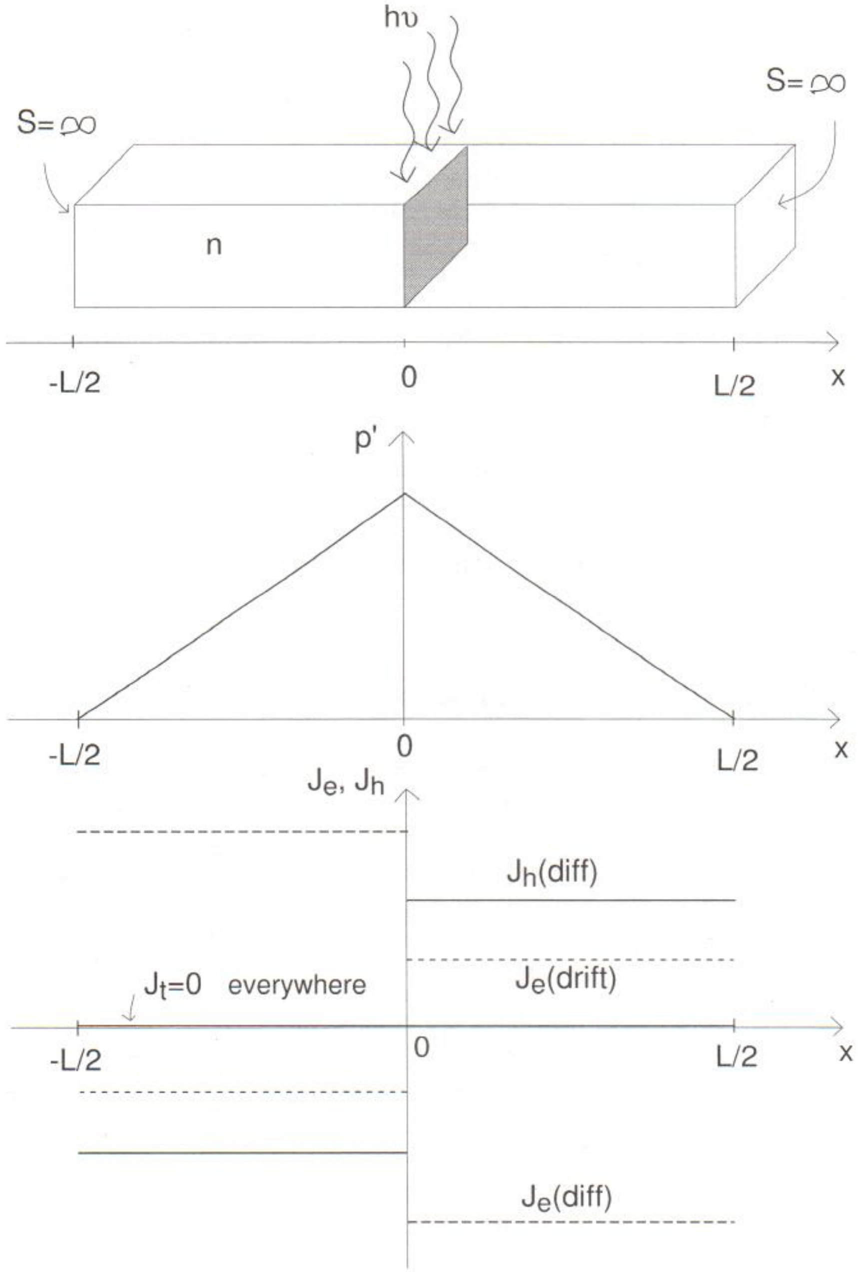


Figure 5.16: Top: a short bar illuminated in the middle by a sheet of generation; middle: excess hole concentration; bottom: electron and hole currents.

A general solution to this equation has the form:

$$p' = Ax + B \tag{5.71}$$

This problem has two boundary conditions. At $x = 0$ the boundary condition is identical to the previous case, Eq. 5.56. At the surfaces, the boundary condition is simply:

$$p'(\pm\frac{L}{2}) = 0 \tag{5.72}$$

For $x \geq 0$, the solution of the differential equation 5.70 is:

$$p' = -\frac{g_l}{2D_h}(x - \frac{L}{2}) \tag{5.73}$$

the result that we expected.

The hole current density is:

$$J_h = q\frac{g_l}{2} \tag{5.74}$$

This is a result that we could have written from the very beginning. If there is no recombination in the body of the bar, half of the generated carriers flow towards one ohmic contact and half flow to the other. The flow rate is then $g_l/2$ and the current density is $qg_l/2$.

Quasi-neutrality demands that $n' \simeq p'$, from which we can get the electron diffusion current to be:

$$J_e(diff) = -q\frac{D_e}{D_h}\frac{g_l}{2} \tag{5.75}$$

In this example, the total current is also zero since the bar is not connected to anything. From this condition, we can get the electron drift current to be:

$$J_e(drift) = \frac{qg_l}{2}\frac{D_e - D_h}{D_h} \tag{5.76}$$

All signs in Eqs. 5.74-5.76 are reversed for $x \leq 0$.

Following a similar procedure to the previous example, the various assumptions can be verified and the maximum value of g_l allowed to maintain low-level injection conditions can be obtained.

Before closing, it is also of interest to compute the velocity at which holes diffuse through the bar. Using Eqs. 5.73 and 5.74, we get for $x > 0$:

$$v_h^{diff}(x) = \frac{J_h(x)}{qp(x)} \simeq \frac{D_h}{\frac{L}{2} - x} \tag{5.77}$$

When comparing this result with that obtained for the long bar in Eq. 5.69, we find that in the short bar, the hole diffusion velocity increases as the carriers approach the contact. This makes sense since the carrier concentration decreases but its slope is constant. A second interesting point is that the hole diffusion velocity diverges at the surface of the semiconductor. This is an artifact of our assumption that $p(x) \simeq p'(x)$. A finite velocity is obtained if we carry out a more careful analysis. In many situations, this is actually not needed and the result of Eq. 5.77 is quite adequate.

5.6.3 Length scales of minority-carrier situations

The previous examples have revealed the existence of two characteristic lengths in minority carrier-type problems: the diffusion length L_{diff} and the sample length L.

L_{diff} is the average length that a carrier diffuses in a bulk semiconductor before it recombines. It was mathematically defined for holes in Eq. 5.54. A similar equation applies to electrons. L_{diff} makes a statement about the balance between diffusion and recombination. The more effective bulk recombination is (by having τ smaller), the shorter L_{diff} becomes.

Since both D and τ depend on doping, the diffusion length is a strong function of doping level. This is shown for Si at room temperature in Fig. 5.17. The lines in this figure come from combining the carrier lifetime data in Fig. 3.17, the mobility data in Fig. 4.3 and from using the Einstein relation to obtain the diffusion coefficient from the mobility. We know that at low and moderate doping levels, the carrier lifetime is not a very tight function of doping level. In consequence, one should expect a wide range of possible values for the diffusion length in this same doping regime. The lines of Fig. 5.17 represent guidelines to reasonable values.

In a given situation, the smallest one of the two characteristic lengths, the sample length or the diffusion length, dominates. Section 5.6.1 showed an example in which the sample length is much larger than the diffusion length. The carrier profiles and the electrostatics of the problem were entirely dominated by the diffusion length. In a situation like this, the semiconductor is said to be *long* or *opaque* from the minority carrier point of view. Section 5.6.2, on the other hand, presented an example in which the diffusion length was much longer than the sample length. In that case, the solution of the problem was entirely dominated by the sample size and bulk recombination was irrelevant. From the minority carrier point of view, the semiconductor in a situation like this is often referred to as *short* or *transparent*.

5.7 Dynamics of majority carrier situations

Time-dependent situations are particularly important in device operation. There are many cases in which we are interested in the dynamic response of a device to a time changing stimulus.

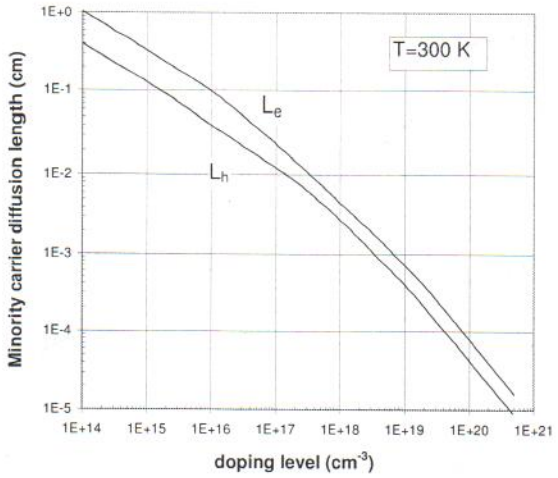


Figure 5.17: Minority carrier diffusion length in Si at room temperature. The lines represent typical values corresponding to the carrier lifetimes given by the lines of Fig. 3.17.

Time-dependent problems are mathematically quite difficult. Few of them result in simple analytical solutions. The goal of this section and the next is to learn to recognize in a given situation the various physical processes at play, to identify the relevant time constants, and to be able to compute order of magnitude values for these time constants.

The equation set that we derived to describe majority carrier situations in Table 5.4 does not contain any time dependent terms. This would suggest that majority carrier situations are quasi-static, meaning that they respond instantaneously to outside stimuli without any memory of its previous state. This is a reasonable approximation in many practical situations. In an ideal integrated resistor such as the one described in Section 5.5.1, if we can disregard its parasitic capacitance, it is usually a fairly good assumption that the current follows the voltage instantaneously.

It is useful to understand where this important result derives from and what are its limits. If we look back at the simplified Shockley equations under quasi-neutrality (Table 5.3) and we assume, as we did in the simplifications leading to Table 5.4, that the carrier concentrations do not change in time, then there are no dynamics left in the description of the situation. The issue is then the quasi-neutrality approximation itself. It is in making this assumption that the time derivative of the volume charge density was dropped. The rational was that if the volume charge density is negligible, then its change in time is also negligible. But how good is this assumption?

In majority carrier situations, there are always small discrepancies to perfect charge neutrality. Net charge appears at contacts (we will study this in Ch. 7), and in the presence of dopant gradients. The quasi-neutrality assumption allows us to neglect this net charge because it is relatively small and/or it is confined to small regions in the scale of the sample size. In a dynamic situation, such as when a voltage step is applied to an integrated resistor, the net charge needs to change to satisfy the changing electrostatics. To accomplish this, some amount of charge needs to be delivered to the appropriate locations in the device. That takes a short but finite time. A

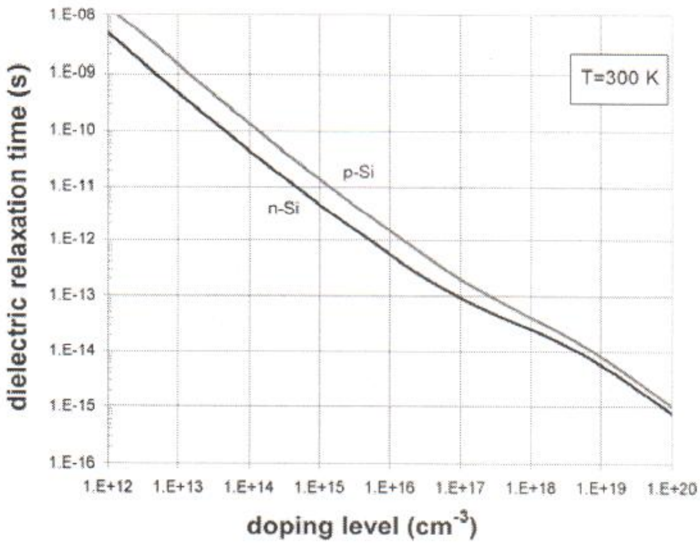


Figure 5.18: Dielectric relaxation time for n- and p-type Si at room temperature.

derivation of the appropriate time for typical situations is performed in Advanced Topic AT5.2. There we find that the time constant for volume charge to relax is called the *dielectric relaxation time* and it is given by:

$$\tau_d = \frac{\epsilon}{\sigma} \tag{5.78}$$

where σ is the conductivity of the semiconductor region in question. The dependencies of this simple equation make sense when we realize that it is the electric field that makes charge move through the drift process.

Fig. 5.18 graphs τ_d for n- and p-type Si at room temperature. As the doping level increases, σ increases and τ_d decreases as a result. τ_d is indeed rather small for medium and highly-doped regions. For example, for Si with a doping level of order 10^{16} cm^{-3}, τ_d is about 1 ps. In microelectronic devices, the doping levels are typically higher and they operate in time scales usually longer than this. We can then conclude that the charge redistribution that takes place in response to outside stimuli in a majority carrier situation occurs on a time scale that is much shorter than our time scale of interest. In this time scale, we can assume that majority carrier situations are indeed quasi-static.

Exercise 5.4: *Up to what frequency can a 1 $\Omega \cdot cm$ n-type Si substrate be considered quasi-static to the application of electric fields?*

The dielectric relaxation time for this substrate is:

$$\tau_d = \frac{\epsilon}{\sigma} = \epsilon\rho = 11.7 \times 8.85 \times 10^{-14} \times 1 = 1.0 \times 10^{-12}\ s = 1\ ps$$

The inverse of the dielectric relaxation time divided by 2π is a reasonable estimate of the maximum frequency at which this substrate can be considered quasi-static:

$$f_d \simeq \frac{1}{2\pi\tau_d} \simeq 150\ GHz$$

This is a frequency substantially beyond most applications of Si today (but perhaps not in the future). Hence, this substrate will respond in a perfectly quasi-static way to the application of electric fields.

Had we worked with a 100 $\Omega \cdot cm$ substrate, the frequency in question would have been 1.5 GHz. This is in the range of many applications today. The response of such a substrate to electric fields will be far from quasi-static and would have to be taken into account in equivalent circuit models of devices.

5.8 Dynamics of minority carrier situations

In minority carrier situations, there are excess carriers which can exhibit substantial memory effects. These typically dominate the dynamics of minority carrier devices. To first order, this can be understood from the fact that carrier lifetimes are long, in the ns to μs range. This argument is misleading, however, since often, the dynamics of minority carriers are not controlled by the recombination process but by their transport through a certain region. In this case, the proper time constant is not the lifetime but the *transit time*. Understanding this is the central goal of this section.

5.8.1 Example 3: Transient in a bar with $S = \infty$

Consider a uniformly doped n-type semiconductor bar of length L as in Fig. 5.19. The two surfaces of the bar are characterized by an infinite surface recombination velocity. For $t \leq 0$, the bar is illuminated with radiation that generates carriers uniformly everywhere at a rate g_l. At $t = 0$, the radiation is switched off. We are interested in the time evolution of the excess carrier profile everywhere in the bar. In particular, we want to identify the dominant time constant of the decay of excess carriers.

We studied time transients in Ch. 3. The types of problems that we dealt with at that time were uniform, that is, nothing changed in space. Their characteristic time constant was the carrier lifetime τ. The present problem might appear similar at first sight. After all, the radiation is of uniform intensity in space and bulk recombination takes place everywhere in the bar. However, the presence of the surfaces with $S = \infty$ changes the situation. Excess carriers in the vicinity of these surfaces can recombine faster than in the bulk if it takes them a shorter time to diffuse

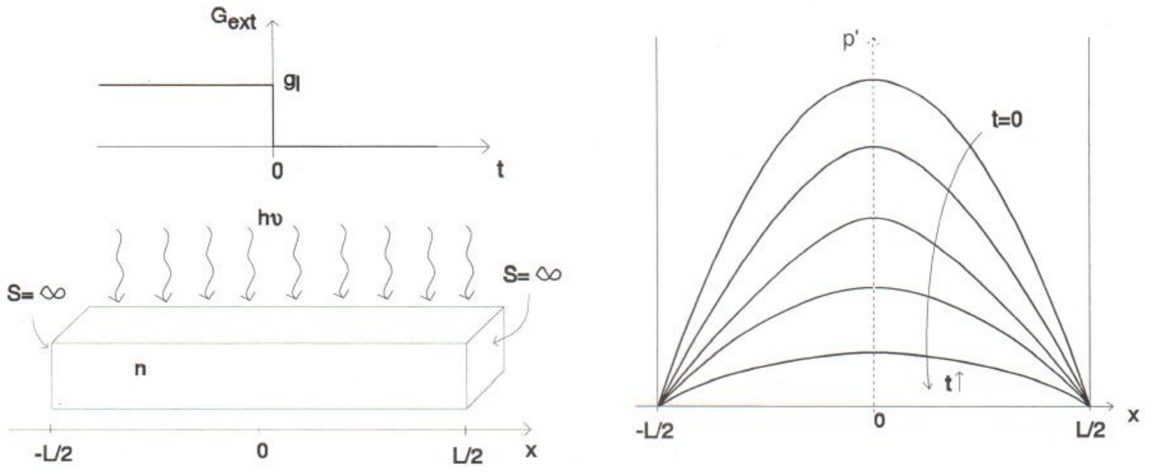


Figure 5.19: Sketch of time decay of excess carrier concentration in a semiconductor bar after a uniform generation function has been turned off. The surfaces are characterized by $S = \infty$.

to the surface. The recognition of this simple fact immediately allows us to conclude that: 1) the characteristic time constant of the decay of the excess carrier concentration is *shorter* than τ, and 2) the time constant associated with surface recombination is directly proportional to the length of the sample and inversely proportional to the diffusion coefficient (the longer the sample and the smaller the diffusion coefficient, the less effective surface recombination is relative to bulk recombination).

The mathematical solution of this problem is interesting. The procedure to follow is general and widely used for different kinds of transient problems. The first step is to compute the steady state excess carrier profile that exists before the pulse is turned off. Only then, the decay of carrier concentration can be studied. The problem has symmetry around $x = 0$, so only for $x \geq 0$ a solution is needed.

For $t \leq 0$, the differential equation that governs this problem is:

$$D_h \frac{d^2 p'}{dx^2} - \frac{p'}{\tau} + G_{ext} = 0 \tag{5.79}$$

Substituting $G_{ext} = g_l$, and dividing all terms by D_h, we get:

$$\frac{d^2 p'}{dx^2} - \frac{p'}{L_h^2} + \frac{g_l}{D_h} = 0 \tag{5.80}$$

A solution to this differential equation is [3]:

[3] A solution of the form $p' = A \exp \frac{x}{L_h} + B \exp \frac{-x}{L_h} + C$ is also valid but results in more complex algebra.

$$p' = A\cosh\frac{x}{L_h} + B\sinh\frac{x}{L_h} + C \tag{5.81}$$

Substitution of this expression into the differential equation 5.80, immediately gives $C = g_l\tau$. Symmetry around $x = 0$ imposes the boundary condition $dp'/dx|_{x=0} = 0$. Applying to Eq. 5.81 results in $B = 0$. The infinitely recombining surface at $x = L/2$ demands that $p'(L/2) = 0$. This implies that $A = -g_l\tau/\cosh\frac{L}{2L_h}$. All together, the steady-state solution at $t = 0$ is:

$$p'(x,0) = g_l\tau(1 - \frac{\cosh\frac{x}{L_h}}{\cosh\frac{L}{2L_h}}) \tag{5.82}$$

For $t \geq 0$, the governing differential equation is:

$$D_h\frac{\partial^2 p'}{\partial x^2} - \frac{p'}{\tau} = \frac{\partial p'}{\partial t} \tag{5.83}$$

In a standard manipulation, we first perform a change of variables. If we define:

$$p' = P\exp(-\frac{t}{\tau}) \tag{5.84}$$

The differential equation that P satisfies becomes:

$$D_h\frac{\partial^2 P}{\partial x^2} = \frac{\partial P}{\partial t} \tag{5.85}$$

This equation can be attacked by the method of separation of variables. Let us postulate that the solution to 5.85 is of the form:

$$P(x,t) = X(x)T(t) \tag{5.86}$$

where X only depends on x, and T only depends on t. Substitution of Eq. 5.86 in 5.85 leads to two separate differential equations:

$$\frac{1}{X}\frac{d^2X}{dx^2} = \frac{1}{D_h}\frac{1}{T}\frac{dT}{dt} = -\frac{1}{\lambda^2} \tag{5.87}$$

The first differential equation is only in terms of the independent variable x. The second one depends only on t. The only way for them to equal each other is if they are also equal to a constant. As we will see in a moment, this separation constant must be negative. To force that, we denote it as $-1/\lambda^2$. We now proceed to look at these two equations separately.

The time-dependent equation has a solution:

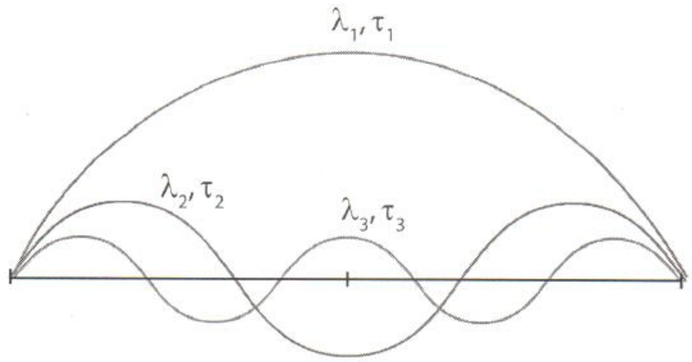


Figure 5.20: Spatial dependence of the first three modes of $p'_n(x,t)$.

$$T = K_1 \exp(-\frac{D_h}{\lambda^2}t) \tag{5.88}$$

which is a decaying exponential. If the separation constant in Eq. 5.87 was positive, the solution 5.88 would be increasing with time, which is unphysical. Similarly, a separation constant of zero would imply that the only time transient is the one identified in Eq. 5.84, which we have already argued is insufficient.

The differential equation in x in 5.87 has a solution:

$$X = K_2 \cos\frac{x}{\lambda} + K_3 \sin\frac{x}{\lambda} \tag{5.89}$$

The boundary conditions that it must satisfy are the same as for $t \leq 0$. $dX/dx|_{x=0} = 0$ implies that $K_3 = 0$. $X(L/2) = 0$ can only be satisfied if $\cos(L/2\lambda) = 0$, which in turn, imposes a restriction on the values that λ can take:

$$\lambda_n = \frac{L}{(2n-1)\pi} \qquad \text{for } n = 1, 2, 3, ... \tag{5.90}$$

We actually find that there are multiple solutions to this problem, each one characterized by a different separation constant given in Eq. 5.90. Assembling Eqs. 5.84, 5.86, 5.89, 5.88, and 5.90 the expression of one of these solutions is:

$$p'_n(x,t) = K_n \exp(-\frac{t}{\tau}) \exp(-\frac{D_h}{\lambda_n^2}t) \cos\frac{x}{\lambda_n} \tag{5.91}$$

where we have absorbed all proportionality constants into one. A sketch of the first three modes is shown in Fig. 5.20.

The most general solution of this problem is then constructed by the superposition of all these solutions:

$$p'_n(x,t) = \exp -\frac{t}{\tau} \sum_n K_n \exp(-\frac{D_h}{\lambda_n^2}t) \cos \frac{x}{\lambda_n} \qquad \text{for } n = 1, 2, 3, ... \tag{5.92}$$

The weight given to each of these components is such that at $t = 0$, $p'(x, 0)$ given in 5.92 equals the solution derived above in 5.82. From Fourier analysis, it is straightforward to derive expressions for K_n:

$$K_n = \frac{1}{L} \int_{-L/2}^{L/2} p'(x,0) \cos[\frac{(2n-1)\pi}{L} x]\, dx = \frac{4 g_l \tau \sinh \frac{L}{2L_h}}{(2n-1)^2 \pi^2 \frac{L_h}{L} + \frac{L}{L_h}} \tag{5.93}$$

Several observations can be made about the solution given in Eq. 5.92. Unlike the case of pure bulk recombination, the complete solution does not exhibit a simple exponential decay. Rather, the complete solution is the sum of many independent exponentially decaying functions, each characterized by a different characteristic time. The nth component exhibits a time constant:

$$\frac{1}{\tau_n} = \frac{1}{\tau} + D_h [\frac{(2n-1)\pi}{L}]^2 \qquad \text{for } n = 1, 2, 3, ... \tag{5.94}$$

All of these time constants are smaller than the carrier lifetime. The higher the order of the component, the shorter the characteristic time constant. In consequence, after the light is switched off, there is a fast decay of the carrier profile that is dominated by the higher order modes. After a short time, the higher order terms become very small and the decay of the carrier profile is dominated by the behavior of the first-order component. This is illustrated at the center of the bar in Fig. 5.21. This interesting time evolution is actually what is obtained in practice, as you can see in the experimental data of Fig. 3.22.

The time constant of the first-order mode is:

$$\frac{1}{\tau_1} = \frac{1}{\tau} + D_h (\frac{\pi}{L})^2 \tag{5.95}$$

Since this is the longest one, this is the dominant time constant of this problem. This time constant has all the features that we expected. It consists of two terms, indicating that there are two recombination processes taking place in parallel. One of the processes is bulk recombination characterized by a time constant τ, the carrier lifetime. The other process is surface recombination characterized by a time constant $L^2/\pi^2 D_h$. The dependences of this surface time constant are exactly what we expected. It is directly proportional to the length of the sample and inversely proportional to the diffusion coefficient. In fact, if you work it out (see Prob. 5.28), you will see that $L^2/\pi^2 D_h$ is the *transit time* of holes through the sample to its surface. That is, the average time it takes for a photogenerated hole to diffuse from the point of generation to the sample surface where it recombines. In a more general way, we could then write:

$$\frac{1}{\tau_1} = \frac{1}{\tau} + \frac{1}{\tau_t} \tag{5.96}$$

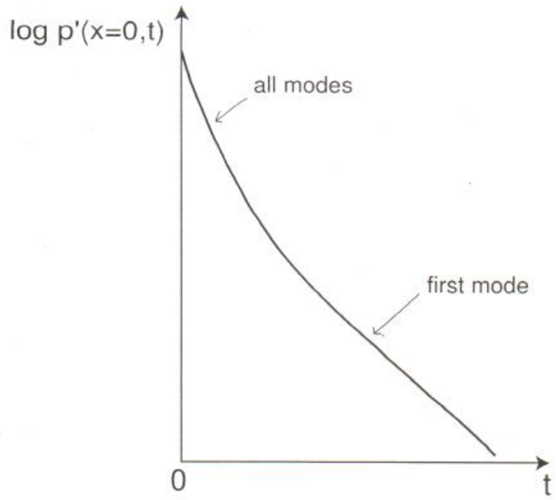


Figure 5.21: Sketch of time decay of excess carrier concentration at the center of a semiconductor bar after a uniform generation function has been turned off. After a fast initial decay, the dominant first mode emerges and controls the rest of the time evolution. The experimental data graphed in Fig. 3.22 shows this behavior.

where τ_t is the transit time. τ_t is given by:

$$\tau_t = \frac{L^2}{\pi^2 D_h} \tag{5.97}$$

The concept of transit time, was already presented in Sec. 4.4. The numerical factor in the expression of the transit time (π^2 in this case) depends on the details of the excess carrier profile ($\cos(\pi x/L)$ for this example; the linear profile discussed in Sec. 4.4 resulted in a numerical factor of 2). The key dependencies, however, are the same. The transit time is always proportional to the *square of the sample length* and inversely proportional to the diffusion coefficient. As a reminder, the L^2 dependence arises from: *i)* the distance that carriers must diffuse is on the order of L, and *ii)* the concentration gradient that drives carrier diffusion is sharper the shorter L is.

Coming back to Eq. 5.95, the time constant of the first-order mode is given by the inverse of the sum of the inverse of the bulk lifetime and the transit time. This suggests that the *smallest* of these two characteristic time constants dominates the decay of the excess carrier concentration in this bar. This is understandable. In this problem there are two recombination processes operating in parallel: bulk recombination and surface recombination. Whichever one is most effective (*i.e.* is characterized by the smallest time constant), dominates the overall decay. This implies, then, that in general $\tau_1 < \tau$. In the limit of slow bulk recombination, $\tau_1 \simeq \tau_t$, *i.e.*, the decay of the excess minority carrier concentration is dominated by their transit to the surface.

The key results of this section apply to other situations with identical boundary conditions. In essence, what we have done here is to examine the time decay of the Fourier components of the minority carrier profile at $t = 0$. In a situation characterized by a different spatial distribution of the generation function but identical minority carrier boundary conditions to those studied in this section, the carrier profile at $t = 0$ can be decomposed in a sum that is composed of the same Fourier terms as Eq. 5.92. The coefficients associated with each time constant are likely to be different to those derived in this section, but the time constant of each of the Fourier terms is identical. In particular, the expression of the time constant of the dominant mode should follow Eq. 5.95.

Exercise 5.5: *Consider a situation like the one depicted in Fig. 5.19. This is an ideal 1D situation involving an n-type Si bar of length $L = 100\ \mu m$ with a doping level $N_D = 10^{17}\ cm^{-3}$ at room temperature. The surfaces located at $-L/2$ and $L/2$ are characterized by $S = \infty$. Estimate the value of the dominant time constant for the decay of excess minority carrier concentration after the light illumination is turned off. Assume low-level injection conditions.*

Using the fit in Fig. 3.17 for the carrier lifetime in n-type Si, for this doping level, we estimate a carrier lifetime of $\tau \simeq 13\ \mu s$. Using the fit to the hole mobility given in Appendix E, we estimate a hole mobility in this bar of $\mu_h \simeq 455\ cm^2/V.s$. This implies a hole diffusion coefficient of $D_h \simeq 12\ cm^2/s$. The time constant associated with the transit of holes to the surfaces where they recombine is then:

$$\tau_t = \frac{L^2}{\pi^2 D_h} = \frac{(0.01\ cm)^2}{\pi^2 \times 12\ cm^2/s} = 0.85\ \mu s$$

The dominant time constant of the excess hole decay is given by:

$$\tau_1 = \frac{1}{\frac{1}{\tau} + \frac{1}{\tau_t}} = \frac{1}{\frac{1}{13\ \mu s} + \frac{1}{0.85\ \mu s}} = 0.8\ \mu s$$

In this situation, the fastest process is hole transit to the surfaces where they recombine. Therefore, the hole transit time nearly completely sets the value of the dominant time constant.

5.9 Transport in space-charge and high-resistivity regions

If we examine a semiconductor device from a space charge point of view, we will notice that there are regions with a very small volume charge, called quasi-neutral regions (QNR), and other regions where there is substantial spatial charge called "space-charge regions" (SCR). For example, in a p-n junction, two quasi-neutral p and n regions are separated by a high resistivity SCR located around the "metallurgical" interface where the doping level changes from p to n. In a MOSFET, a SCR also exists underneath the inversion layer.

As we have extensively discussed, QNRs are characterized by a high carrier concentration and low resistivity. SCRs, on the other hand, exhibit very low carrier concentrations and therefore have a high resistivity. Carrier transport through SCRs is of a markedly different nature than through QNRs. The dielectric relaxation time in a high-resistivity region can be very long. For example, $10\ \Omega \cdot cm$ Si has a dielectric relaxation time of about $4\ ns$. This means that majority carriers can take a very long time, relative to other relevant time constants, to screen out a charge perturbation. Similarly, the Debye length can be very long in an SCR. For a carrier concentration of $10^{12}\ cm^{-3}$, the Debye length is about $4\ \mu m$. This implies that SCRs can sustain net charge over substantial dimensions.

SCRs belong to a general class of high-resistivity regions with very low carrier concentrations.

$\frac{\partial \mathcal{E}}{\partial x} \simeq \frac{q}{\epsilon}(N_D - N_A)$ or $\mathcal{E} = \mathcal{E}'$
$J_e = -qnv_e^{drift}(\mathcal{E}) + qD_e\frac{\partial n}{\partial x}$
$J_h = qpv_h^{drift}(\mathcal{E}) - qD_h\frac{\partial p}{\partial x}$
$\frac{\partial n}{\partial t} = G_{ext} - U(n,p) + \frac{1}{q}\frac{\partial J_e}{\partial x}$
$\frac{\partial p}{\partial t} = G_{ext} - U(n,p) - \frac{1}{q}\frac{\partial J_h}{\partial x}$
$J_t = J_e + J_h$

Table 5.6: Equation set for space-charge and high-resistivity regions. In the entry that corresponds to Gauss' law (top line), the equation on the left applies to space-charge regions. The equation on the right applies to high-resistivity regions.

In these, transport is of a rather different nature than in QNR's. Additionally, and largely due to the low carrier concentrations, large electric fields are often present (that is the case of SCRs, in fact) or are applied from the outside.

A treatment of transport in high-resistivity regions has to be based on the Shockley equations *before* the quasi-neutral approximation was made (see Fig. 5.8). That is, we start with the equation set of Table 5.2. This brings us back to a system of equations that is highly coupled. Our strategy for uncoupling these equations again is to look at simplifying Gauss' law, although in a different way than in the QNR case. The one key feature of high-resistivity regions that we can exploit here is the fact that due to the low carrier concentrations that exist, *the electric field is independent of the carrier concentrations.* The electric field is typically set by ionized dopants (the case of SCRs) or by an external applied voltage.

A corollary of this is that *the behavior of electrons and holes in a high-resistivity region is largely uncorrelated.* The reason for this is that the coupling between the electron and hole profiles is made by the electric field. This is most clear in a quasi-neutral region in which the need to maintain $\rho \simeq 0$, tightly binds up n and p. In high-resistivity regions, on the contrary, the electric field is independent of the carrier concentrations and hence n and p are independent of each other.

A simplified set of Schockley equations for SCR and high-resistivity regions is in Table 5.6. There are two entries for Gauss' law (top line). The first one applies to space-charge regions in which the electric field is determined by the ionized impurity distributions. The second one corresponds to high-resistivity regions in which the electric field is set by the application of an external voltage.

Typically, this equation set is to be solved in the following way. The electric field is first obtained using the appropriate version of Gauss' law, as explained in the previous paragraph. Then, the electron and hole profiles are determined by solving the equation that results from plugging the current equation into the corresponding continuity equation. Finally, the current equation is used to derive the current.

The procedure is particularly easy for a steady-state situation. For electrons, for example, integrating the electron continuity equation from a point x_1 to another point x_2 yields:

$$J_e(x_1) - J_e(x_2) = -q \int_{x_1}^{x_2} [G_{ext} - U(n,p)]dx \tag{5.98}$$

In situations in which there is external generation, we can often neglect the net recombination inside the region in question since n and p tend to be relatively small. In this case, we have:

$$J_e(x_1) - J_e(x_2) = -q \int_{x_1}^{x_2} G_{ext}\, dx \tag{5.99}$$

from which the current can be quickly obtained.

An example is the best way to illustrate these procedures.

5.9.1 Example 4: Drift in a high-resistivity region under external electric field

This is a one-dimensional problem. Consider a high-resistivity chunk of n-type semiconductor of length L with metallic contacts applied to it. Let us assume that these contacts are ideal, that is, the bulk properties of the semiconductor extend all the way to the interface with the metal. In equilibrium, the semiconductor is charge neutral with $n_o = N_D$ and $p_o = n_i^2/n_o$.

Let us now apply a voltage V between the two contacts. This produces a uniform electric field $\mathcal{E} = V/L$ across the sample. This case appears similar to the situations that we studied in section 5.5. In response to the electric field carriers drift and current flows through the structure. In general, the current density is given by: $J = q(n_o v_e^{drift} + p_o v_h^{drift})$, where the drift velocities v_e^{drift} and v_h^{drift} depend on the magnitude of the electric field. If N_D is small, this current can be very small. For example, if $n_o = 10^{12}\ cm^{-3}$, the maximum current density in a Si sample at room temperature is obtained under velocity saturation conditions and is about 1.6 A/cm^2 (for a sense of proportion, typical microelectronic devices operate with current densities as high as $10^5\ A/cm^2$).

Let us now illuminate the sample with a very narrow beam of photons that produces g_l electron-hole pairs per unit area per second uniformly across the entire section of the sample at a location x_o, as sketched in Fig. 5.22. What is the resulting carrier concentration? How much current flows through the sample?

The electric field that exists inside the sample separates the photogenerated carriers. The photogenerated electrons drift to the positive terminal, while the holes drift to the negative terminal. If both recombination is negligible, under steady state circumstances, the electron current density is simply:

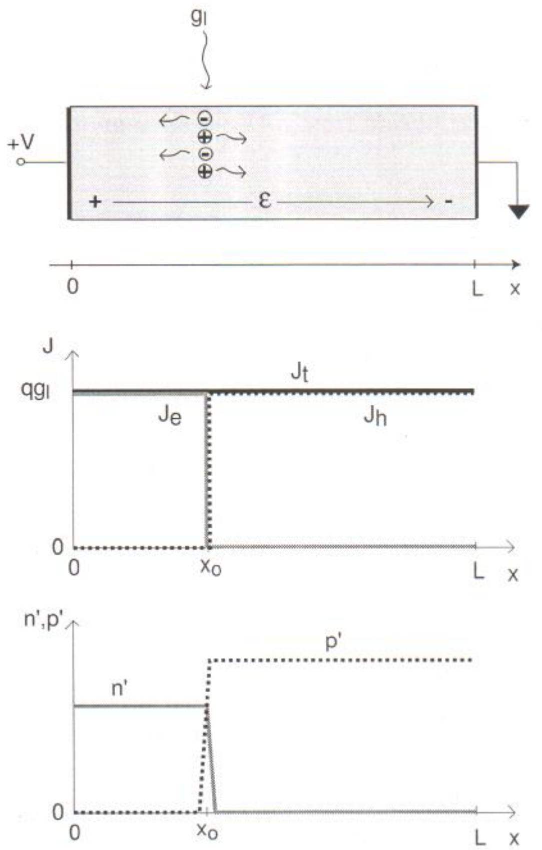


Figure 5.22: High-resistivity semiconductor bar under an applied electric field and a steady-state photogeneration at x_o. The middle diagram shows the carrier current densities. The bottom diagram shows the excess carrier concentrations.

$$J_e = qg_l \qquad \text{for } x < x_o \tag{5.100}$$
$$J_e = 0 \qquad \text{for } x > x_o \tag{5.101}$$

Similarly, the hole current density is:

$$J_h = 0 \qquad \text{for } x < x_o \tag{5.102}$$
$$J_h = qg_l \qquad \text{for } x > x_o \tag{5.103}$$

In consequence, the total current density is:

$$J_t = qg_l \tag{5.104}$$

everywhere, as required by the continuity equations. This is sketched in Fig. 5.22.

The excess carrier densities are not difficult to compute. Excess electrons are only to be found to the left of the generation point. In that region, since the electron current is due to electron drift and is constant, the electron concentration is simply uniform in space. Hence:

$$n' \simeq \frac{g_l}{v_e^{drift}(\mathcal{E})} \qquad \text{for } 0 < x < x_o \tag{5.105}$$

$$n' \simeq 0 \qquad \text{for } x_o < x < L \tag{5.106}$$

Similarly for holes:

$$p' \simeq 0 \qquad \text{for } 0 < x < x_o \tag{5.107}$$

$$p' \simeq \frac{g_l}{v_h^{drift}(\mathcal{E})} \qquad \text{for } x_o < x < L \tag{5.108}$$

Since for a given electric field, in general $v_h^{drift} < v_e^{drift}$, it follows that $p' > n'$, as sketched in Fig. 5.22.

Around x_o some diffusion takes place against the electric field. In consequence the excess carrier profile softly drops to zero. The actual shape can be computed in a similar way as in Section AT5.3.1.

The excess carrier situation depicted in Fig. 5.22 is no longer quasi-neutral. The intense electric field has separated the photogenerated carriers. A charge dipole has been established across the generation point.

In the derivation of Eqs. 5.105-5.108, we did not need to make the low-level injection approximation. In fact, since the doping level is so low, it is very likely that the carrier concentrations greatly exceed the doping level. The one assumption that we have implicitly made is that the carrier concentrations are not so large that they modify the electric field that was set up inside the bar by the application of the voltage. Without this assumption, the problem would be a lot more difficult since the field is modified by the carrier concentrations which in turn are determined by the net magnitude of the field. These kinds of situations are known as *space-charge limited transport* and their study is beyond the objectives of this book.

5.9.2 Comparison between SCR and QNR transport

The best way to appreciate the peculiarities of SCR transport is to compare it with transport in a quasi-neutral region. Consider the example of Fig. 5.23. Shown are two uniformly-doped n-type semiconductor bars with ohmic contacts at their ends. A voltage across both ohmic contacts is applied. At some location in the bar, a narrow beam of light generates electron-hole pairs at a constant rate. The bar on the left has a very low doping level and a high resistivity, as discussed in the previous subsection. The bar on the right has a low resistivity and can be considered quasi-neutral and under low-level injection. Let us assume that the bar is "short" as compared to the minority carrier diffusion length. Both are steaty-state situations.

Below each bar the excess carrier concentrations and the current densities are displayed. The case on the left has been discussed in the previous subsection in detail. The electric field separates the photogenerated carriers and a current density $J_t = qg_l$ is established along the bar and the outside circuit.

The case on the right requires a bit more thinking. This is a mixed minority-majority carrier situation. We certainly know that there is no carrier separation in this case, since quasi-neutrality imposes $n' \simeq p'$. What about the current? Since the sample has significant doping, there must certainly be a majority carrier current that results from the application of the voltage, as studied in Section 5.5. But does carrier generation also result in external current? The best way to answer this question is to go back to the general equation set for quasi-neutral carrier transport of Table 5.3. If, for simplicity, we assume that the electric field is small enough that mobility-type drift applies, summing J_e and J_h we obtain the following expression for the total current:

$$J_t = J_e + J_h = q(n\mu_e + p\mu_h)\mathcal{E} + qD_e\frac{dn}{dx} - qD_h\frac{dp}{dx} \tag{5.109}$$

Since the bar is uniformly doped, n_o and p_o do not depend on position and $\mathcal{E}_o = 0$. Also, since the bar is n-type, $n_o \gg p_o$. Eq. 5.109 then simplifies to:

$$J_t \simeq qn_o\mu_e\mathcal{E}' + qD_e\frac{dn'}{dx} - qD_h\frac{dp'}{dx} \tag{5.110}$$

Let us now integrate this equation along the bar, from $x = 0$ to $x = L$:

$$J_tL \simeq qn_o\mu_eV + qD_e[n'(L) - n'(0)] - qD_h[p'(L) - p'(0)] \tag{5.111}$$

Since at $x = 0$ and $x = L$ the bar has ohmic contacts where excess carrier concentrations cannot be sustained, $n'(0) = n'(L) = p'(0) = p'(L) = 0$. Hence, solving for the current, we find:

$$J_t = \frac{qn_o\mu_e}{L}V \tag{5.112}$$

This is precisely the result that we obtained in Section 5.5 for an identical bar but in the absence of carrier generation. Hence, in the low-resistivity sample, carrier photogeneration does not result in external photocurrent through the circuit. Only a majority-carrier type current flows as a result of the application of a voltage. In contrast, carrier separation in the high-resistivity sample results in a photocurrent that is directly proportional to the generation rate, as given by Eq. 5.104. In other words, in the high-resistivity sample, the photogenerated carriers are spatially separated. Because of this and also as a consequence of the low background carrier concentration, recombination cannot take place and all the photogenerated carriers get extracted

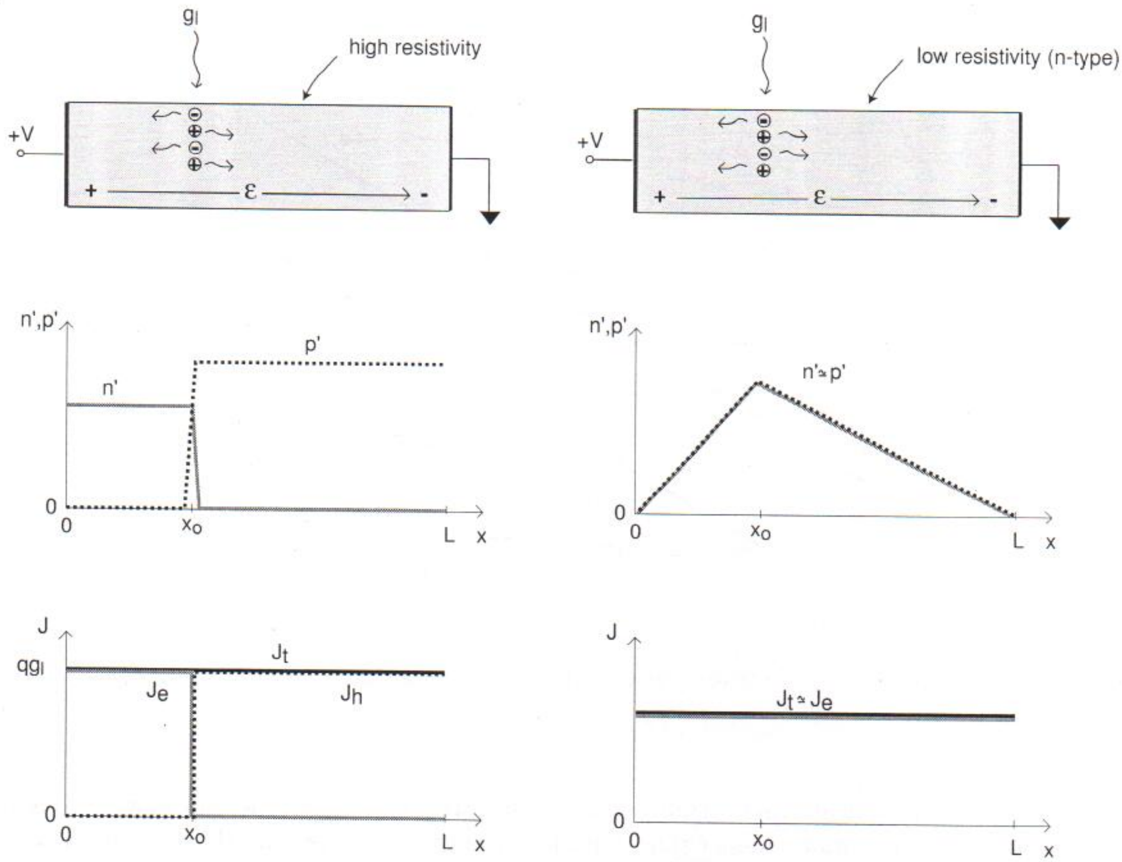


Figure 5.23: Comparison of SCR transport vs. QNR transport. On the left is a high-resistivity semiconductor bar. On the right is a low-resistivity semiconductor bar that is shorter than the diffusion length. Both are under an applied electric field and a steady-state photogeneration at x_o. The middle set of diagrams shows the excess carrier concentrations. The bottom diagram shows the carrier current densities.

out by the electric field. In contrast, in the low-resistivity sample, all the photogenerated carriers recombine at the ohmic contacts and no photogenerated current results. [4]

5.10 Carrier multiplication and avalanche breakdown

If a high electric field is present in a semiconductor sample, impact ionization might take place. Our analysis so far has not accounted for it. The goal of this section is to understand its implications.

As we discussed in Ch. 3, the driving force behind impact ionization is the carrier flux. When

[4]Note that there is a small *photoconductive* current in the low-resistivity sample that arises from the small modification of the conductivity of the sample as a result of carrier generation. This current is much smaller than qg_l, the maximum photogenerated current that can be produced.

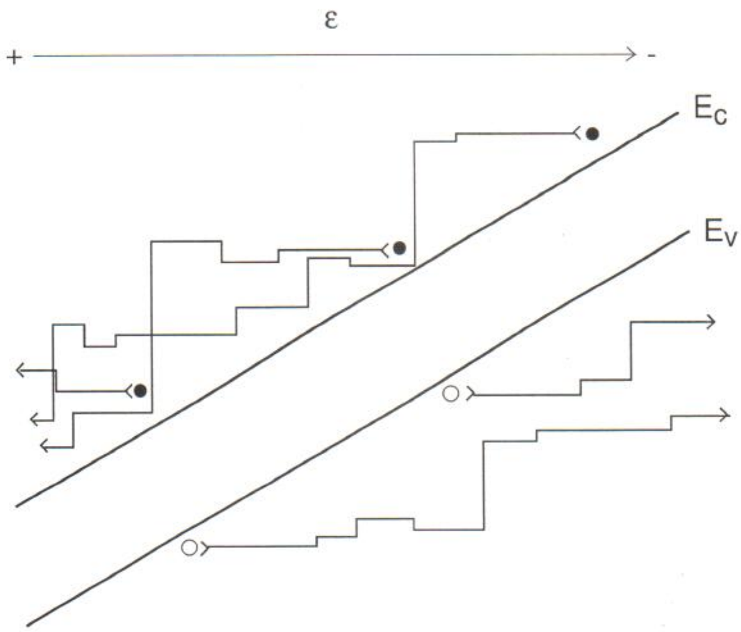


Figure 5.24: Sketch of carrier multiplication produced by impact ionization in a semiconductor region with a uniform electric field.

impact ionization is significant in a certain semiconductor region, the current that flows through that region can be substantially higher than expected from the analysis carried out so far in this chapter. In addition to the current contributed by the primary carriers, our focus up to this point, we must account for the secondary carriers that are generated by impact ionization. If the electric field is high enough, there might well be tertiary carriers, generated by the secondary carriers, and so on. This is sketched in Fig. 5.24. This process is called carrier multiplication and it is much like an atomic fission chain reaction in which neutrons, one of the products of the fission of an atom, can cause the fission of additional atoms.

The analogy between carrier multiplication and a nuclear chain reaction can be extended one step further - carrier multiplication can go "critical." If the electric field is high enough, the secondary carriers, both electrons and holes, generated by the impact ionization of the primary carriers can themselves generate more carriers. At the same time, the primary carriers continue generating more secondary carriers as they drift in the high field region. If the electric field is high enough, a runaway situation or *carrier avalanche* can arise in which the total current grows exponentially. If left unchecked, the current density or the power dissipation can reach levels that cause device destruction. This is called **avalanche breakdown**.

Avalanche breakdown is the dominant breakdown mechanism in microelectronic devices. The importance of understanding and correctly modeling avalanche breakdown cannot be overemphasized. A miscalculation here might well mean severe device degradation or even destruction. Even if the maximum current that can flow through a device is limited by some means, such as an outside circuit, so that device destruction is averted, a mode of operation in which avalanche breakdown occurs in a device is largely useless since large and uncontrolled currents flow [5]. In the design of just about any microelectronic device, the need to avoid entering, even getting close to, avalanche breakdown imposes design criteria that limit the performance of the device. For example, for a given base design, the frequency response of a bipolar transistor is essentially set by the design of the collector. This is also what controls the breakdown voltage of the device. The performance/breakdown trade-off is one of the most important design compromises that a device designer must carefully weigh.

Carrier multiplication, even if mild enough so that it does not lead to avalanche breakdown, is an important phenomenom that demands careful assessment in a device. In devices for analog applications, carrier multiplication becomes accompanied by substantial noise. In MESFETs, it results in unwanted gate current. In floating-body silicon-on-insulator (SOI) MOSFETs, it leads to the so-called "kink" effect.

One might think that carrier multiplication could be an excellent amplification mechanism, to implement a "semiconductor photomultiplier" for electrical and optical signals. This would be the case if impact ionization was only triggered by one type of carrier. Consider, for example, a situation where electrons are flowing through a certain semiconductor region. If the electric field increases to the point that significant electron impact ionization takes place, more electrons are produced. In consequence, the total electron current increases in a manner that is proportional to the incoming one. The trouble is with the holes generated by impact ionization. If along the way they also acquire enough energy from the electric field so that they can produce impact ionization events themselves, the tertiary electrons that are generated are not in phase with the primary electrons that started the process. This "blurs" the signal that is carried by the primary electrons. This is why carrier multiplication is very "noisy". Only if the impact ionization coefficient of one type of carrier is much higher than the other, carrier multiplication can be exploited as a gain mechanism. This can be done in specially constructed "superlattices" of two or more carefully selected semiconductors.

The general mathematical treatment of carrier multiplication needs to use the complete set of equations of Table 5.1. Simplifications are possible in steady state situations in which the electric field is either set by the doping distribution or an externally applied voltage and is not modified by the carrier concentrations. In these circumstances, the continuity equations can be simplified if impact ionization is the dominant generation mechanism and recombination is negligible. This is often reasonable since the high electric field maintains the carrier concentrations low. In the simplified analysis here, we are also going to neglect carrier diffusion. Substituting the generation function for impact ionization from Eq. 4.109 into the continuity equations 5.3 and 5.4, and neglecting the time-dependent and recombination terms, we get:

$$-\frac{dJ_e}{dx} = \frac{dJ_h}{dx} = \alpha_e|J_e| + \alpha_h|J_h| \tag{5.113}$$

where α_e and α_h are, respectively, the electron and hole impact ionization coefficients. The absolute signs are to insure that the generation functions are always positive. The two continuity

[5] There are some exceptions to this. Certain voltage reference circuits are based on diodes biased at avalanche breakdown.

$\frac{\partial \mathcal{E}}{\partial x} \simeq \frac{q}{\epsilon}(N_D - N_A)$ or $\mathcal{E} = \mathcal{E}'$
$J_e \simeq -qnv_e^{drift}$
$J_h \simeq qpv_h^{drift}$
$-\frac{dJ_e}{dx} = \alpha_e\vert J_e\vert + \alpha_h\vert J_h\vert$ or $\frac{dJ_h}{dx} = \alpha_e\vert J_e\vert + \alpha_h\vert J_h\vert$
$\frac{dJ_t}{dx} = 0$
$J_t = J_e + J_h$

Table 5.7: Equation set for one-dimensional steady-state situation with imposed electric field and where carrier multiplication is the dominant generation mechanism.

equations are actually redundant, since in steady state the total current $J_t = J_e + J_h$ is independent of position. Table 5.7 summarizes the carrier equation set for steady-state carrier multiplication under a fixed electric field.

The best way to conceptually understand carrier multiplication and its implications is to study an example.

5.10.1 Example 5: Carrier multiplication in a high-resistivity region with uniform electric field

Consider a high resistivity uniformly-doped region again under an external electric field, as shown in Fig. 5.25. A narrow but penetrating beam of light impinges on the right-hand side of the sample generating g_l electron-hole pairs per cm^2 every s. This is a case similar to the example discussed in Section 5.9.1, but with $x_o = L$. Following the analysis carried out in Section 5.9.1, the holes are instantly collected at the metal contact but the electrons drift through the high-resistivity region until they reach the other end of the structure.

In the absence of significant carrier multiplication, the current through the high-field region is entirely supported by electrons and is given by $J_t = J_e = qg_l$. If carrier multiplication is taking place, as the photogenerated electrons drift down the sample, they generate secondary electrons and holes which can also, in turn, generate more carriers. In these conditions, the current profile is obtained by solving the differential equation 5.113. Since with our choice of axis and electric field, both J_e and J_h are positive, we have:

$$-\frac{dJ_e}{dx} = \alpha_e J_e + \alpha_h J_h \tag{5.114}$$

It is advantageous to write this equation in terms of the total current J_t:

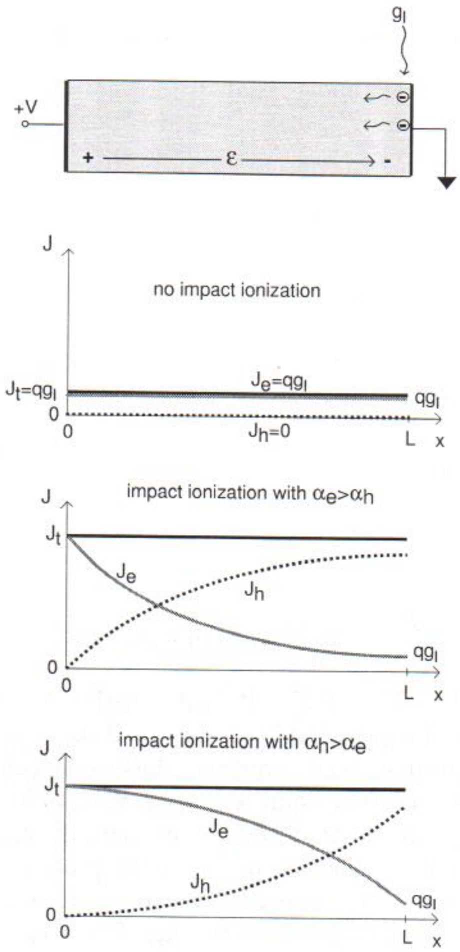


Figure 5.25: Top: sketch of a high-resistivity uniformly-doped sample with carrier multiplication started by a sheet of electrons generated at one end. Below, the resulting electron, hole and total currents are sketched in both cases $\alpha_e > \alpha_h$ and $\alpha_h > \alpha_e$.

$$-\frac{dJ_e}{dx} = (\alpha_e - \alpha_h)J_e + \alpha_h J_t \tag{5.115}$$

This equation has to be solved subject to the boundary condition that $J_e(L) = qg_l$.

The solution to this differential equation is:

$$J_e(x) = \frac{\alpha_h J_t}{\alpha_h - \alpha_e}[1 + e^{(\alpha_h - \alpha_e)x}] + Ae^{(\alpha_h - \alpha_e)x} \tag{5.116}$$

where A is a constant to be determined and J_t is unknown.

A is obtained by demanding that at $x = 0$, the total current is solely supported by electrons.

This is because the sign of the electric field prevents any holes from being collected at that contact. Hence, $J_e(x=0) = J_t$, which substituted in Eq. 5.116 results in $A = J_t(\alpha_e + \alpha_h)/(\alpha_e - \alpha_h)$.

We must also demand, as mentioned above, that $J_e(x=L) = qg_l$. This gives the expression of J_t:

$$J_t = qg_l \frac{\alpha_h - \alpha_e}{\alpha_h - \alpha_e e^{(\alpha_h - \alpha_e)L}} \tag{5.117}$$

Plugging A and J_t into Eq. 5.116 yields the final result:

$$J_e(x) = qg_l \frac{\alpha_h - \alpha_e e^{(\alpha_h - \alpha_e)x}}{\alpha_h - \alpha_e e^{(\alpha_h - \alpha_e)L}} \tag{5.118}$$

which satisfies the boundary conditions.

The hole current is simply:

$$J_h = J_t - J_e = qg_l \frac{\alpha_e[e^{(\alpha_h - \alpha_e)x} - 1]}{\alpha_h - \alpha_e e^{(\alpha_h - \alpha_e)L}} \tag{5.119}$$

All the currents are sketched in Fig. 5.25. It is interesting to discuss the shape of these currents depending on the relative magnitude of α_e and α_h. If $\alpha_e > \alpha_h$, electrons tend to be the source of most impact ionization events. In consequence, the electron current grows exponentially as the electrons travel from the generation point at $x = L$ to $x = 0$. If $\alpha_h > \alpha_e$, the contrary situation takes place. It is the hole current that grows exponentially since they are responsible for the bulk of the generation events. In both cases, however, the process is started by the electrons generated at $x = L$. In both cases too, the contribution to the total current carried out by the electrons increases from right to left as required by the sign of the electric field.

It is interesting to define a **multiplication coefficient**, M, as the ratio of the total current flowing through the structure to the current that started the multiplication process $J_e(L) = qg_l$:

$$M = \frac{\alpha_h - \alpha_e}{\alpha_h - \alpha_e e^{(\alpha_h - \alpha_e)L}} \tag{5.120}$$

There are two limits to M. If the electric field is small, both ionization coefficients are very small. This allows us to linearize the exponential in the denominator of M and obtain $M \simeq 1$. In this case, carrier multiplication is negligible.

At the other extreme, at high electric fields, since both terms in the denominator of Eq. 5.120 are positive, it is entirely possible for the denominator to go to zero and for M to diverge. When this happens, J_t grows without bounds and *avalanche breakdown* takes place. In the example considered here, this occurs when:

$$(\alpha_h - \alpha_e)L = \ln \frac{\alpha_h}{\alpha_e} \tag{5.121}$$

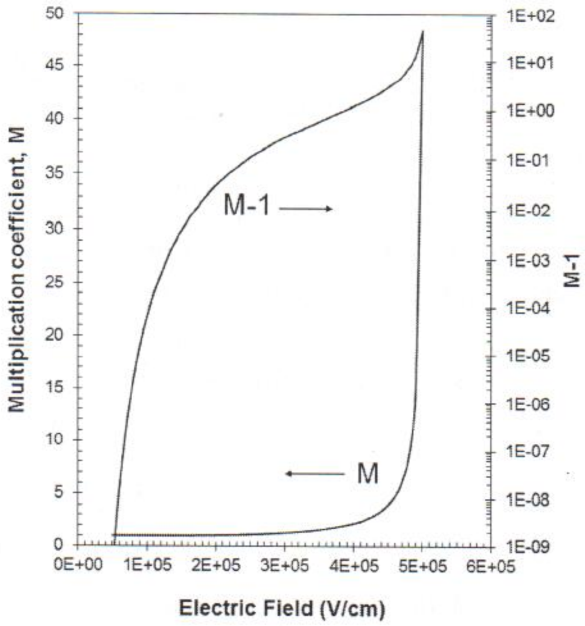


Figure 5.26: Multiplication coefficient M, and $M-1$ as a function of electric field for a 0.2 μm long Si sample with uniform electric field.

For a sample of given length L, this condition can be reached by increasing the voltage applied to the sample and hence the electric field. If $\alpha_h > \alpha_e$, for example, when the field is small, both impact ionization coefficients are small and $(\alpha_h - \alpha_e)L$ is very small. As the field increases, both coefficients increase and their difference increases too. At a certain voltage, the condition 5.121 is satisfied and the sample goes into breakdown. The same happens if $\alpha_e > \alpha_h$. The voltage at which this occurs is the breakdown voltage, BV.

Fig. 5.26 shows an example of a calculation of the multiplication coefficient for a 0.2 μm long Si sample under a uniform electric field. At an electric field of $\mathcal{E}_b = 5 \times 10^5\ V/cm$ the sample goes into avalanche breakdown. This is called the *critical breakdown field*. The breakdown voltage of this sample is $BV = 10\ V$.

The critical electric field is not a fundamental parameter of a material. Depending on the field distribution and the sample size, the critical field can be rather different. The only reliable way to compute the breakdown voltage in a given situation is to carry out a similar analysis to the one presented in this example.

It often occurs that we are interested in impact ionization in situations in which carrier multiplication is not very significant. This is the case, for example, when analyzing the substrate current in MOSFETs. In these cases, the multiplication coefficient is not a good parameter since its value is very close to unity for a wide range of electric fields. Instead, it is more useful to focus on $M-1$, which reflects only the current produced by carriers generated through impact ionization. $M-1$ depends very strongly on the electric field. Fig. 5.26 shows $M-1$ as a function of electric field for the example discussed in this section.

Exercise 5.6: *Consider a 0.2 μm long high-resistivity region of a Si sample under an applied electric field of $\mathcal{E} = 5 \times 10^5\ V/cm$ at room temperature. a) Calculate the multiplication coefficient. b) If there is a triggering current of 1 μA, calculate the impact ionization current that is produced and the total current that flows through the sample. c) If we now increase the length of this high-resistivity region without changing the electric field, calculate the critical length that results in avalanche breakdown.*

a) The first step in solving this problem is to get the appropriate values of the ionization coefficients from the analytical expressions given in Appendix E. For an field of $\mathcal{E} = 5 \times 10^5\ V/cm$, the ionization coefficients are $\alpha_e = 6.0 \times 10^4\ cm^{-1}$ and $\alpha_h = 3.9 \times 10^4\ cm^{-1}$.

We can now plug into the expression of the multiplication coefficient:

$$M = \frac{\alpha_h - \alpha_e}{\alpha_h - \alpha_e e^{(\alpha_h - \alpha_e)L}} = \frac{3.9 \times 10^4 - 6.0 \times 10^4}{3.9 \times 10^4 - 6.0 \times 10^4 e^{(3.9 \times 10^4 - 6.0 \times 10^4)0.2 \times 10^{-4}}} \simeq 50$$

b) The impact ionization current is given by the triggering current times $M - 1$. Then:

$$I_{ii} = (M-1)I_{trig} = (50-1) \times 1\ \mu A = 49\ \mu A$$

The total current that flows through the sample is M times the triggering current:

$$I_t = MI_{trig} = 50 \times 1\ \mu A = 50\ \mu A$$

c) Avalanche breakdown occurs when the following condition is met:

$$(\alpha_h - \alpha_e)L_{crit} = \ln \frac{\alpha_h}{\alpha_e}$$

Solving for L_{crit}:

$$L_{crit} = \frac{1}{\alpha_h - \alpha_e} \ln \frac{\alpha_h}{\alpha_e} = \frac{1}{3.9 \times 10^4 - 6.0 \times 10^4} \ln \frac{3.9 \times 10^4}{6.0 \times 10^4} = 0.21\ \mu m$$

5.11 Summary

- The systems of equations of Table 5.1 describe carrier behavior in semiconductors. These equations can be solved exactly using CAD tools in three dimensions. Appropriate simplifications for a few important families of problems can yield key insight regarding the dominant physics.

- Concept of quasi-neutrality: absence of substantial volume charge. Quasi-neutrality implies that the divergence of the total current density is zero everywhere. Quasi-neutrality is more likely the higher the mobile carrier concentrations. Quasi-neutrality is a pervasive situation in semiconductor devices.

- Majority-carrier situations: characterized by the application of an external voltage without perturbing the carrier concentrations from their equilibrium values. In this family of problems, minority carriers play no role. The current through the semiconductor is determined by majority-carrier drift. The relevant time constant for majority-carrier situations is the dielectric relaxation time which is often much shorter than the time scale of interest.

- Minority-carrier situations are those with bottlenecks dominated by minority carrier phenomena, typically diffusion, drift, recombination and generation. Dominant processes can be identified by estimating the magnitude of the key length scales: diffusion length and the sample length. In dynamic problems, key time constants are the carrier lifetime and the transit time. The identification and estimation of the dominant length scale and time scale of minority carrier problems is a very valuable skill of great practical significance to microelectronic device engineers.
- Space-charge regions: regions with very low carrier concentrations, high resistivity and a substantial electric field. Carrier drift is the most significant transport mechanism.
- Impact ionization in space charge regions produces carrier multiplication and can result in avalanche breakdown. This often limits the maximum voltage that can be handled by a device (the breakdown voltage). An ability to estimate the breakdown voltage and an understanding of its leading dependencies are essential skills for a device engineer.

5.12 Further reading

This chapter has covered a lot of territory. There is no single book with a treatment that is parallel to the one here. Several books, however, contain complementary material.

A First Course in Differential Equations with Applications by D. G. Zill, Prindle, Weber and Schmidt, 1979 (ISBN 0-87150-266-6, QA372.Z54). This is a text on the theory, methods of solution and applications of ordinary differential equations. This book is very clearly written and contains many examples from different fields of physics and engineering. The techniques taught in this book come handy to solve many different types of minority-carrier problems.

Conduction of Heat in Solids by H. S. Carslaw and J. C. Jaeger, Oxford University Press, 1959 (ISBN 0-19-853303-9, QC321.C321). The mathematics of minority carrier transport share a lot with that of conduction of heat in solids. This classical text (first published in 1946) presents a very complete mathematical treatment of heat conduction in solids in several coordinate systems. This might be useful in some specific circumstances since many minority carrier-type problems have a dual heat conduction problem.

AT5.1 Continuity equations in integral form

The continuity equations derived in Section 5.1 are particularly intuitive when written in an integral form. We can easily do this by extending Eq. 5.1 to a finite volume V enclosed by a surface S:

$$\frac{\partial}{\partial t}\int_V n dV = \int_V (G-R)dV + \frac{1}{q}\int_S \vec{J_e}.d\vec{S} \tag{5.122}$$

Similarly for holes:

$$\frac{\partial}{\partial t}\int_V p dV = \int_V (G-R)dV - \frac{1}{q}\int_S \vec{J_h}.d\vec{S} \tag{5.123}$$

These equations clearly state that the total number of carriers inside a certain volume changes in time if there is net carrier generation or if there is net flow of carriers through its surface.

The continuity equation for charge is also very intuitive when expressed in an integral form. If we multiply both of the above equations by q and we subtract the first one from the second one and use Eq. 5.5, we get:

$$\frac{\partial}{\partial t}\int_V \rho\, dV = -\int_S \vec{J_t}.d\vec{S} \tag{5.124}$$

This now makes clear that if there is net charge increasing in time in a certain volume of a semiconductor it is because there is net current flowing into that volume (negative surface integral).

In steady state, the time derivatives in Eqs. 5.122, 5.123, 5.124 become zero. From Eq. 5.124 we can conclude that there cannot be net current into any region of the semiconductor. From Eqs. 5.122 and 5.123, if we substitute $G_{ext} - U$ for $R - G$, we can write:

$$-\int_S \vec{J_e}.d\vec{S} = \int_S \vec{J_h}.d\vec{S} = q\int_V (G_{ext} - U)dV \tag{5.125}$$

This states that in steady state, the integral of the carrier current extended over the surface of an enclosed volume is equal (with the appropriate sign) to the net generation of carriers taking place inside the enclosed volume. This is an observation that derives directly from the continuity equations and that comes handy in a few situations.

AT5.2 Dielectric relaxation

A uniformly doped semiconductor in thermal equilibrium, in the absence of any surface effects, is charge neutral at every point in space. Suppose that through some external influence we can