4

Generation and Recombination processes

4.1 Introduction

The concentrations of mobile charges (electrons and holes) in semiconductors are governed by statistical laws that provide their equilibrium average values. Equilibrium means that the average values are maintained by a continuous fluctuations of mobile charges that are subject to creation and annihilation phenomena. Such events, at least involve the transition from states in the conductance band, the valence band and the doping atoms energy levels.
These processes, called generation and recombination, are of outmost importance for the properties of the devices. They operate in order to maintain the equilibrium condition, represented by the law mass action, when additional charges are injected in a volume of the material.
The generation of electrons in the conduction band and holes in the valence band occurs in intrinsic and doped semiconductors with an important difference: in intrinsic semiconductor electrons and holes are generated together, while in doped semiconductors to the generation of holes and electrons corresponds the generation of a fixed counter charge. It is straightforward that the process of generation is reversible and then electrons and holes can recombine together in both intrinsic and doped materials.
Then, the density of charges in a volume may change as a consequence of both the current and the generation and recombination events. Then total variation of charges in a material volume is described by the continuity equation that is a useful tool to determine the charges balance.

4.2 The continuity equation

Let us consider a volume of a semiconductor extended from x to x+dx and with a section of area A. A current (e.g. of electrons) $J_n(x)$ is injected in the volume through the area A at the coordinate x, and a current $J_n(x+dx)$ leaves the volume at the coordinate x+dx. Inside the volume, generation and recombination phenomena may also occur at rates G_n and R_n respectively. The continuity equation describes the time behaviour of the total amount of charges (N) inside the volume. This is defined as:

$$N = \frac{\partial n}{\partial t} \cdot A dx \tag{4.1}$$

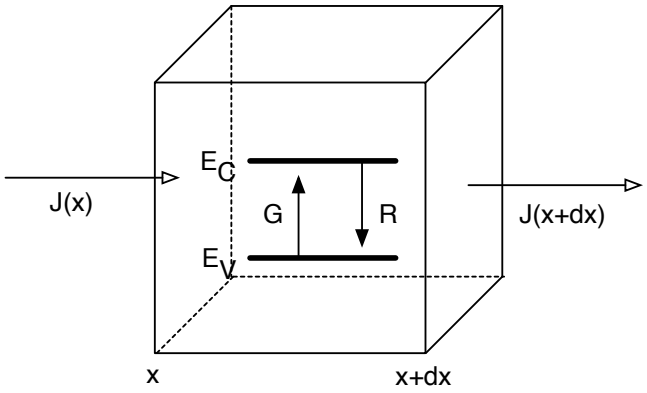


Fig. 4.1. *Charge balance in a volume of material. Due to the current the charges can enter and leave the volume, furthermore they can be generated and recombined.*

in case of electrons, the balance of charges can be written as:

$$\frac{\partial n}{\partial t}Adx = \frac{J_n(x)}{-q}A - \frac{J_n(x+dx)}{-q}A + G_nAdx - R_nAdx \tag{4.2}$$

After replacing $J_n(x+dx) - J_n(x) = \frac{\partial J_n}{\partial x}dx$, the equation is normalised by Adx so it does not depend on the volume. Thus the continuity equation for electrons is:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + (G_n - R_n) \tag{4.3}$$

symmetrically, the continuity equation for holes is:

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} + (G_p - R_p) \tag{4.4}$$

These formulas can be further expanded replacing the current that, in a semiconductor, is the sum of drift and diffusion currents. For electrons: $J_n = q\mu_n n\mathcal{E} + qD_n\frac{dn}{dx}$ then the explicit form of the continuity equation is:

$$\frac{\partial n}{\partial t} = \mu_n n\frac{\partial \mathcal{E}}{\partial x} + \mu_n\mathcal{E}\frac{\partial n}{\partial x} + D_n\frac{\partial^2 n}{\partial x^2} + (G_n - R_n) \tag{4.5}$$

while for holes:

$$\frac{\partial p}{\partial t} = -\mu_p p\frac{\partial \mathcal{E}}{\partial x} - \mu_p\mathcal{E}\frac{\partial p}{\partial x} + D_n\frac{\partial^2 p}{\partial x^2} + (G_p - R_p) \tag{4.6}$$

The previous set of equations describes the continuity of the charges when the mobility and the diffusion coefficient are constant. It is worth to remind that this is the case of a uniformly doped material. A typical situation is the steady state condition that corresponds to the time invariant concentration of charges ($\frac{\partial n}{\partial t} = 0$).

4.3 Generation and recombination phenomena

The generation and the recombination of mobile charges are peculiar phenomena in semiconductors. These processes are manifested as the production of a pair of mobile charges (generation) or their disappearance (recombination). Ultimately, this implies the transition of charges between the valence and the conductance band.
The main action of generation and recombination processes is to restore the equilibrium, namely the action mass law. The equilibrium condition is perturbed anytime the charges are altered by an external cause. Typical cases are the illumination with a flux of photons and the injection or extraction of a current from or to a nearby volume of the material.
The transition of an electron from the valence to the conduction bands requires an amount of energy at least equal to the energy band gap but in some cases also a change of the momentum . The change of momentum is not relevant in those materials where the top of the valence band coincides, in the k space, with the bottom of the conduction band. Otherwise, when the maximum and minimum of the two bands are not coincident in the space k, the transition requires a momentum change ($\Delta k \neq 0$).
$\Delta k = 0$ is the case of direct band-gap which is found in semiconductors formed by III-V elements of the periodic table such as Gallium Arsenide (GaAs) or Aluminium Arsenide (AlAs). $\Delta k \neq 0$ is the case of an indirect band gap typical for semiconductors of the IV group such as silicon and germanium.
The probability of events where energy and momentum simultaneously change is small. The variation of energy and momentum requires a sort of a three bodies interaction. In practice the additional momentum necessary for the transition is provided by the lattice vibration. Lattice vibrations in crystals are represented as quasi-particles called phonons which are endowed with energy and momentum.
The small probability indicates that transitions from one band to the other occur at a slow rate, then they are not efficient to maintain the equilibrium in case of fast events such as the illumination with a pulse of light. In these materials the generation and recombination actually occur with the assistance of impurity states located in the band gap.
Doping is a typical source of states in the band gap. Doping states are close to the respective band according to the character of the dopant: either acceptor or donor. Besides dopants, that are intentionally added, a number of naturally occurring impurities are also found in the crystal, they may introduce additional electron energy states in the band gap. Thus the energy gap is not totally forbidden.
Inter gap states may be found at any energy level, an example of them are the surface states. In the bulk of the material the density of inter gap states is smaller respect to the surface but it is not negligible. It is important to consider that defects in real materials are inevitable, then a density of states inside the band gap always exists even in intrinsic semiconductors.
In the rest of this section the generation and recombination phenomena driven by inter gap states are discussed.

4.3.1 Generation and recombination rates

In order to derive a quantitative description of generation and recombination, let us consider a semiconductor characterized by a density of inter gap states (N_T) at the same energy (E_T). These states are spatially localized and they are the manifestation of the defects of the material (impurities or lattice defects). In acceptable materials, the defects are so sparse that their states are non

interacting. Then, like the doping states, even the defect states do not degenerate into bands.
The interaction of the defect states with the electrons of the conduction band and the holes of the valence band is detailed by four different processes.
Each process is described by the rate of occurrence whose dimensions are $\frac{events}{time\ cm^3}$:
r_1: capture of electrons;
r_2: emission of electrons;
r_3: capture of holes;
r_4: emission of holes.

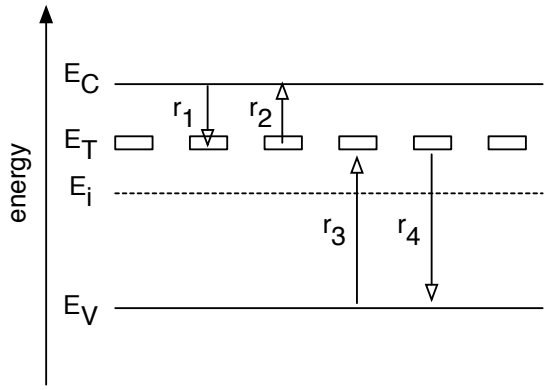


Fig. 4.2. *Arrangement of the defect states at the energy level E_T inside the band gap. The four processes of capture and emission of electrons and holes are shown.*

Let us calculate the rate of capture and emission in the case of electrons. The same procedure holds to determine the rates of the processes involving the holes.

The rate of capture

Since the defects are spatially localized, the capture may occur only after the physical encounter of an electron with the defect. Given a volume of material let us consider a flux of electrons impinging into the volume. The rate of captured electrons is the product of the flux of incoming electrons times the density of empty capture states times the intrinsic capability that a state can capture an electron. This last quantity (indicated as σ) can be considered as the affinity between the electron and the defect, and it depends on the nature of the defect itself.

$$r_1 = F \cdot N_T(1 - f_{FD}(E_T)) \cdot \sigma_n \tag{4.7}$$

The flux of electrons is the total current, however since the thermal velocity is much greater than the velocity of electrons due to drift and diffusion, the amount of flux of electrons onto a surface inside the material is due to the thermal current. Indeed, at microscopic level and even under either

an electric field or a concentration gradient, the instantaneous speed of the electrons is the thermal velocity. The flux of electrons impinging onto a plane is calculated by the classical kinetic theory of gas where the flux is proportional to the thermal velocity according to the Knudsen law : $F = \frac{1}{4} v_{th} n$ where n is the density of electrons. The average thermal velocity v_{th} is calculated from the Maxwell distribution of velocity $v_{th} = \sqrt{\frac{8kT}{\pi m^*}}$.
The density of empty capture states is the total density of capture states times the probability that they are not occupied by electrons. This probability is the complementary Fermi-Dirac function at the energy of the states: $N_{T,empty} = N_T(1 - f_{FD}(E_T))$. Thus, the rate of capture of electrons is:

$$r_1 = \frac{1}{4} v_{th} n \cdot \sigma_n \cdot N_T(1 - f_{FD}(E_T)) \tag{4.8}$$

The quantity σ_n is the affinity between the impurity and the electron. It is called the cross-section of capture and it has the dimension of an area. Sometimes σ_n is described as a sort of the equivalent area of a target, then largest is the area easiest is the capture. Anyway, it defines the occurrence of the capture event without any relation with the actual area of the defect. σ_n of gold atoms, a typical capture state in silicon, is of the order of $10^{-15} cm^2$. Beryllium has the largest cross section of capture in silicon $\sigma_n = 10^{-10} cm^2$. The density of defects, always in silicon, is of the order of $10^{-15}\ cm^{-3}$.

The rate of emission

The emission is the release, back to the conduction band, of previously captured electrons. The rate of emission is simply proportional to the density of electrons in the capture states times the intrinsic probability of emission. Differently from the previous case, since the conduction band has a very large density of free states the density of states at destination is irrelevant. The density of filled capture states is given by the Fermi-Dirac function: $N_{T,fill} = N_T f_{FD}(E_T)$. The rate of emission of electrons is:

$$r_2 = N_T f_{FD}(E_T) \cdot e_n \tag{4.9}$$

Where e_n is the intrinsic probability of emission. It has the dimension of the inverse of a time, and it corresponds to the inverse of the average lifetime of an electron in a capture state.
It has to be noted that σ_n and e_n are not exactly probabilities (they should bd dimensionless and bounded between 0 and 1) but they rather describe the chance of occurrence of the respective phenomena.

Equilibrium

The cross-section of capture and the constant of emission are one each other related, and their relationship defines the effect of the capture state on the population of charge carriers. In steady state condition, the mean densities of electrons and holes are constant, then the rates of capture and emission are the same: $r_1 = r_2$.

$$\frac{1}{4} v_{th} n \cdot \sigma_n \cdot N_T(1 - f_{FD}(E_T)) = N_T f_{FD}(E_T) \cdot e_n \tag{4.10}$$

from which e_n is calculated:

$$e_n = \frac{1}{4} v_{th} n \cdot \sigma_n \cdot \left(\frac{1}{f_{FD}(E_T)} - 1 \right) \tag{4.11}$$

Since the capture state is in the band gap, the Boltzmann approximation of the Fermi-Dirac function is not valid and the complete form of the Fermi-Dirac function has to be used. This leads to:

$$e_n = \frac{1}{4} v_{th} n \cdot \sigma_n \cdot exp\left(\frac{E_T - E_F}{kT} \right) \tag{4.12}$$

The constant of emission depends on the distance between the energy level of the defect state and the Fermi level.
Replacing the Fermi level with: $E_F = E_i + E_F - E_i$ (where E_i is the intrinsic Fermi level) and the density of electrons with $n = n_i exp(\frac{E_F - E_i}{kT})$, we obtain:

$$e_n = \frac{1}{4} v_{th} \cdot \sigma_n \cdot n_i \cdot exp\left(\frac{E_T - E_i}{kT} \right) \tag{4.13}$$

Namely, the emission does not depend on the Fermi level, but rather it depends on the energy of the defect with respect to the intrinsic Fermi level namely with respect to the centre of the band gap.
Similar results are obtained for the rates of the processes involving the holes.

$$r_3 = \frac{1}{4} v_{th} p \cdot \sigma_p \cdot N_T f_{FD}(E_T) \tag{4.14}$$

$$r_4 = N_T (1 - f_{FD}(E_T)) \cdot e_p \tag{4.15}$$

The equilibrium condition ($r_3 = r_4$) leads to the expression of the constant of emission of holes:

$$e_p = \frac{1}{4} v_{th} \cdot \sigma_p \cdot n_i \cdot exp\left(\frac{E_i - E_T}{kT} \right) \tag{4.16}$$

In conclusion, the chance of emission of electrons increases as the level of the defect is close to the conduction band, and the chance of emission of holes increases as the level of the defect is close to the valence band.
As a consequence, the same capture state if not located at the centre of the band gap behaves differently respect to electrons and holes.

4.3.2 Traps and recombination centers

Although the probability of capture of electrons and holes may be equal, their emission rates are different because they depend on the position of the capture state with respect to the intrinsic Fermi level. Of course, electrons and holes are re-emitted with the same rate only when $E_T = E_i$.
At the equilibrium the concentration of electrons and holes are stationary, this condition is achieved when: $r_1 = r_2$ and $r_3 = r_4$.
Let us suppose that, for some reason, the concentration of holes suddenly increases. Then the rate of capture of holes (r_3) increases. As a consequence, the rate of emission of holes (r_4) should increase. However, the capture of holes affects also the rate of capture of electrons. Indeed, an increase of

holes in the defect states means an increase of empty states and then an increase of the rate of capture of electrons (r_1). The processes r_3 and r_1 are exclusive, thus the consequence of the capture of a hole is either the capture of an electron or the release of a hole. Which of them prevails depend on the magnitude of the related rates and there are two extreme situations: $r_1 \gg r_4$ and $r_4 \gg r_1$.
In the first case, at the end of the process the number of electrons in the capture states does not change, the hole in excess is eliminated, and one electron disappears from the conductance band.
In the second case, at the end of the process the number of electrons in the capture states does not change, the excess hole is only temporarily removed and released back to the valence band, while the electrons in conductance band are not affected by the process.
Then, if $r_1 \gg r_4$ the excess hole is removed by one electron of the conductance band and such a state is called a *recombination center*. If $r_4 \gg r_1$ the excess hole is captured and re-emitted after a delay time, and such a state is called a *trap*.
In other words, traps respond to the variation of the concentration of holes restoring the pristine value, while in case of recombination centers the concentration of holes is still restored to the equilibrium value but the concentration of electrons decreases.
In order to behave as a trap the probability of emission of a hole has to be larger than the probability of capture of an electron. e_n depends on the distance between the energy of the defect state from the intrinsic Fermi level. The largest rate of emission of holes is exhibited by those states whose energy is close to the top of the valence band, and, symmetrically, the states more close to the conduction band have the largest probability to emit electrons. These conditions are fulfilled by the acceptor and donor states. Thus, it can be concluded that dopant atoms are traps while recombination centers correspond to different kinds of impurities and defects.

The effect of traps and recombination centers to the current

The presence of traps or recombination centers affects the charge transport. To illustrate the effect of recombination centers, let us consider the case of a volume of a semiconductor populated by a density of recombination centers and that these centers are filled with electrons. When the current of holes is injected, the holes in exces are recombined (they disappear) but also the concentration of electrons decreases. As a consequence, a steady current cannot exist unless the disappeared electrons are supplied by the adjacent portion of material. If this process is possible, the current of holes is transformed into a current of electrons flowing in the opposite direction. Since the charges have different sign, a net current is observed.
Figure 4 shows the role of recombination centers in the continuity equation.

The case of traps is shown in figure 4.5. In this case the current of holes is not modified. In practice the traps extend the transit time of the holes in the volume, then traps participate to the definition of the mobility. Since donors and acceptors are traps, this is consistent with the fact that the mobility decreases with the increase of doping.
The defects resulting in electronic states inside the band gap are important elements of the semiconductors. They have two extreme effects on the mobile charges concentrations: traps and recombination centers.
Traps affects the transit time of charges in the volume, then traps are an important constituents of the mobility of the material. On the other hand, the effect of the recombination center is completely different. Recombination centers remove the excess of one type of charges consuming the concentration of the other charge carriers. If the decreased charges can be replaced by an external source (typically a ohmic contact) the recombination centers transform the current of one charge

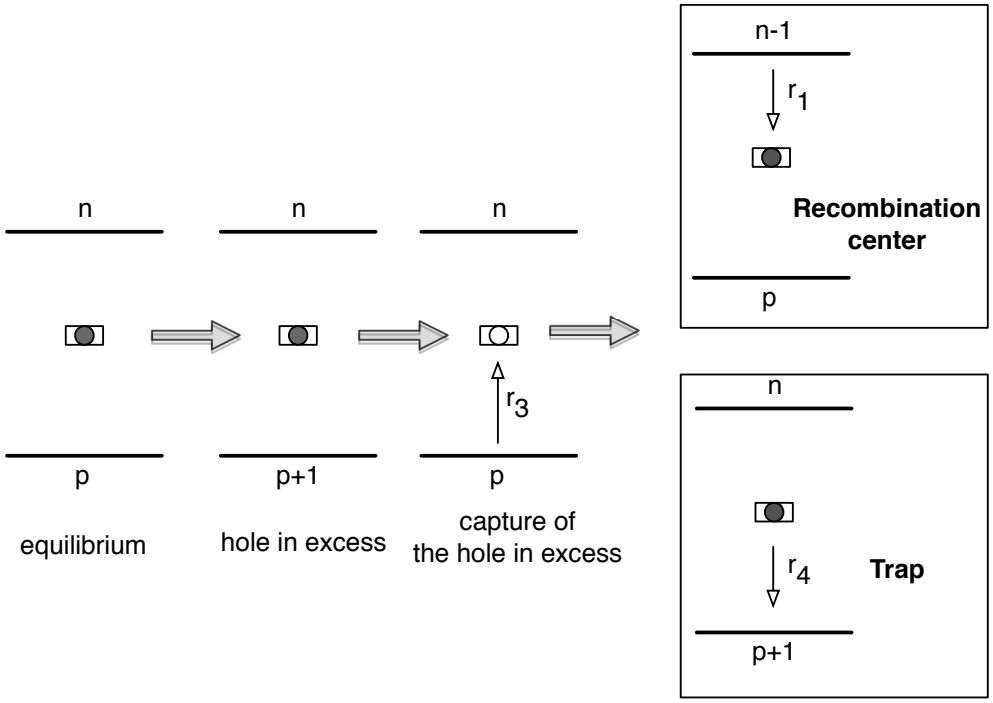
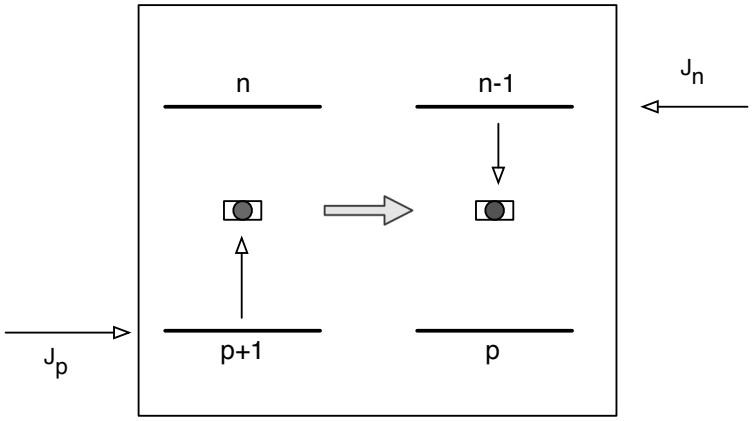


Fig. 4.3. *The reaction of trap and recombination centre reaction to an excess of holes. The black dot in the capture state indicates an electron while the white dot indicates a hole.*

carrier into the current of the other charge carrier.
This mechanism of charge transport is fundamental for instance in PN junctions where the current injected from one side of the junction to the other is formed by only one of the two charge carriers. For this reason it is important to define an analytical tool allowing to treat the recombination and generation phenomena in a semiconductor.

4.4 The Shockley-Hall-Read model of recombination

In doped semiconductors, the processes of generation and recombination can be easily modelled. As noted in the above section, the energy level of the recombination centers is distant from the conduction and the valence bands. Thus, in a N-type semiconductor, the energy level of the recombination centers likely lie below the Fermi level. And oppositely, in case of P-type semiconductors it lies above the Fermi level. As a consequence, In a N-type semiconductor, at such a level, $f_{FD} \approx 1$ and in a P-type semiconductor $1 - f_{FD} \approx 1$. Thus, the centers of recombination are always filled with majority charges.
In a N-type semiconductor, the equilibrium conditions ($r_1 \gg r_3$ and $r_2 \gg r_4$ and $r_1 = r_2$ and $r_3 = r_4$) are simultaneously fulfilled even if the holes are few, and their capture is a rare event. In such a condition, any increase of concentration of holes elicits a large increase of r_3 while to

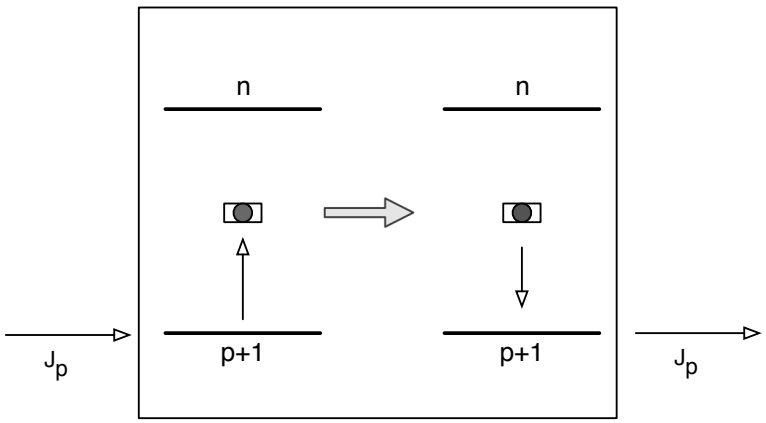


steady-state case

$$\text{holes}: \frac{\partial p}{\partial t} = 0 \rightarrow R_p = 1; G_p = 0 \rightarrow \frac{\partial J_p}{\partial t} = -q$$

$$\text{electrons}: \frac{\partial n}{\partial t} = 0 \rightarrow R_n = 1; G_n = 0 \rightarrow \frac{\partial J_p}{\partial t} = +q$$

Fig. 4.4. *The continuity equations (eq. 4.3 and 4.4) describe the effect of the recombination centers on a current of holes injected in the volume. The current of holes has a negative gradient then it decays in the volume while a current of electrons has a positive gradient then it grows in the volume as the holes disappear.*



steady-state case

$$\text{holes}: \frac{\partial np}{\partial t} = 0 \rightarrow R_P = 1; G_p = 1 \rightarrow \frac{\partial J_p}{\partial t} = 0$$

Fig. 4.5. *The effect of traps on current of holes injected in the volume can be appraised by means of the continuity equation. The current of holes has a null gradient then it does not change across the volume.*

maintain the equilibrium only a modest, practically negligible, change of r_1 is required. Eventually, the material is ready to recombine any change of the minority charges.
An additional argument can be obtained considering the condition for which a capture state behave as a trap or as a recombination center for a hole. It has been seen in the previous section that a trap occurs when $r_4 \gg r_1$, let us evaluate than the ratio between the two rates:

$$\frac{r_4}{r_1} = \frac{N_T(1 - f_{FD}(E_T)) \cdot e_p}{\frac{1}{4} v_{th} n \cdot \sigma_n \cdot N_T(1 - f_{FD}(E_T))]} = \frac{\frac{1}{4} v_{th} \cdot \sigma_p \cdot n_i \cdot exp(\frac{E_i - E_T}{kT})}{\frac{1}{4} v_{th} n \cdot \sigma_n \cdot N_T(1 - f_{FD}(E_T))]} \tag{4.17}$$

If the semiconductor is doped, $n = N_D$ and the holes are the minority charges. When $\sigma_n = \sigma_p$ (this condition will be discussed again below) The ratio between the ratio of holes emission and electrons capture is:

$$\frac{r_4}{r_1} = \frac{n_i \cdot exp(\frac{E_i - E_T}{kT})}{N_D} \tag{4.18}$$

Then the position of the energy level of the capture state with respect to the intrinsic Fermi level is:

$$E_i - E_T = kT \cdot ln\left(\frac{r_4}{r_1}\frac{n_D}{N_i}\right) \tag{4.19}$$

If $N_D = 10^{16}\ cm^{-3}$ in order to have $r_4 > 10 \cdot r_1$ $E_i - E_T > 478\ meV$. In silicon, this means that only those states whose energy is less than 80 meV above the valence band behave as traps of holes. This is a negligible portion of the band gap, then all the states in the band gap are recombination centers for the holes. Of course the same argument holds for electrons in a P type material. Then we can conclude that, in practice, the totality of the states in the band gap are recombination centers for the minority charges.
The difference between the rates of recombination and generation is the generation-recombination function (U).

$$U = R - G = r_1 - r_2 = r_3 - r_4 \tag{4.20}$$

U can be calculated from the definition of the four rates (eq. 4.7 and 4.8) and the chance of emission (eq. 4.11 and 4.12). The calculation introduces the following relevant quantities:

$$\tau_{n0} = \frac{1}{N_t v_{th} \sigma_n};\ \tau_{p0} = \frac{1}{N_t v_{th} \sigma_p} \tag{4.21}$$

These are the times of recombination for electrons and holes. These quantities define the typical time scale of capture of the excess charges. They depend on the temperature (v_{th}), the cross-section of capture , and the concentration of the recombination centers.
The calculation results in the recombination function U (a detailed description of the calculus is shown in the appendix to this chapter):

$$U = \frac{np - n_i^2}{\tau_{n0}[p + n_i exp(\frac{E_i - E_T}{kT})] + \tau_{p0}[n + n_i exp(\frac{E_T - E_i}{kT})]} \tag{4.22}$$

The numerator of U is different from zero if $np \neq n_i^2$ namely when the material is in non equilibrium. The function U reacts to a non equilibrium condition and the denominator expresses the

magnitude of the reaction.
The capture of an electron corresponds to the fact that a wandering electron is bonded to a previously empty orbital of the impurity atom, while the capture of a hole is the transfer of an electron previously bond to a orbital of the impurity to a nearby silicon atom. Since the recombination event always involve a pair of electrons and holes, the probability of the processes are rather similar, so we can assume $\sigma_n = \sigma_p = \sigma$ and th $\tau_{n0} = \tau_{p0} = \tau_0$. From eq. 4.20, considering typical defects density around $10^{-15}\ cm^{-3}$ and a cross-section for electrons and holes capture $\sigma = 10^{-15}\ cm^2$ we can calculate that $\tau_0 \approx 100\ ns$.
The recombination function can then be simplified as

$$U = \frac{np - n_i^2}{\tau_0[p + n + 2n_i cosh(\frac{E_T - E_i}{kT})]} \tag{4.23}$$

Where the definition of hyperbolic cosine has been used ($coshx = \frac{e^x + e^{-x}}{2}$).
U gets its maximum value when the hyperbolic cosine is at the minimum, and this condition is met when $E_T = E_i$. This means that the maximum efficiency of the recombination is achieved when E_T is close to E_i, namely when the energy level of the recombination centers is around to the middle of the band gap. As a consequence, in doped semiconductors the most efficient states are always filled with majority charges.
Typical centers of recombination in silicon are atoms of gold and copper whose energy levels are about 0.03 eV and 0.01 eV above the the intrinsic Fermi level respectively.
The function U describes the reaction of the semiconductor to non equilibrium situations: when $np > n_i^2 \rightarrow U > 0$ the recombination prevails and on the other hand, when $np < n_i^2 \rightarrow U < 0$ and the generation prevails.

4.4.1 Example of application of the SHR model: the dynamics of generation-recombination phenomena

The generation-recombination function (GR function) of eq. 23 can be further simplified in some particular cases. The most interesting and frequent of them is the *low injection limit* where the density of excess charges is small with respect to the concentration of the majority charges. This condition ensures that there is enough majority charges in the centers of recombination to recombine any excess of minority charges.
In order to elucidate the use of the GR function let us consider two examples of symmetrical and asymmetrical non equilibrium conditions.

Electron-hole pairs creation

Let us consider a semiconductor where the thermal equilibrium ($n_0 p_0 = n_i^2$) is altered by the simultaneous creation of an excess of electrons and holes. This can be obtained for instance by a flash of light whose wavelength is sufficiently small to create electron-hole pairs. Note that this condition is met when the energy of the photons is greater than the energy gap ($\frac{hc}{\lambda} > E_{gap}$).
Let n' and p' be the instantaneous excesses of charges with respect to the equilibrium concentrations: n_0 and p_0. Then the total concentrations of charges are: $n = n_0 + n'$ and $p = p_0 + p'$.
Let us calculate the fate of the electrons when the semiconductor is not biased and the charges are

homogeneously created in the volume of the semiconductor. These conditions means that the drift and the diffusion currents are both null ($J_n = 0$). Applying the continuity equation we find:

$$\frac{dn}{dt} = \frac{d(n_0 + n')}{dt} = \frac{dn'}{dt} = G - R = -U = -\frac{np - n_i^2}{\tau_{n0}[p + n + n_i cosh(\frac{E_T - E_i}{kT})]} \tag{4.24}$$

The numerator of the GR function can be written as:

$$np - n_i^2 = (p' + p_0)(n' + n_0) - n_0 p_0 = p'n' + n'p_0 + p'n_0 \tag{4.25}$$

Under the low injection limit, $p'n'$ is the negligible product of two small quantities. Low injection limit also means that p' and n' are much smaller than p_0 and n_0 respectively. Furthermore, since electron-hole pairs are simultaneously created and the recombination times of electrons and holes are equal, we have $p' = n'$ at any time.
U function can be further simplified considering that the most efficient centers of recombination are located at $E_T = E_i$, and then the hyperbolic cosine can be replaced by 1. This assumption is supported by the fast growth of the hyperbolic cosine; as an example the function is ten times its minimum value when the argument is about 3. Then the hyperbolic cosine contribution to the GR function becomes negligible if $E_T - E_i > 3kT$ that is a very small quantity with respect to the whole energy gap.
Hereafter, we can assume that all the effective centers of recombination lie at $E_T \approx E_i$ and the hyperbolic cosine is replaced by 1.
Then U function is:

$$U = \frac{n'(p_0 + n_0)}{\tau_0(p_0 + n_0 + 2n_i)} \tag{4.26}$$

If the semiconductor is N-type: $p_0 \ll n_0$ and $n_i \ll n_0$; on the contrary, if it is P-type: $p_0 \gg n_0$ and $n_i \ll p_0$. In both cases the GR function for the electrons can be written as:

$$U = \frac{n'}{\tau_0} \tag{4.27}$$

Then, besides than on the density of excess charges, U function depends only on τ_0 namely on the inverse of the product between the density of the centers of recombination (N_T), the thermal velocity (v_{th}), and the intrinsic probability of capture (σ_0).
The U function determines, through the continuity equation, the fate of the excess charges created by the flash of light.

$$\frac{dn'}{dt} = -U = -\frac{n'}{\tau_0} \tag{4.28}$$

The equation can be integrated in dn' from $n'(0)$ and $n'(t)$ and in dt from 0 to t. The result is:

$$ln\left(\frac{n'(t)}{n'(o)}\right) = -\frac{t}{\tau_0} \rightarrow n'(t) = n'(o)exp\left(-\frac{t}{\tau_0}\right) \tag{4.29}$$

The charges are completely eliminated after about four times the recombination time (τ_0). After this interval, another independent process of creation-recombination can take place. This time is important for the response time of photoconductors.

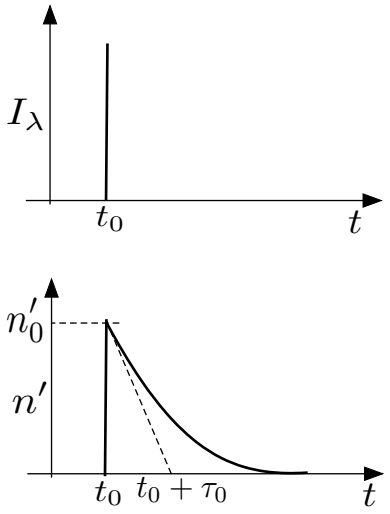


Fig. 4.6. *The charges created at time t_0 by a flash of light decay exponentially with a time scale tau_0.*

Non symmetrical charge injection

Let us consider now the case when only the density of one of the two charge carriers is altered. The excess charge can be due by the injection of a hole or a electron current from a nearby volume inside the material.
For this example let us consider an injection of holes in a N-type semiconductor. then $n = n_0$ and $p = p_0 + p'$.
The same conditions about the doping of the material and the low injection limit still hold (p' and $p_0 \ll n_0$ and $n_i \ll n_0$). Furthermore, the centers of recombination are located around the band gap center and then the hyperbolic cosine is replaced by 1. The generation-recombination function can be simplified as in the previous case:

$$U = \frac{p'}{\tau_0} \tag{4.30}$$

The same expression is also found in the case of excess electrons injected in a P-type semiconductor. Finally, we have introduced a simple and compact version of the GR function that can be used in the continuity equation for any deviation from equilibrium valid only when the low-injection limit is fulfilled. This equation is an important tool to calculate the current in PN junction based devices such as the PN diode and the Bipolar Junction Transistor.

4.4.2 The generation-recombination function for direct band gap materials

The GR function previously derived is valid when the processes of generation and recombination are mediated by inter-gap states. This is the typical process occurring in non direct band gap materials where the probability of direct transitions from the valence to the conductance band and vice versa are low.

On the other hand, in direct band gap semiconductors the band-to-band transition is highly probable. Direct band gap means that the top of the valence band is aligned, in the k-space, with the bottom of the conductance band. This is the case of the semiconductors of the III-V group of the periodic table such as gallium arsenide (GaAs), indium phosphide (inP), or aluminum arsenide (AlAs).
At thermal equilibrium, the recombination rate is simply proportional to the concentration of electrons and holes: $R = \beta n_0 p_0$. At the equilibrium, $R = G$, the generation rate can be defined as $G = \beta n_0 p_0$ where β is a constant of the phenomenon.
In case of a symmetrical excess of charges such as that described in section 4.1.1 the recombination is increased but the generation remains fixed at the equilibrium value. then:
$R = \beta(n_0 + n')(p_0 + p')$ and $G = \beta n_0 p_0$.
The generation-recombination function is:

$$U = R - G = \beta(n_0 + n')(p_0 + p') - \beta n_0 p_0 \tag{4.31}$$

If the low-injection limit is fulfilled, we have that $n'p'$ is negligible, then:

$$U = \beta(n_0 + p_0)n' = \frac{n'}{\tau_0} \tag{4.32}$$

The above equation holds in a N-type semiconductor where p_0 is negligible. Then $\tau_o = (\beta n_0)^{-1}$ is the the recombination time. An identical expression is obtained in case of a P-type material.
Under the low-injection limits the generation-recombination function in indirect and direct band gap semiconductors are analytically identical. Note that the recombination time depends on the number of recombination centers which are N_T when the recombination is mediated by the inter-gap centers, and n_0 in the case of band-to-band recombination.
Eventually, when the low-injection limit is satisfied, the continuity equation can be solved independently from the kind of semiconductor.

4.4.3 Appendix: the SHR generation-recombination function

The generation-recombination function can be calculated considering that at the equilibrium the rate of generation and recombination of electrons and holes are equal and the both the conditions defines the function U:

$$\begin{cases} U = r_1 - r_2 \\ U = r_3 - r_4 \end{cases} \tag{4.33}$$

replacing eq. 4.7 and 4.8 in the first equation and eq. 4.13 and 4.14 in the second it becomes:

$$\begin{cases} U = v_{th} n N_T (1 - f)\sigma_n - N_T f e_n \\ U = v_{th} p N_T f \sigma_p - N_T (1 - f) e_p \end{cases} \tag{4.34}$$

Replace in both equations e_n and e_p with eq. 4.11 and 4.15 respectively. The first equation gives:

$$U = v_{th} n N_T (1 - f)\sigma_n - N_T f exp\left(\frac{E_T - E_i}{kT}\right) \tag{4.35}$$

From which the Fermi-Dirac function is calculated

$$f = \frac{v_{th} n N_T \sigma_n - U}{v_{th} n N_T \sigma_n (n + n_i exp(\frac{E_T - E_i}{kT}))} \tag{4.36}$$

Introducing the time scale of electrons capture:

$$\tau_{n0} = \frac{1}{v_{th} n N_T \sigma_n} \tag{4.37}$$

The eq. 4.35 is:

$$f = \frac{n - U\tau_{n0}}{n + n_i exp(\frac{E_T - E_i}{kT})} \tag{4.38}$$

Repeat the same calculation for the second equation in eq. 4.33, introducing this time the time scale of holes capture:

$$\tau_{p0} = \frac{1}{v_{th} n N_T \sigma_p} \tag{4.39}$$

It leads to a second expression for the Fermi-Dirac function:

$$f = \frac{U\tau_{n0} + n_i exp(\frac{E_i - E_T}{kT})}{p + n_i exp(\frac{E_i - E_T}{kT})} \tag{4.40}$$

The two expressions of the Fermi-Dirac function are obviously equal:

$$\frac{n - U\tau_{n0}}{n + n_i exp(\frac{E_T - E_i}{kT})} = \frac{U\tau_{n0} + n_i exp(\frac{E_i - E_T}{kT})}{p + n_i exp(\frac{E_i - E_T}{kT})} \tag{4.41}$$

From which the U function is easily calculated:

$$U = \frac{np - n_i^2}{\tau_{n0}[p + n_i exp(\frac{E_i - E_T}{kT})] + \tau_{p0}[n + n_i exp(\frac{E_T - E_i}{kT})]} \tag{4.42}$$

Assuming that $\sigma_n = \sigma_p$ also $\tau_{n0} = \tau_{p0}$. Introducting τ_0, the unique characteristic time of capture of holes and electrons, and considering the definition of the hyperbolic cosine ($coshx = \frac{e^x - e^{-x}}{2}$) we obtain the simplified version of the generation-recombination function:

$$U = \frac{np - n_i^2}{\tau_{n0}[p + n + 2n_i cosh(\frac{E_T - E_i}{kT})]} \tag{4.43}$$

5

PN Junction

5.1 Introduction

The PN junction is one of the fundamental blocks of semiconductor devices, in particular it makes diodes, bipolar junction transistors and a plethora of sensors, not least the photodetectors.
The ideal PN junction is a homogeneous piece of semiconductor characterized by an abrupt behavior of the dopant atoms: at $x < 0$ the material is P-type doped by a uniform distribution of acceptors N_A and at $x > 0$ it is N-type doped by a uniform distribution of donors N_D. Obviously, this is just a ideal approximation of real PN junctions. Actually, the distribution of dopant atoms is not uniform and then the junctions are not sharp. Although the departures from reality, the ideal model is surprisingly accurate to describe most of the electric properties of the junction.
PN junctions can be formed by any couple of P and N semiconductors. To this regard, the noteworthy cases are the homojunctions formed by the same semiconductors, and the heterojunctions formed by different semiconductors. In this chapter homojunctions are treated, while heterjunctions are discussed in chapter 7. In terms of the band diagram, a homojunction is made by two materials sharing the same affinity and band gap but with different work functions. Hereafter, the materials properties are specified with suffixes p and n with the obvious meaning of P-type and N-type respectively.

5.2 The PN junction at the equilibrium

The formation of the junction can be ideally thought as the instantaneous union of two separated materials kept at thermal equilibrium. The process takes place at the same temperature, then the joint material reaches the equilibrium following the same steps outlined in chapter 2 except that this time both electrons and holes have to be considered. Then, electrons migrate from the material with the smallest work function towards that with the largest work function; and vice versa for the motion of holes. In other words, the electrons depart from the N-type material to migrate into the P-type material and the holes are transferred from the P-type to the N-type. Charges transfer proceeds until the Fermi level in the whole device becomes uniform.
The charges injected from one material to the other are minority charges in the destination regions. Hence, they are quickly recombined by the majority charges. On the other hand, majority transfer leaves their own material. Both these processes lead to the formation of a density of fixed charges

made by the ionized dopant atoms. Eventually, a double layer of oppositely charged regions (positive in the N-type side and negative in the P-type) is formed. The most immediate consequence of these charges is an electric field that prevent any further transfer of charges and then establish the equilibrium in the system. It is worth to remind once again that equilibrium means a balance between opposite currents. The equilibrium band diagram is shown in figure 2.
At the equilibrium, the whole system is formed by two unperturbed bulk regions connected by a space charge region (in particular a depletion layer). The total charge density in any volume is the sum of the concentration of the four kinds of available charges: $Q = q(p - n + N_D - N_A)$. The charge densities in each zone are listed in Table 1.

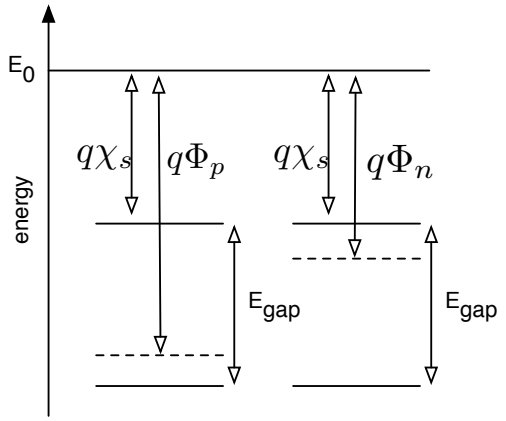


Fig. 5.1. *Band diagrams of the two constituents of a ideal PN junction.*

Table 5.1. Charge condition in the three regions of the semiconductor.

region	size	holes density	electrons density	total charge density
P neutral zone	$x < -x_p$	$p_p = N_A$	$n_p = \frac{n_i^2}{N_A}$	$\rho = 0$
N neutral zone	$x > x_n$	$p_n = \frac{n_i^2}{N_D}$	$n_p = N_D$	$\rho = 0$
Space charge region	$-x_p < x < x_n$	$p < N_A$	$n < N_D$	$\rho \neq 0$

The electric field and the potential are calculated from the Poisson equation solved in the space charge region where the total charge is different from zero. Also in this case the electrostatic characteristics of the junction are calculated under the deep depletion hypothesis assuming that the total charge in the depletion layer is only contributed by the donor and acceptor atoms. Then the Poisson equation is:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_s}(N_D - N_A) \tag{5.1}$$

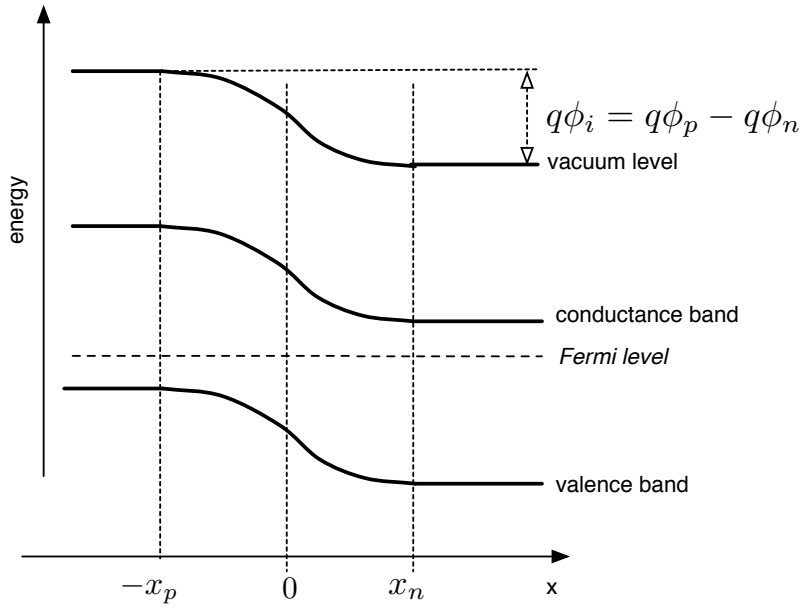


Fig. 5.2. *Band diagrams at the equilibrium.*

The deep depletion hypothesis applied to the step junction results in the charge distribution depicted in figure 3. The striking feature of this distribution is the sharp transition from the neutral zone to the depletion layer. Let us remark again that this is a ideal representation of the real distribution of charges, later a modification of the model will be introduced to remove at least the abrupt transitions at the borders of the space charge region.

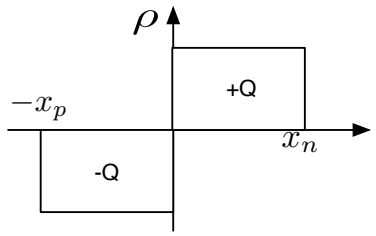


Fig. 5.3. *Distribution of total charges in the deep depletion approximation.*

The amounts of charge in the two sides of the depletion layer are: $+Q = +qN_D x_n$ and $-Q = -qN_a x_p$. Since the total charge has to be null, the following relationship holds

$$N_D x_n = N_A x_p \tag{5.2}$$

Equation 5.2 shows that the extent of the depletion layer is inversely proportional to the doping. Namely, the depletion layer extends more towards the less doped side of the junction. Obviously, the two sides of the depletion layers are symmetrical only when the concentrations of the dopant atoms are the same.
The distribution of charges shown in figure 3 is the input to calculate the electric field and the potential.
The electric field is given by:

$$\frac{d\mathcal{E}}{dx} = \frac{\rho}{\epsilon_s} \tag{5.3}$$

in the N-type side of the depletion layer ($0 < x < x_n$), under the hypothesis of constant doping, the electric field is:

$$\int_{\mathcal{E}(x)}^{\mathcal{E}(x_n)} d\mathcal{E} = \int_x^{x_n} \frac{qN_D}{\epsilon_s} dx \rightarrow \mathcal{E}(x_n) - \mathcal{E}(x) = \frac{qN_D}{\epsilon_s}(x_n - x) \tag{5.4}$$

since the bulks of the semiconductors are neutral zones, the boundary condition is $\mathcal{E}(x_n) = 0$, then the electric field in the N-type side of the depletion layer is:

$$\mathcal{E}_n(x) = \frac{qN_D}{\epsilon_s}(x - x_n) \tag{5.5}$$

Similarly, in the P-type part of the space charge region where:

$$\int_{\mathcal{E}(-x_p)}^{\mathcal{E}(x)} d\mathcal{E} = -\int_{-x_p}^{x} \frac{qN_A}{\epsilon_s} dx \rightarrow \mathcal{E}(x) - \mathcal{E}(x_p) = -\frac{qN_D}{\epsilon_s}(x + x_p) \tag{5.6}$$

with the boundary condition $\mathcal{E}(-x_p) = 0$ the electric field in the P-type side is:

$$\mathcal{E}_p(x) = -\frac{qN_A}{\epsilon_s}(x + x_p) \tag{5.7}$$

Since the material is homogeneous, the electric permittivity is a constant, and the electric field is continuous at the interface ($\mathcal{E}_n(0) = \mathcal{E}_p(0)$) where it takes its absolute value is maximum:

$$\mathcal{E}_{max} = -\frac{qN_A}{\epsilon_s} x_p = -\frac{qN_D}{\epsilon_s} x_n \tag{5.8}$$

It is worth to note that eq. 5.8 is similar to eq. 5.2 and it provides an alternative expression of the null total charge condition.
The potential is calculated from the electric field ($d\phi = -\mathcal{E}dx$) in the two sides of the depletion layer.
The boundary conditions for the calculation of the potential are: $\phi(x \geq x_n) = \phi_n$ and $\phi(x \leq -x_p) = -\phi_p$. Where ϕ_n and $-\phi_p$ are the potentials in the neutral zone as defined by eq. 2.74 where $\phi = E_F - E_i$.

$$x < 0 \rightarrow \phi(x) = -\phi_p + \frac{qN_A}{2\epsilon_s}(x + x_p)^2 \tag{5.9}$$

$$x > 0 \rightarrow \phi(x) = \phi_n - \frac{qN_D}{2\epsilon_s}(x - x_n)^2 \tag{5.10}$$

Note that the potential in x=0 is null only if the concentrations of acceptors and donors are the same ($N_A = N_D$). The null potential occurs where the semiconductor becomes intrinsic ($p = n = n_i$). Noteworthy, in case of asymmetric doping, the potential becomes null inside the depletion layer of the less doped semiconductor. Thus, in this side of the depletion layer, the region between the condition where $\phi = 0$ and the interface is populated by more minority charges than majority charges. Such a situation is called inversion and it is a fundamental phenomenon in metal-oxide-semiconductors junctions.
The net drop of potential across the whole junction is the built-in potential. It can be calculated from the potentials in the unperturbed neutral zones, $\phi_{bi} = \phi_n - \phi_p$, that depend on the concentration of majority charges in the respective regions (see eqs. 2.73 and 2.74).

$$\phi_n = \frac{kT}{q} ln\left(\frac{N_D}{n_i}\right);\ \phi_p = -\frac{kT}{q} ln\left(\frac{N_A}{n_i}\right); \tag{5.11}$$

from which:

$$\phi_i = \frac{kT}{q}\left[ln\left(\frac{N_D}{n_i}\right) + ln\left(\frac{N_A}{n_i}\right)\right] = \frac{kT}{q} ln\left(\frac{N_A N_D}{n_i}\right) \tag{5.12}$$

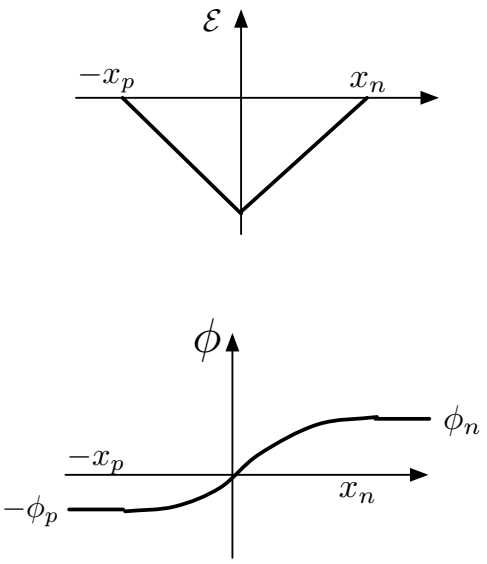


Fig. 5.4. *Electric field and potential in a PN junction at the equilibrium.*

As an example, if $N_D = N_A = 10^{16} cm^{-3}$ the built-in potential is about 0.72 V.
The total size of the depletion layer descends from the condition of continuity of the potential in x=0.

$$\phi_n - \frac{qN_D}{2\epsilon_s}x_n^2 = \phi_p + \frac{qN_A}{2\epsilon_s}x_p^2 \tag{5.13}$$

Then, introducing the built-in potential we obtain:

$$\phi_i = \phi_n - \phi_p = \frac{q}{2\epsilon_s}(N_D x_n^2 + N_A x_p^2) \tag{5.14}$$

Replacing x_n with the expression derived from the charge neutrality condition (eq. 5.2): $x_n = \frac{N_D}{N_A}x_p$ the depletion layer in the P region can be calculated:

$$x_p = \sqrt{\frac{2\phi_i\epsilon_s}{q}}\frac{N_D}{\sqrt{N_D N_A^2 + N_A N_D^2}} \tag{5.15}$$

Replacing x_p the depth of the depletion layer in N region is found:

$$x_n = \sqrt{\frac{2\phi_i\epsilon_s}{q}}\frac{N_A}{\sqrt{N_D N_A^2 + N_A N_D^2}} \tag{5.16}$$

Eventually, the whole depletion layer is

$$x_d = x_n + x_p = \sqrt{\frac{2\phi_i\epsilon_s}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)} \tag{5.17}$$

5.2.1 Removal of the charge discontinuity at the depletion layer border

According to the hypothesis of deep-depletion, the concentration of the majority charges becomes abruptly negligible at the borders of the depletion layer. The relationship between charges density and potential is ruled by the Boltzmann expression, then even a small bending might result in a large change of electrons (and holes) concentration. This makes plausible the deep depletion hypothesis inside the depletion layer, but around the border of the depletion layer a smooth transition has to exist. Figure 5 shows a qualitatively more realistic behavior of mobile and total charges.
In this section a quantitatively expression of the behavior of the potential around the border of the space charge region is evaluated. It is important to remark that this behavior is common to any depletion layer disregarding the material where it takes place. In particular, the same conclusions hold also for the metal-semiconductor junction described in chapter 2.
To calculate the behavior of the potential around the border of the depletion layer let us consider the Poisson equation. The calculation is shown here for the N-type material, of course, on the other side and absolutely symmetrical calculation can be carried out.
The total charge in the N-type material is contributed by donors and electrons. The concentration of donors is constant (by hypothesis) and the concentration of the electrons depends, through the Boltzmann equation, on the potential. Then the Poisson equation can be written as:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_s}(N_D - n) = -\frac{q}{\epsilon_s}\left[N_D - n_i exp\left(\frac{q\phi}{kT}\right)\right] \tag{5.18}$$

At the border of the depletion layer the actual potential is smaller than ϕ_n, and it can be written as: $\phi = \phi_n - \phi'$. The deep depletion hypothesis results in $\phi = \phi_n$ at $x = x_n$. The smooth transition

of the density of electrons from 0 to N_D is represented by the potential ϕ'.
Replacing the potential in the Poisson equation we get:

$$\frac{d^2(\phi_n - \phi')}{dx^2} = -\frac{q}{\epsilon_s}\left[N_D - n_i exp\left(\frac{q\phi_n}{kT}\right) exp\left(-\frac{q\phi'}{kT}\right)\right] \tag{5.19}$$

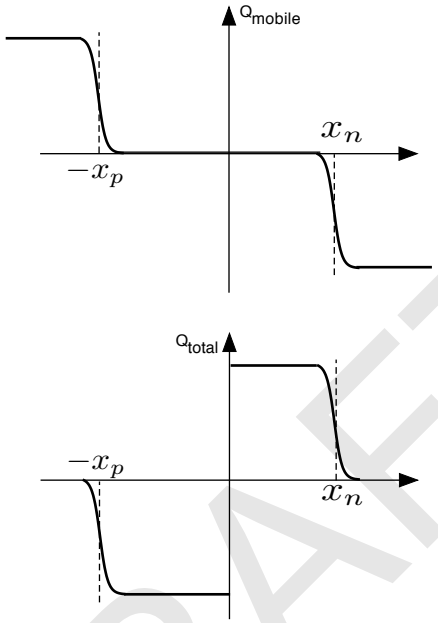


Fig. 5.5. *Qualitative behavior of mobile charge (electrons and holes) and total charge: mobile charges plus acceptors and donors.*

Considering that ϕ_n is constant and the first exponential defines the concentration of electrons in the neutral zone (see eq. 5.11), namely the donor concentration we obtain:

$$-\frac{d^2\phi'}{dx^2} = -\frac{qN_D}{\epsilon_s}\left[1 - exp\left(-\frac{q\phi'}{kT}\right)\right] \tag{5.20}$$

Around x_n, ϕ' is a small perturbation, so the exponential can be replaced with its first order approximation ($e^{-x} = 1 - x$).

$$\frac{-d^2\phi'}{dx^2} = -\frac{qN_D}{\epsilon_s}\left(1 - 1 + \frac{q\phi'}{kT}\right) = -\frac{q^2N_D}{\epsilon_s kT}\phi' \tag{5.21}$$

The previous equation contains the Debye length defined as $L_D = \sqrt{\frac{\epsilon_s kT}{q^2 N_d}}$
Then the Poisson equation can be written as:

$$\frac{d^2\phi'}{dx^2} = \frac{\phi'}{L_D^2} \tag{5.22}$$

whose solution is: $\phi' = A\ exp(\frac{x}{L_D}) + B\ exp(-\frac{x}{L_D})$. The constants A and B depends on the boundary conditions. The first boundary conditions is $\phi'(0) = \phi'_0$, which is the deviation with respect to ϕ_n at $x = x_n$. For sake of simplicity, the origin of the coordinate x is translated to $x = x_n$. For the second boundary condition let us consider that the length of the bulk is much greater than the Debye Length. the the perturbative potential $\phi'(x)$ vanishes inside the semiconductor. The two boundary conditions provides that $A = 0$ and $B = \phi'_0$. Eventually, the excess potential due to the smooth behaviour of total charge is $\phi' = \phi'_0\ exp(-\frac{x}{L_D})$.
The additive term for the electric field is then given by:

$$\mathcal{E}' = -\frac{d\phi'}{dx} = \frac{\phi'_0}{L_D} exp\left(-\frac{x}{L_D}\right) \tag{5.23}$$

Even the electric field decays exponentially beyond x_n and x_p towards the bulk of the semiconductor. The Debye length is the length scale of the exponential decay. In practice the electric field becomes negligible at about four Debye lengths from the boundary of the depletion layer calculated with the deep depletion approximation.
The Debye length depends on the permittivity and the doping. For silicon, in case of $N_A = N_D = 10^{16} cm^{-3}$ the Debye length is $L_D = 40\ nm$. This figure has to be compared with the corresponding depletion layer width (eq. 5.17) is $x_d = 332 nm$.
The deep depletion approximation can be applied. However, outside the depletion layer the electric field is small but non zero. Such a small field can be neglected in the description of electronic devices but it is of great importance in case of light detectors such as photodiodes and solar cells.
Both the depletion layer size and the Debye length depends on the doping concentration, it may be of interest to determine the relationship between these two quantities.
For this scope let us consider a symmetric diode ($N_D = N_A$) where the potential at the origin is null. Then $\phi_0 = \phi_n - \frac{qN_D}{2\epsilon_s} x_n^2 = 0$. Replacing the expression for ϕ_n we obtain:

$$x_n^2 = 2\frac{\epsilon_s kT}{q^2 N_D} ln\left(\frac{N_D}{n_i}\right) = 2L_D^2 ln\left(\frac{N_D}{n_i}\right) \tag{5.24}$$

Then the ratio between the depletion layer size and the Debye length is:

$$\frac{x_n}{L_D} = \sqrt{2ln\left(\frac{N_D}{n_i}\right)} \tag{5.25}$$

The ratio between the depletion layer size and the Debye length is rather independent from the doping concentration. At $N_D = 10^{16} cm^{-3}$ the ratio is 5.25, if N_D is 100 times larger it becomes only 6.06.

5.2.2 Actual configurations

Figure 5.6 shows the simplified sketch of a planar PN junction. The actual area of the device is defined by the density of the field lines connecting the two terminals. As a first approximation the device can be restricted to the separation surface parallel to the surface of the semiconductor.
The device can be fabricated from a P-type semiconductor, through a suitable lithographic step where a circular area is defined at the surface from which N-dopant atoms are diffused. Most used methods for doping are ion implantation and gaseous deposition. In the first case, dopant atoms are ionized and accelerated towards the semiconductor surface, in the second case the dopant atoms

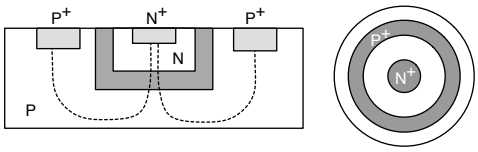


Fig. 5.6. *Schematic arrangement of a planar PN junction diode. To preserve symmetry, metal contacts are ring shaped, the current flows from the junction towards the lateral contact.*

are vaporized in gas phase and absorbed by the semiconductor through a solubility process. Diffusion enables the doping atoms to cover a relevant volume of the semiconductor. Either implanted or absorbed atoms are let diffuse under controlled temperature. At low concentrations, the dopant profile is properly described by the diffusion current and the continuity equation. These are the first and the second Fick's laws.

$$J = -D\frac{\partial C(x,t)}{\partial x}; \qquad \frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial x^2} \tag{5.26}$$

The solution of the equations depends on the boundary conditions. In case of ion implantation, the amount of atoms is limited and as the dopants penetrate the material the concentration at surface decreases. The dopant profile has an exponential behavior. In case of absorbed atoms, the concentration at surface is maintained constant by the atoms in gas phase, and the concentration inside the material evolve as the complementary error function (erfc) which is defined as: $erfc = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$. In both cases, the implantation of donors in the P-type material results in a non constant profile, the junction occurs where the added dopant concentration becomes smaller than the background dopant concentration. Figure 5.6 shows an example of the dopant concentration profile in case of doping from gas phase.

The behaviour around the junction of the doping atoms can be approximated by a liner function. Note that the assumption of a linear behavior of the charges in the depletioin layer changes the exponent of the dependence of the electric field and the potential from the distance. In particular, the electric field becomes dependent on the square of the distance and the potential on the third power. Qualitatively, this introduces only a small perturbation to the electrostatic quantities calculated with the hypothesis of a constant donor concentration in each zone.

5.3 The current in the PN junction

In order to calculate the current to voltage relationship of a PN junction let us consider the ideal device sketched in figure 7. The device is made by A space charge region and two neutral regions in the P and N type sides. A ohmic contact is provided at each end of the device. The contact is provided through a heavily doped layer. The ideal diode is mono dimensional and instead of the current the current density J is calculated, so the measurable current is $I = JA$ where A is the cross-sectional area of the device.

The voltage V_A is applied between the electrode of the P-type material and the electrode of the N-type material. For sake of simplicity, the N-type material is grounded. As discussed in chapter 2, the applied voltage is almost completely found across the depletion layer. This is a fundamental

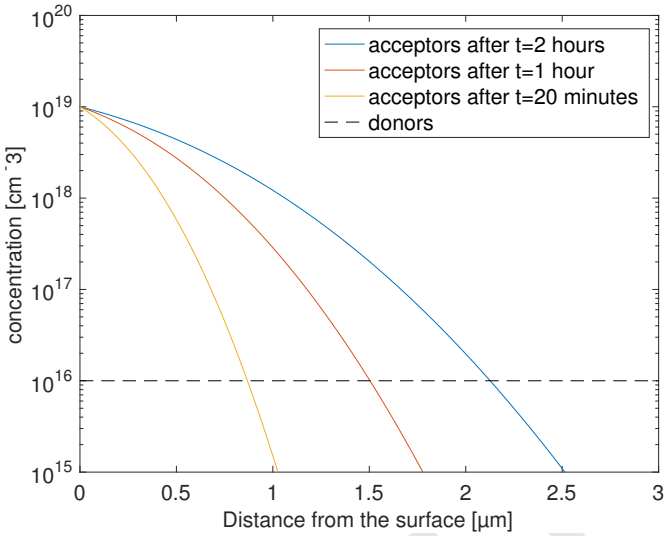


Fig. 5.7. *Example of solution of the diffusion equations in case of gaseous deposition. The example shows the diffusion of Boron atoms in a silicon wafer doped with* 10^{16} cm^{-3} *donors.The process occurs at 1100* $^\circ$ *C at this temperature the diffusion constant of boron is* $D = 2.96 \cdot 20^{-13}$ cm^2/s. *The concentration of boron atoms at the surface is* $C_0 = 10^{19}$ cm^3. *The solution of the diffusion equation is* $C(x,t) = C_0 \grave{e} rfc(\frac{x}{2sqrtDt}$ *where t is the time. The curves are calculated at t=2 hours, 1 hour and 20 minutes. The junction point, where the doping character of the material switches from N to P, occurs at a distance in the range* $1 - 2\mu m$ *approximately.*

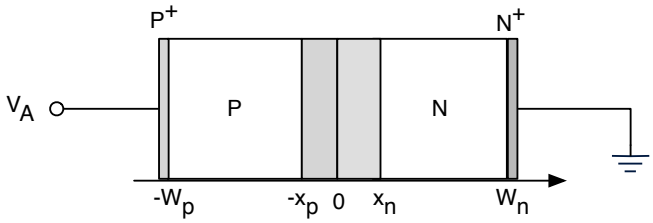


Fig. 5.8. *Ideal sketch a PN junction. The device is biased through two ohmic contacts and a voltage* V_A *is applied to the device. For sake of simplicity, the N type material is grounded. The interface between the two materials is the origin of the coordinate x along which the current flows.*

assumption to study the current in the device, as a consequence the electric field in the neutral zones of the device is null, then no drift current can take place in the bulk of the two pieces of semiconductor.

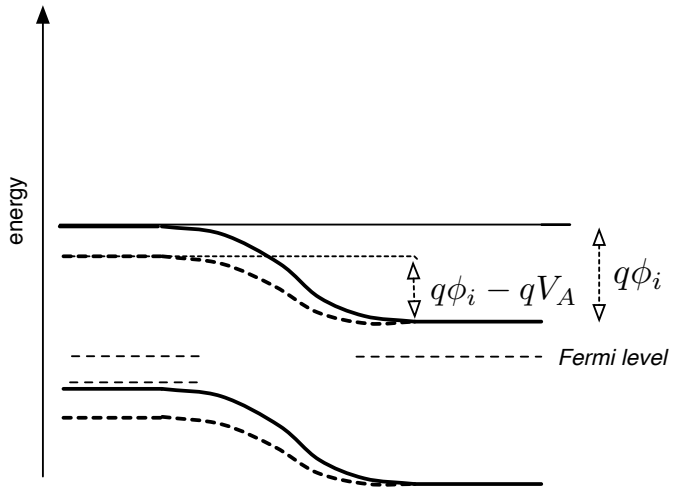


Fig. 5.9. *Band diagram change due to the voltage V_A applied according to the condition in figure 6. The Fermi level is no more uniform due the non equilibrium condition. The Fermi level in the two neutral zones is different due to the different electric potential of the charges.*

The energy of the electrons in the P-type semiconductor is shifted by a quantity $-qV_A$, while the energy of the electrons in the N-type semiconductor remains unchanged. The built-in potential becomes $q\phi_i = q(\phi_i - V_A)$ and all the quantities depending on the built-in potential changes as a consequence. For instance, the depletion layer size is changed into:

$$x_d = x_n + x_p = \sqrt{\frac{2(\phi_i - V_A)\epsilon_s}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)} \tag{5.27}$$

The built-in potential (ϕ_i) keeps the whole system in equilibrium where the total current across each section of the device is zero. In the equilibrium condition, the concentrations of majority and minority charges in the neutral zones are ruled by the mass action law.
The changes of the built-in potential barrier depend on the sign of the applied voltage. The condition $V_A > 0$ is called *forward bias*. This corresponds to the case when the potential barriers are lowered. The change of the potential makes no more valid the mass action law. The lowering of barrier increases the concentration of the minority charges in both the neutral zones. Indeed, the built-in potential prevents the majority charges to diffuse in the other region where their concentration is much smaller.
Then, with respect to the equilibrium condition, there will be an excess of minority charges in the neutral zones

$$n_p \rightarrow \frac{n_i^2}{N_A} + n' \; ; \; p_n \rightarrow \frac{n_i^2}{N_D} + p' \tag{5.28}$$

where n_p is the concentration of electrons in the P-type materials, and p_n is the correspondent concentration of holes.
This excess of charges is responsible of the current. However, charges collection at the electrodes is not straightforward because there is no electric field in the regions between the depletion layer and the electrodes. Rather, the current is driven by the recombination phenomena.
With the opposite bias ($V_A < 0$), in a condition called *reverse bias*, the potential barrier is increased favouring the transfer of the minority charges from one material to the other. Due to the different numerosity of majority and minority charges, it is clear that the two currents are quantitatively different. This difference leads to the rectifier character of the PN junction.
In definitive, under direct bias the large population of majority charges are transferred to the other region where they give place to an excess of minority charges that are recombined according to the Shockley-Hall-Read (SHR) recombination law. While, under reverse bias the small population of minority charges are transferred to the other region to produce a almost negligible increase of majority charges. Eventually, the properties of the PN junction depends on the fate of the minority charges.
The total current in the device, is made by two main contributions. The dominant component is the current due to the increase of minority charges in the neutral zone, the second additional term, which is dominant in reverse bias, is due to the processes of generation and recombination in the space charge region.
The first contribution to the current is *the ideal current of the diode.*

5.3.1 Ideal current

The ideal current is the consequence of the changes of the concentration of the minority charges at the border of the neutral zones. The following calculation is based on the quasi-equilibrium condition, which states that even under bias, the concentration of electrons and holes can still be calculated using the equations of the statistical equilibrium but replacing the built-in potential with the actual barrier.
Another important condition to calculate the ideal current is the low-injection limit. This condition allows to consider the simplified version of the generation-recombination function (eq. 3.26 and 3.28 of chapter 3).
The ideal current stems from the modulation of the concentration of electrons and holes at the border between the depletion layer and the two neutral zones.
At the equilibrium ($V_A = 0$) the majority charges at the coordinates x_n and $-x_p$ are:

$$n_{n0}(x_n) = N_D; \quad p_{p0}(-x_p) = N_A; \tag{5.29}$$

and the corresponding minority charges are:

$$n_{p0}(-x_p) = n_{n0}(x_n) \; exp\left(-\frac{q\phi_i}{kT}\right) = N_D \; exp\left(-\frac{q\phi_i}{kT}\right) = \frac{n_i^2}{N_A} \tag{5.30}$$

and

$$p_{n0}(x_n) = p_{p0}(-x_p) exp\left(-\frac{q\phi_i}{kT}\right) = N_A exp\left(-\frac{q\phi_i}{kT}\right) = \frac{n_i^2}{N_D} \tag{5.31}$$

The subscripts p and n indicate the sides of the junction while the subscript 0 indicates the equilibrium .
Under bias ($V_A \neq 0$) the concentration of the minority charges at the borders of the neutral zone is altered into:

$$n_p(-x_p) = N_D \; exp\left(-\frac{q(\phi_i - V_A)}{kT}\right) > \frac{n_i^2}{N_A} \tag{5.32}$$

and

$$p_n(x_n) = N_A \; exp\left(-\frac{q(\phi_i - V_A)}{kT}\right) > \frac{n_i^2}{N_D} \tag{5.33}$$

The primary consequence of the applied voltage is the creation of an excess of minority charges at the borders of the neutral zones. The concentration of minority charges is kept constant by the applied voltage, then even if these charges are removed by recombination their concentration at x_n and $-x_p$ remains constant.
The concentration of excess charges can be written as:

$$n'_p(-x_p) = n_p - n_{p0} = N_D \; exp\left(-\frac{q(\phi_i - V_A)}{kT}\right) - N_D \; exp\left(-\frac{q\phi_i}{kT}\right) = N_D \; exp\left(-\frac{q\phi_i}{kT}\right) \; [exp\left(\frac{qV_A}{kT}\right) - 1] \tag{5.34}$$

and

$$p'_n(x_n) = n_p - n_{p0} = N_A \; exp\left(-\frac{q(\phi_i - V_A)}{kT}\right) - N_A \; exp\left(-\frac{q\phi_i}{kT}\right) = N_A \; exp\left(-\frac{q\phi_i}{kT}\right) \; [exp\left(\frac{qV_A}{kT}\right) - 1] \tag{5.35}$$

At the equilibrium, the density of the minority charges is about 11 orders of magnitude smaller than the density of majority charges. Thus, until the excess charges do not reach this level the low-injection condition is fulfilled.
The excess of minority charges (electrons in the P-type material and holes in the N-type material) are created at the border of the neutral zones where the electric field is null. Then, the current is collected at the electrode is a diffusion current, this means that in order to measure a constant current a constant gradient of minority charges has to be established in the neutral zones.
The relevant phenomena responsible for the current is the recombination, in the two regions. The excess of minority charges activates the recombination processes that tend to restore the equilibrium. Under the low-injection limit hypothesis, the generation-recombination function can be written in a simplified form and the U function in the two sides of the junction is:

$$U_n = (R - G)_n = \frac{n'_p}{\tau_n}; \quad U_p = (R - G)_p = \frac{p'_n}{\tau_p}; \tag{5.36}$$

Let us do the calculation of the current due to the holes injected in the N-type material. The calculus are symmetrical for the electrons in the other side of the device.
To calculate the effect of the recombination, let us consider the continuity equation for holes.

$$\frac{\partial p_n}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - U \tag{5.37}$$

Since the electric field is null, the current is a diffusion current: $J_p = -qD_p\frac{\partial p_n}{\partial x}$. The concentration of holes is $p_n = p'_n + p_{n0}$. The doping is uniform, namely p_{n0} is uniform in time and in space.

Replacing the recombination function, the diffusion current and the total holes concentration in the continuity equation we obtain:

$$\frac{\partial p'_n}{\partial t} = D_p \frac{\partial^2 p'_n}{\partial x^2} - \frac{p'_n}{\tau_p} \tag{5.38}$$

Let us restrict our analysis to the steady state, namely to the d.c. current. Some considerations about the transient phonemena are given at the end of this chapter. A steady current means that $\frac{\partial p'_n}{\partial t} = 0$, and the continuity equation for the holes injected in the N-type region is:

$$\frac{\partial^2 p'_n}{\partial x^2} = \frac{p'_n}{L_p^2} \tag{5.39}$$

The quantity $L_p = \sqrt{D_p \tau_p}$ is called the *recombination length*. It is a measure of the penetration depth of the holes inside the N-type semiconductor.
The solution of the continuity equation is a linear combination of two exponential functions:

$$p'_n(x) = A\ exp\left(\frac{x - x_n}{\sqrt{D_p \tau_p}}\right) + B\ exp\left(-\frac{x - x_n}{\sqrt{D_p \tau_p}}\right) \tag{5.40}$$

That can be alternatively written in terms of hyperbolic functions:

$$p'_n(x) = A^* sinh\left(\frac{x - x_n}{L_p}\right) + B^* cosh\left(\frac{x - x_n}{L_p}\right) \tag{5.41}$$

Where $A^* = \frac{A-B}{2}$ and $B^* = \frac{A+B}{2}$. The constants have to be calculated from the boundary conditions. The steady-state profile of the concentration of holes is achieved by an equilibrium between the diffusion of the holes from the space charge region towards the interior of the semiconductor and the recombination that tends to eliminate the holes in excess with respect to the thermal equilibrium, the whole process is ruled by the diffusion length. A short L_p is found in materials characterized by an efficient recombination of holes. It is worth to remind that the recombination time τ_p is inversely proportional to the density of recombination centres and their cross-section of capture of holes (see eq. 17 of chapter 3), and D_p is proportional to the mobility and then inversely proportional to the doping.
The holes distribution is the necessary ingredient to calculate the current. It is important to keep in mind that the profile is steady in time but individual charges are continually injected from the depletion layers and recombined by the recombination centers in the neutral zone.
The current can be directly calculated from the definition of the diffusion current and it is proportional to the derivative of the profile of concentration of the excess holes. The profile of concentrations of excess charges can be calculated imposing the boundary conditions to equation 5.40. The first boundary condition is the concentration of excess holes at x_n, while the second boundary condition is fixed by the position of the electrode. At the electrode indeed all the excess charges are eliminated, then $p'_n(W_B) = 0$.
This means that:

$$p'_n(x = x_n) \rightarrow B^* = p'_n(x_n) \quad p'_n(W_B) = 0 \rightarrow A^* = -\frac{p'_n(x_n)}{tanh(\frac{W_B}{L_p})} \tag{5.42}$$

Finally the distribution of excess holes is given by:

$$p'_n(x) = p'_n(x_n)\left[cosh\left(\frac{x-x_n}{L_p}\right) - \frac{sinh(\frac{x-x_n}{L_p})}{tanh(\frac{W_B}{L_p})}\right] \tag{5.43}$$

the analytical behavior depends on the distance between the electrode and the depletion layer with respect to the diffusion length. The function approximates an exponential as $W_B \gg L_p$ and a linear function as $W_B \ll L_p$. Figure 9 shows the behavior of the excess holes profile at different W_B/L_p ratios.

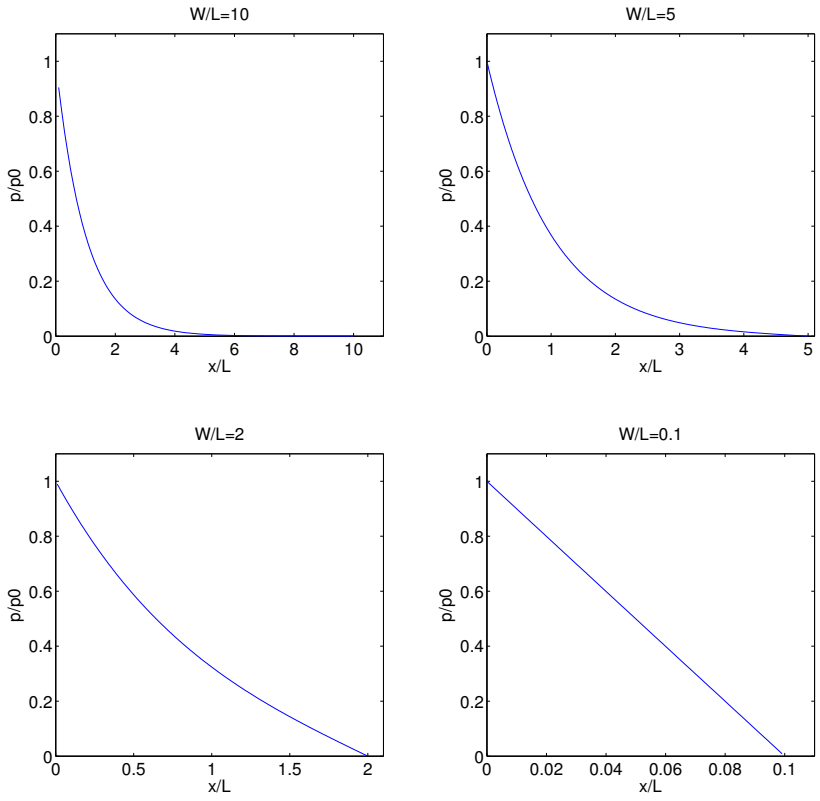


Fig. 5.10. *Decay of the excess holes in a N-type semiconductor calculated from equation 5.42 at different W_B/L_p ratios. Observe that as $W \gg L$ the behavior tends to an exponential function and, on the other hand, when $W \ll L$ it tends to be linear.*

As a consequence, also the current depends on the depth of the neutral zone defined by the distance between the border of the depletion layer, from where the excess charges are injected, and the electrode where the charges are collected by the external circuit. To this regard, it is convenient

to discuss the two limit cases called long base diode, and short base diode where the solution of the continuity equation is rather simplified. The nomenclature is clearly referred to the bipolar transistor, since the PN junction is the fundamental block of this device.
In the following the current will be calculated in each of these extreme situations.

Long base diode

The condition of long base diode is met when the distance between the electrode and the depletion layer is much larger than the diffusion length. According to the symbols in figure 5.6 the long base diode corresponds to the following conditions: $W_B - x_n \gg L_p$ in the N-type material and $-W_E + x_p \gg L_n$ in the P-type material. The two zones are independent, so it is possible to make a diode where the bases are long and short respectively.
Since L_p is much shorter than the distance to be travelled to reach the electrode, all the excess holes are recombined before to leave the semiconductor. This condition applied to eq. 39 results to A=0 and the solution of the continuity equation is limited to the decreasing exponential function. Then the constant B in eq. 39 is calculated from the other boundary condition: $p'_n(x_n)$. Eventually, the steady state profile of concentration of holes is:

$$p'_n(x) = p'_n(x_n)\ exp\left(-\frac{x - x_n}{L_p}\right) \tag{5.44}$$

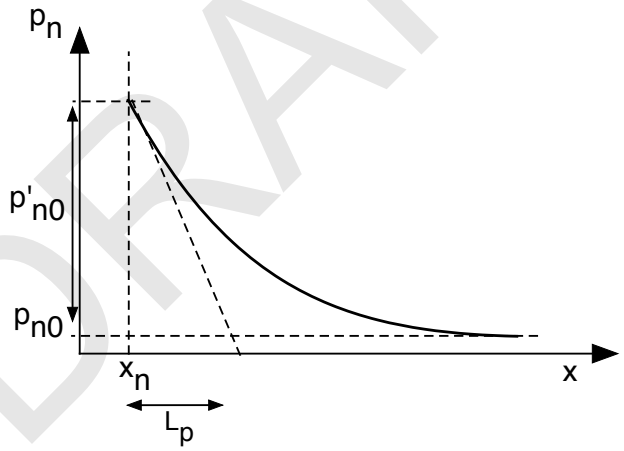


Fig. 5.11. *Steady state distribution of the concentration of excess holes in the neutral region of the N-type semiconductor. The excess charge p'_n are summed over the constant equilibrium holes concentration (p_{n0}.*

Replacing $p'_n(x_n)$ with its expression (from eq. 34) we have.

$$p'_n(x_n) = N_A\ exp\left(-\frac{q\phi_i}{kT}\right)\ \left[exp\left(\frac{qV_A}{kT}\right) - 1\right] = \frac{n_i^2}{N_D}\ \left[exp\left(\frac{qV_A}{kT}\right) - 1\right] \tag{5.45}$$

namely:

$$p'_n(x) = \frac{n_i^2}{N_D} \left[exp\left(\frac{qV_A}{kT}\right) - 1\right] exp\left(-\frac{x - x_n}{L_p}\right) \tag{5.46}$$

From this profile the diffusion current is calculated:

$$J_p(x) = -qD_p \frac{\partial p'_n}{\partial x} = q\frac{D_p}{L_p}\frac{n_i^2}{N_D} \left[exp\left(\frac{qV_A}{kT}\right) - 1\right] exp\left(-\frac{x - x_n}{L_p}\right) \tag{5.47}$$

Obviously the current of the excess holes is not constant in the space, but it follows the profile of the density of the excess holes. In particular, as the holes travel inside the material their current decreases, and it vanishes well before to reach the electrode. In this condition, no current should be recorded in the external circuit, but it has to be considered the properties of the recombination.
Actually, the recombination of holes is achieved consuming the majority charges (electrons) so the gradient of holes is complemented by a specular gradient of electrons. As a consequence there is a complementary current of electrons originated from the electrode so that the sum of the currents of electrons and holes is constant throughout the material: $J = J_n + J_p$. It is interesting to note that the electrons necessary for the recombination are provided through the ohmic contact by the external circuit that has the function of a "electrons reservoir".
Since their sum is constant, the two currents can be calculated everywhere between the depletion layer and the electrode. The most convenient location for the calculus is x_n where the current of holes is maximum and the current of electrons is null.
The total current flowing in the N-type material due to the holes injected from the P-type side of the junction is:

$$J_N = J_p(x_n) = q\frac{D_p}{L_p}\frac{n_i^2}{N_D}\left[exp\left(\frac{qV_A}{kT}\right) - 1\right] \tag{5.48}$$

A similar calculus can be done for the electrons injected in the P-type material that give rise to the current flowing in the P-type side of the junction.

$$J_P = J_n(-x_p) = q\frac{D_n}{L_n}\frac{n_i^2}{N_A}\left[exp\left(\frac{qV_A}{kT}\right) - 1\right] \tag{5.49}$$

Finally, the total current is the sum of the two currents, note that only in the case of equal doping the two currents have the same magnitude.

$$J_{tot} = J_N + J_P = qn_i^2\left(\frac{D_n}{N_A L_n} + \frac{D_p}{N_D L_p}\right)\left[exp\left(\frac{qV_A}{kT}\right) - 1\right] \tag{5.50}$$

The majority charges transferred from one material to the other are also continuously provided by the electrodes, so the total current is constant everywhere in the whole device.
In the ideal current only the current generated by the changes in the concentration of the minority charges at the interface between the space charge region and the neutral zone is considered. Additional phenomena occurring in the depletion layer are necessary to be taken into account for a more complete picture.

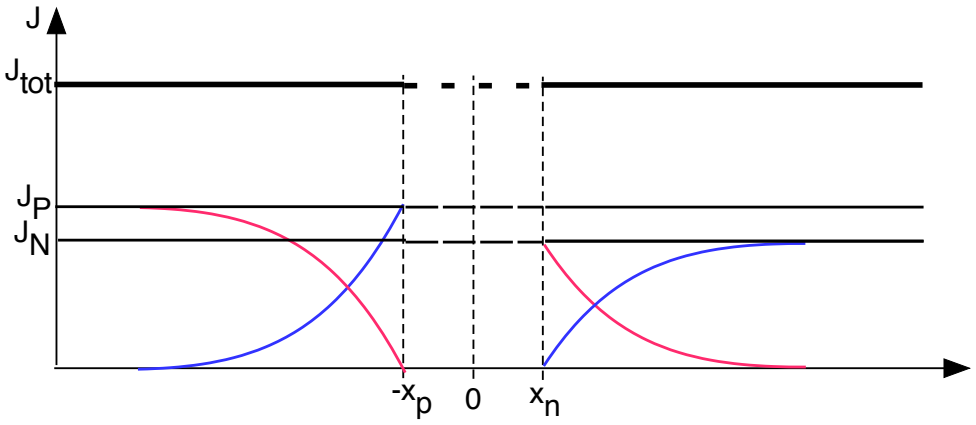


Fig. 5.12. *Currents in a long base diode due to the change of minority charges at the border to the neutral zones. Currents of holes are indicated in red, and blue indicates the currents of electrons. The current in the depletion layer is dashed to indicate that additional processes occurring into the depletion layer have to be considered for a complete picture of the currents across the PN junction.*

Short base diode

The other geometric limit occurs when the distance between the electrode and the depletion layer is shorter than the diffusion length. Namely: $W_B - x_n \ll L_p$ in the N-type material and $-W_E + x_p \ll L_n$ in the P-type material. The solution can be directly derived from eq. 5.40 observing that the short base diode corresponds to a small argument for the exponential functions. Then the solution of the continuity equation can be linearized.
On the other end, the same solution can be obtained considering that in a short base diode the recombination processes carried out by the recombination centers are negligible. Rather, all the injected holes recombines at the ohmic contact. In practice, the excess holes only recombines with the electrons of the metal contact, and the holes current entering at the surface of the contact is compensated by an electron current flowing from the external circuit to the contact.
Under this assumption, $U_p \approx 0$ while the current is still the diffusion current (even when small the semiconductor is a neutral zone):

$$\frac{\partial p'_n}{\partial t} = D_p \frac{\partial^2 p'_n}{\partial x^2} = 0 \tag{5.51}$$

The above equation is settled to zero to indicate the steady state case. The general solution of the previous equation is: $p'_n = A + Bx$. Applying the boundary conditions: $p'(x_n) = p'_{n0}$ and $p'_n(W_B) = 0$ the steady-state profile of holes is:

$$p'_n = p'_{n0}\left(1 - \frac{x - x_n}{W'_B}\right) = \frac{n_i^2}{N_D}\left[exp\left(\frac{qV_A}{kT}\right) - 1\right]\left(1 - \frac{x - x_n}{W'_B}\right) \tag{5.52}$$

Where W'_B is the distance between the contact and the border of the depletion layer.
The linear profile of holes gives rise to a constant diffusion current whose expression is:

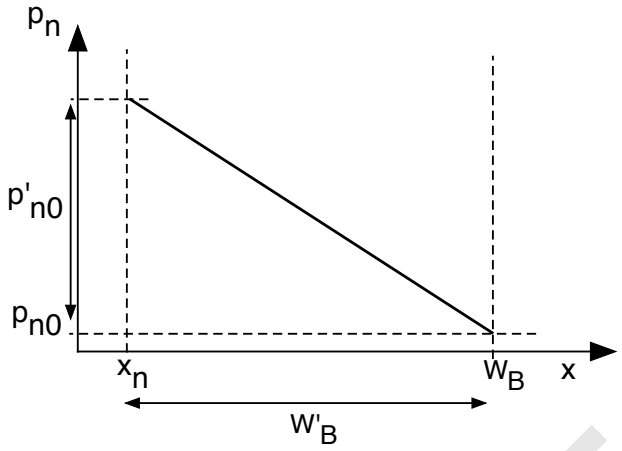


Fig. 5.13. *Profile of holes in the N-type region under forward bias and in a short base diode.*

$$J_p = qD_p \frac{n_i^2}{N_D W'_B} \left[exp\left(\frac{qV_A}{kT}\right) - 1 \right] \tag{5.53}$$

Since there is no recombination, the current is only contributed by the injected holes and there is not a compensating electron current.
A similar behaviour can be calculated for the electrons injected in the P-type material so that the total current is:

$$J_{tot} = J_N + J_P = qn_i^2 \left(\frac{D_n}{N_A W'_E} + \frac{D_p}{N_D W'_E} \right) \left[exp\left(\frac{qV_A}{kT}\right) - 1 \right] \tag{5.54}$$

Note that this equation is formally similar to the current in the long base diode, the only difference is in the typical length parameter involved in the equation. In the short base case the distances between the ohmic contact and the depletion layer (W'_E and W'_B) replace the diffusion lengths (L_n and L_p) of the long base diode.
The dependence of the depletion layer size from the applied voltage introduce a slight dependence of the inverse current from V_A. Another consequence of the short base diode is that since W'_E and W'_B are smaller than L_n and L_p, respectively, the inverse current is larger in the short base diode. In both the cases the inverse current is more contributed by the less doped material.
However, as it will be seen in the next section. the inverse current is actually dominated by the processes of generation and recombination in the depletion layer.

Evaluation of the approximations

Before to continue to calculate the second contribution to the diode current, it is important to discuss the extent of validity of the assumptions on which the ideal current has been calculated. These assumptions are the quasi-equilibrium and the neutral zone.
The quasi-equilibrium condition requires that the applied voltage elicits a small perturbation with

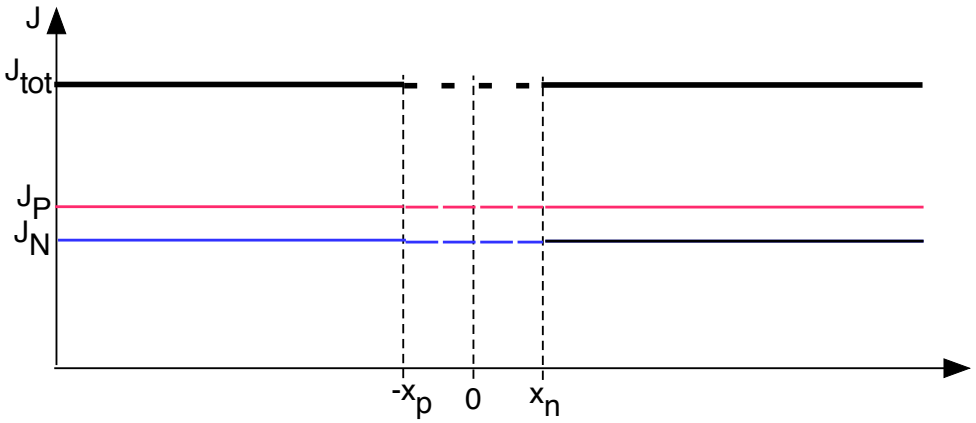


Fig. 5.14. *Currents in a short base diode due to the change of minority charges at the border to the neutral zones. Currents of holes are indicated in red, and blue indicates the currents of electrons. The current in the depletion layer is dotted to mean that additional processes occurring into the depletion layer have to be considered for a complete picture of the currents across the PN junction.*

respect to the equilibrium condition so that it is still valid to continue to calculate the concentrations of electrons and holes using the statistical laws. The neutral zone condition assumes that the voltage drops in the bulks of the semiconductors are negligible.
Numerical examples can help to give reason of the two assumptions.
Let us consider a junction between two pieces of silicon whose doping concentrations are: $N_D = N_A = 10^{16} cm^{-3}$. With such doping levels, the mobility of holes and electrons is: $\mu_n = 1200 \frac{cm^2}{Vs}$ and $\mu_p = 500 \frac{cm^2}{Vs}$. From the mobility, the diffusion constants at T=300K is calculated: $D_n = 31.2 \frac{cm^2}{s}$ and $D_p = 13 \frac{cm^2}{s}$. Let us consider $\tau_n = \tau_p = \tau_0$ and $\tau_0 = 1/(N_T \sigma v_{th})$. Where N_T is the density of the recombination centers, σ is their cross-section and v_{th} is the thermal velocity. With $N_T = 10^{15}\ cm^{-3}$, $\sigma = 10^{-15}\ cm^2$ and $v_{th} = 10^7\ cm/s$ the recombination time is $\tau_0 = 10^{-7} s$. From the diffusion constant and the recombination time we have: $L = \sqrt{D\tau}$ which is: $L_n \approx 17 \mu m$ and $L_p \approx 11 \mu m$. The built-in potential and the depletion layer are calculated with the eq. 5.12 and eq. 5.17: $\phi_i = 0.71V$ and $x_d \approx 0.43 \mu m$.
At the equilibrium, the total current across the depletion layer is zero, however it is the result of the algebraic sum of the drift and the diffusion currents. An estimation of the average diffusion current of holes in the depletion layer is obtained considering the net change of concentrations across the depletion layer:

$$J_{p0} = -qD_p \frac{\Delta p}{x_d} \tag{5.55}$$

where $\Delta p = 10^{16} - 10^4 cm^{-3}$, the the diffusion current of holes is of the order of $J_{p0} \approx 482 \frac{A}{cm^2}$. Under forwards bias, applying the formula of the ideal current with $V_A = 0.6V$ we find that the forward current is $J_{tot} = 0.65 \frac{A}{cm^2}$. Namely, the forward current is more than one hundred times smaller than the currents kept in balance by the thermal equilibrium. Thus forward bias introduces only a small perturbation to the equilibrium state, and then the quasi-equilibrium condition is

plausible.
To verify the neutral zone assumption, it is necessary to calculate the voltage drop in the bulk of the semiconductor. The resistivity are calculated from the mobility and the doping densities ($\rho_N = 1/qN_D\mu_n$ and $\rho_P = 1/qN_A\mu_p$): $\rho_N \approx 0.51\Omega cm$ and $\rho_P \approx 1.25\Omega cm$. To calculate the actual resistance it is necessary to introduce the dimensions of the device. In the case of a long base diode (namely, the worse condition where the resistance is larger) we can assume $W = 100\mu m$ larger enough to extinguish the recombination processes. As the area let us consider $A = 10^{-5}cm^2$.
Then the resistances are: $R_N = 520\Omega$ and $R_P = 1250\Omega$, and the current is $I = J_{tot}A = 3.3 \cdot 10^{-5}A$. Then the voltage drops are $\Delta V_N = R_N I = 0.0029V$ and $\Delta V_P = R_P I = 0.0070V$. Less than 2% of the applied voltage falls in the neutral zones, so more than 98% is found across the space charge region.
Of course as the current increases the voltage drop in the neutral zone cannot be neglected. In this case the actual voltage across the space charge region is $V_A - RI$ where R is the total resistance of the two neutral zones. In this condition the voltage is simply: $V_A = RI + \frac{kT}{q} ln(\frac{I}{I_0} - 1)$.

5.3.2 Generation and recombination current

The current of the ideal diode depends only on the alteration of the concentration of minority charges at the border of the depletion layer due to the applied bias. The ideal current model developed in the previous section accounts for the behavior of the current in forward bias, however to complete the picture it is necessary to take into consideration the processes occurring in the depletion layer. Both in forward and reverse bias the depletion layer is altered because either majority (in forward bias) and minority (in reverse bias) charges across the region altering the equilibrium. It is known that anytime the equilibrium is altered the generation and recombination processes tend to restore it, this holds also in the depletion layer and then the alteration induced by injected charges has some consequence in the total current in the device. It is worth to remind that the density of the recombination centre is constant throughout the material and it does not depend on the doping.
At the equilibrium ($V_A = 0$) the concentration of holes and electrons is significantly smaller than the corresponding doping ($p \ll N_A$ and $n \ll N_D$) but the mass action law is valid ($np = n_i^2$) and then the recombination function is null.
Under bias ($V_A \neq 0$) the concentration of charges change and then the recombination function is no more zero. According to the sign of the applied voltage, the charges changes are different and then the sign of the recombination function is different.
In case of forward bias, the depletion layer is narrowed and due to the injection of majority charges the concentration of holes and electrons increases, as a consequence the recombination processes prevails. On the other hand, in reverse bias, the size of the depletion layer increases and due to the direction of the electron field the residual majority charges in the depletion layer are dragged away from the region. The decrease of mobile charges elicits the generation phenomena.
In order to evaluate these phenomena, it is necessary to calculate the generation-recombination function (U) in these two conditions. Since the concentration of holes and electrons in the depletion layer is very small, the variation induced by bias is not a low-injection process, so the complete form of the U function has to be considered (eq. 4.22 of chapter 4) still using the assumption that the efficient recombination centers lie close to the intrinsic Fermi level, and than the hyperbolic cosine is replaced by 1. Furthermore, let us continue to consider $\sigma_n = \sigma_p$, namely a unique recombination time (τ_0). So the U function is the same for both the sides of the space charge region.

$$U = R - G = \frac{np - n_i^2}{\tau_0[p + n + 2n_i]} \tag{5.56}$$

Generation and recombination processes give rise to an additional current that is calculated from the continuity equation.

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_{GR}}{\partial x} - U \tag{5.57}$$

In case of a steady current, all the derivatives with respect to time are null and then the current is simply given by:

$$J_{GR} = q\int_{-x_p}^{x_n} U dx \tag{5.58}$$

To evaluate U, let us start to estimate the product np. This can be easily calculated at the borders of the neutral zones where the charges concentrations are known.
At the border of the P-type zone at $x = -x_p$ it is:

$$n(-x_p)p(-x_p) = (n_{p0} + n'_p)p_{p0} = \left[\frac{n_i^2}{N_A} + \frac{n_i^2}{N_A}\left[exp\left(\frac{qV_A}{kT}\right) - 1\right]\right] \; N_A = n_i^2 exp\left(\frac{qV_A}{kT}\right) \tag{5.59}$$

similarly at the border of the N-type:

$$n(x_n)p(x_n) = n_{n0}(p_{n0} + p'_n) = \left[\frac{n_i^2}{N_D} + \frac{n_i^2}{N_D}\left[exp(\frac{qV_A}{kT}\right) - 1]\right] \; N_D = n_i^2 exp\left(\frac{qV_A}{kT}\right) \tag{5.60}$$

The product np is the same at the two extremities of the depletion layer. It can take any value inside the region but the concentrations of n and p in the depletion layer changes in the opposite direction, thus it is plausible to assume that their product remains constant. Replacing np in the recombination function we obtain:

$$U = R - G = \frac{n_i^2[exp(\frac{qV_A}{kT}) - 1]}{\tau_0[p + n + 2n_i]} \tag{5.61}$$

It is easy to see that, as previously anticipated, if $V_A > 0$ then $U > 0$ and recombination dominates, and if $V_A < 0$ then $U < 0$ and the generation is the dominant phenomenon.
The function U can be further simplified considering the conditions when it takes its maximum value. Since p and n appears at the denominator of the U function, the recombination and generation phenomena are more efficient where the sum p+n takes its minimum value. At the same time, we have just assumed that the product np is constant. The following two relations lead to calculate the conditions that maximise the function U

$$\frac{d(p+n)}{dx} = 0 \text{ and } \frac{d(pn)}{dx} = 0 \tag{5.62}$$

Developing the derivatives we find:

$$\frac{dp}{dx} + \frac{dn}{dx} = 0 \text{ and } p\frac{d(n)}{dx} + n\frac{d(p)}{dx} = 0 \tag{5.63}$$

From which we obtain that the recombination function is maximum where $p = n$. The behaviour of p,n, and p+n in the depletion layer calculated at the equilibrium are shown in figure 5.14 in case of a symmetrical doping. Under bias, p and n are modified shifting either upward or downward, however their analytical behaviour is preserved.
In the following calculations, let us approximate the U function with its maximum value.

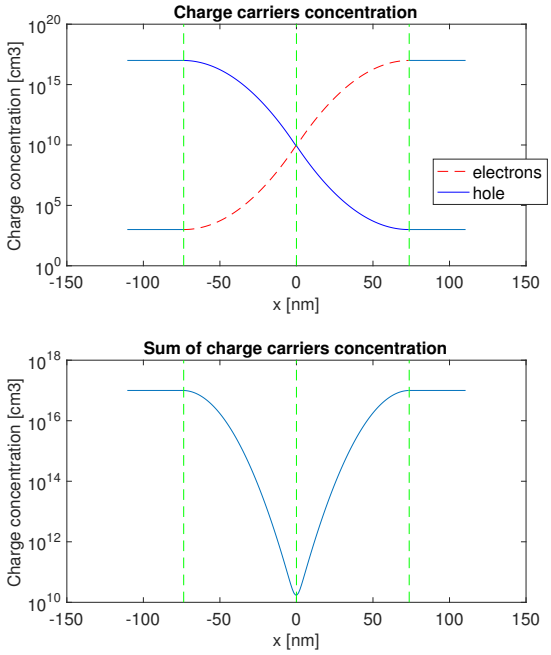


Fig. 5.15. *Charge carriers concentrations in the depletion layer in case of symmetrical doping of* $10^{17}\ cm^{-3}$. *The charge distributions are calculated applying eq. 2.73 and 2.74 where the potential is the built-in potential (eq. 5.9 and 5.10). The sum of electrons and holes get its minimum exactly when p=n. It can be shown that this condition holds also if* $N_a \neq N_D$.

The concentration of holes and electrons where U is maximum is: $p = n = \sqrt{pn} = n_i exp(\frac{qV_A}{2kT})$. Replacing all these conditions in the U function we have:

$$U = \frac{n_i^2\left(exp(\frac{qV_A}{kT}) - 1\right)}{\tau_0[p + n + 2n_i]} = \frac{n_i^2\left(exp(\frac{qV_A}{kT}) - 1\right)}{\tau_0 2n_i\left(exp(\frac{qV_A}{2kT}) + 1\right)} = \frac{n_i\left(exp(\frac{qV_A}{kT}) - 1\right)}{2\tau_0\left(exp(\frac{qV_A}{2kT}) + 1\right)} \tag{5.64}$$

Under forward bias, if $V_A \gg kt/q$, $exp(\frac{qV_A}{kT}) \gg -1$ and $exp(\frac{qV_A}{2kT}) \gg +1$. Then, under forward bias:

$$U = \frac{n_i}{2\tau_0} exp(\frac{qV_A}{2kT}) \tag{5.65}$$

Then, integrating from $-x_p$ to x_n we have the GR current for forward bias (J_{GR}^{FB}):

$$J_{GR}^{FB} = \frac{qn_i}{2\tau_0} x_d exp\left(\frac{qV_A}{2kT}\right) \tag{5.66}$$

The calculations has been performed extending the maximum recombination rate to the whole depletion layer, this leads to a slight overestimate of the actual GR current. However, the simplified calculation correctly describes the exponential dependence from the applied voltage.
Under reverse bias, if $V_A \ll kt/q$, $exp(\frac{qV_A}{kT}) \ll -1$ and $exp(\frac{qV_A}{2kT}) \ll +1$. Then, under reverse bias:

$$U = -\frac{n_i}{2\tau_0} \tag{5.67}$$

Then the reverse bias GR current (J_{GR}^{RB}) is by:

$$J_{GR}^{RB} = -\frac{qn_i}{2\tau_0} x_d \tag{5.68}$$

It is worth to remind that x_d depends on the root square of the applied voltage, then even if not explicitly shown, the current still depends on the applied voltage.
Forward and reverse GR currents can be combined together to give the GR contribution to the total current:

$$J_{GR} = \frac{qn_i}{2\tau_0} x_d \left(exp\left(\frac{qV_A}{2kT}\right) - 1\right) \tag{5.69}$$

Eventually, the total current in the diode is the sum of the ideal and the GR currents:

$$J = J_{ideal} + J_{GR} \tag{5.70}$$

In practice, in order to take into consideration the two currents a non ideality term η is introduced. The value of η is between 1 and 2. Then the current-voltage relationship is written as:

$$I = I_0 \left(exp\left(\frac{qV_A}{\eta kT}\right) - 1\right) \tag{5.71}$$

Figure 13 shows a numerical example. It is related to a silicon PN junction made with $N_D = N_A = 10^{18} cm^{-3}$. The device in this example is in the long base configuration, the recombination lengths are 36 μm in the P-type material and 55 μm in the N-type side.
Since the diffusion current is proportional to n_i^2 while the GR current is proportional to n_i, the relative magnitude of the GR current with respect to the ideal current depends on n_i, which is a function of the energy gap (see eq. 1.30). Then, in small band gap semiconductors such as germanium n_i is large, and the diffusion current dominates over the GR current, as a consequence the ideality factor is close to 1.
Further deviations from the ideality occurs at large applied voltage when the low-injection condition is no more fulfilled. A number of effects occurs under high injection. Among then it is worth to mention that, as mentioned in the previous section, when the current is large the voltage drops in the neutral zone are no more negligible, and this leads to a reduction of slope of the I-V relationship.

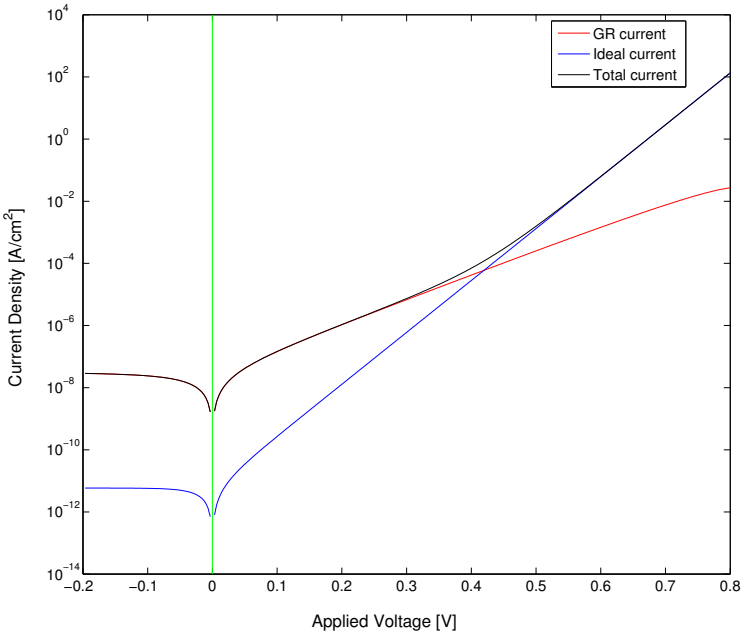


Fig. 5.16. *Numerical example of GR and ideal currents in a silicon PN diode. GR current dominates in reverse bias regime and at small forward biases. In order to allow for a logarithmic representation the currents are plotted as absolute values.*

5.4 Capacitive effects

In chapter 2 the capacitance associated to the metal-semiconductor junction was discussed. In the same way, also in the PN junction capacitance associated to the depletion layer is found. The value of this capacitance can be calculated considering the relationship between the charge in the space charge region and the applied voltage. The connection between these two quantities is contained in the width of the depletion layer that depends on the applied voltage:

$$x_d = \sqrt{\frac{2\epsilon_s}{q}\left(\frac{1}{N_A}+\frac{1}{N_D}\right)(\phi_i - V_A)} \tag{5.72}$$

Since the charge in the two regions is the same ($Q = \mid qN_A x_p \mid = \mid qN_D x_n \mid$), the capacitance can be calculated considering either the P-type or the N-type side of the depletion layer.

$$C = \frac{dQ}{dV_A} = qN_D\frac{x_n}{dV_A} = qN_A\frac{x_p}{dV_A} \tag{5.73}$$

Let us consider the N-type side. the relationship between the extension of the depletion layer in one side and the total size of the depletion layer is:

$$x_p = \frac{N_A}{N_D} x_n; \quad x_d = x_n + x_p; \quad x_n = \frac{N_A}{N_D + N_A} x_d \tag{5.74}$$

Then:

$$C = \frac{dQ}{dV_A} = q \frac{N_D N_A}{N_D + N_A} \frac{dx_d}{dV_A} \tag{5.75}$$

Calculating the derivative of x_d and gathering under the square root all terms except ϵ_s we obtain the same result of the junction capacitance (C_j) of the metal-semiconductor junction:

$$C_j = \frac{\epsilon_s}{x_d} \tag{5.76}$$

It is worth to remind that, as for the other quantities, the above capacitance is a density of capacitance whose units are $Farad/cm^2$. The actual capacitance of the device is obtained multiplying the density of capacitance for the device area.
Such a capacitance is associated to any depletion layer wherever it occurs. In terms of the equivalent circuit a solid-state device always contains resistances and capacitances.
It has to be noted that in forward bias, the depletion layer becomes smaller and the capacitance tends to be large, as a consequence the associated impedance ($Z = \frac{1}{j\omega C}$) becomes negligible. On the other hand, the capacitance becomes important under reverse bias.

Minority charges storage and the diffusion capacitance density

The peculiar current transport processes in the PN junction is manifested by an additional capacitance located in each neutral zone. This capacitance accounts for the charge accumulated in the neutral zones and proportional to the current. Indeed, in order to achieve a steady current is necessary to establish a constant profile of electrons (in the P-type side) and holes (in the N-type side). It is known that charges do not accumulate instantaneously, then from a zero current situation to a forward bias condition there is transit time that is necessary to accumulate the charge profiles. The amount of charges is a function of the applied voltage, then there is a capacitance effect. The time required to accumulate and to deplete these charges determines the response time of the PN junction.
In order to calculated the stored charge let us consider the excess holes in the N-type region.

$$Q_p = q \int_{x_n}^{W_B'} p_n'(x) dx \tag{5.77}$$

In the case of a long base diode we have to replace the excess holes profile (eq. 42), the solution of the integral can be simplified applying the long base diode condition ($W_B - x_n \gg L_p \rightarrow exp(-\frac{W_B - x_n}{L_p} \approx 0$. Then the amount of excess holes is:

$$Q_p = q \frac{n_i^2}{N_D} L_p \left(exp\left(\frac{qV_A}{\eta kT} \right) - 1 \right) \tag{5.78}$$

Replacing the holes current (eq. 45) we obtain a simple relationship between the excess charges and the current:

$$Q_p = J_p \tau_p \tag{5.79}$$

In other words, the current is given by the steady stored amount of charge divided by the recombination time. The in order to achieve the same current the necessary amount of excess charge depends on the recombination time.
In the case of the short base diode, the excess charge profile is given by eq. 49 and the amount of excess charges then is:

$$Q_p = q \frac{n_i^2}{N_D} \left(exp\left(\frac{qV_A}{\eta kT} \right) - 1 \right) \frac{W_B - x_n}{2} \tag{5.80}$$

Considering the current in the short base diode (eq. 50) we obtain the following expression for the amount of stored charges

$$Q_p = J_p \frac{(W_B - x_n)^2}{2D_p} \tag{5.81}$$

were $\tau_{tr} = \frac{(W_B - x_n)^2}{2D_p}$ is the transit time of charges in the neutral zone, namely, it is the time necessary to a charge to travel from the border of the depletion layer to the electrode in case of a diffusion current resulting from a linear profile of excess charges.
This can be easily demonstrated applying the definition of electric current to the diffusion current ($qp\frac{dx}{dt} = qD_p\frac{dp}{dx}$) and considering that for a stationary diffusion current a linear profile of charge is necessary: $p(x) = kx$:

$$D_p \frac{dp}{dx} = p\frac{dx}{dt} \rightarrow \quad \int_0^t dt = \frac{1}{D_p} \int_0^L \frac{p}{\frac{dp}{dx}} dx \rightarrow \; t = \frac{1}{D_p} \int_0^L \frac{kx}{k} dx = \frac{L^2}{2D_p} \tag{5.82}$$

The total charge is proportional to the current and the proportionality term has the dimension of a time. This is the recombination time in case of the long base diode and the transit time in the case of the short base diode. A fast device requires less charge to be accumulated and this is obtained with many efficiency recombination centres in a long diode or with a large doping in case of the short diode.
Eventually, we can conclude that in both cases the charge is proportional to the time that the charge spend in the neutral zone or before recombination or before to be collected by the contact. Since this charge is modulated by the applied voltage it defines a capacitance, it is called diffusion capacitance (C_d). It is worth to observe that since a steady diffusion current requires a steady charge profile, the diffusion capacitance is necessary to the diffusion current. Then we can conclude that a capacitance is associated to any semiconductor and that it is not possible to have a pure resistive semiconductor material.
The diffusion capacitance is defined as $C_d = \frac{dQ_p}{dV_A}$, and the charge can be written as $Q_p = Q_{p0}(exp(\frac{qV_A}{kT}) - 1)$ where Q_{p0} is different in the cases of long or short diode.
Of course the same calculations applies to the electrons stored in the P-type side of the device.Then the diffusion capacitance densities are:

$$C_d^p = \frac{q}{kT} Q_{p0} exp\left(\frac{qV_A}{kT} \right); \quad C_d^n = \frac{q}{kT} Q_{n0} exp\left(\frac{qV_A}{kT} \right) \tag{5.83}$$

Since the doping can be different, the two capacitances may also be different.
Diffusion capacitances are important in forward bias and, as mentioned before, they rule the response time of the device. An equivalent circuit based on lumped elements can be derived considering that the conductivity and the three capacitances depend on the same voltage (V_A). Then the equivalent circuit of the PN junction is formed by a resistor and three capacitors in parallel as in fig. 14.

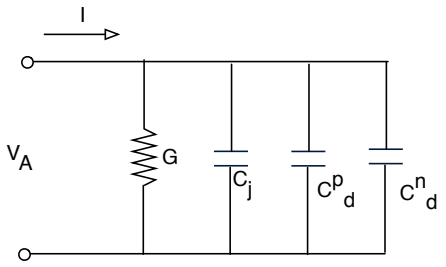


Fig. 5.17. *Equivalent circuit of the PN junction. The conductance G can be calculated from the I/V relationship:* $G = \frac{dI}{dV_A} = \frac{q}{kT} I_0 exp(\frac{qV_A}{kT})$.

The diffusion capacitances make evident the fact that the steady-state is reached only when an amount of charge is steadily present in the neutral zones. The diffusion capacitance rules the behaviour of the diode under forward bias. In particular the transitory times necessary to turn on or turn off a diode depends on the diffusion capacitance. To illustrate these effects let us consider a couple of examples related to the switch of a long-base diode.
In the first example, let us consider the circuit in figure 5.16 where a PN diode is biased by an ideal current generator. The switch is turned on at $t = 0$. At $t = 0^-$ the diode is not biased, then it is in the equilibrium condition with $I_D = 0$. In order to fulfil the Kirchhof law, at $t = 0^+$ the current becomes immediately I_0. In this condition, the voltage across the diode can be calculated inverting the diode characteristics: $V_D = \frac{kT}{q} ln(\frac{I_s}{I_0})$. The former is valid when $I_s \gg I_o$ namely under forward bias. Actually, the previous value is the steady-state value of the voltage. Indeed, let us recall that the ideal current of the diode has been calculated solving the continuity equation under the steady-state assumption ($\frac{dJ}{dt} = 0$). The charge necessary for the current cannot be accumulated instantaneously, then the voltage reaches a steady state value only after a certain time.
Apparently, it is not possible to obtain that the current is immediately equal to I_s and the time of charge accumulation. However, it has to be considered, that the current is not related to the amount of charge but to the derivative of the charge profile, and in particular the derivative calculated at the edge of the space charge region (x_n and $-x_p$). Then the total amount of charge can grow with time according to the current/voltage relationship of the capacitance, but the maximum current can be immediately achieved because it depends on the derivative. Figure 15b shows the progression of charge profile, figure 15c an 15d the voltage and current across the diode respectively.

In order to discuss the effects when the diode is switched off let us consider the example in figure 16a. At $t < 0$ the switch T_1 is off and the switch T_2 is on. In this condition the diode is in forward regime and let us suppose that the diode has been kept in this condition for a time sufficient to establish the steady state. At $t = 0$ T_1 is instantaneously switched on T_2 off, then the diode is found

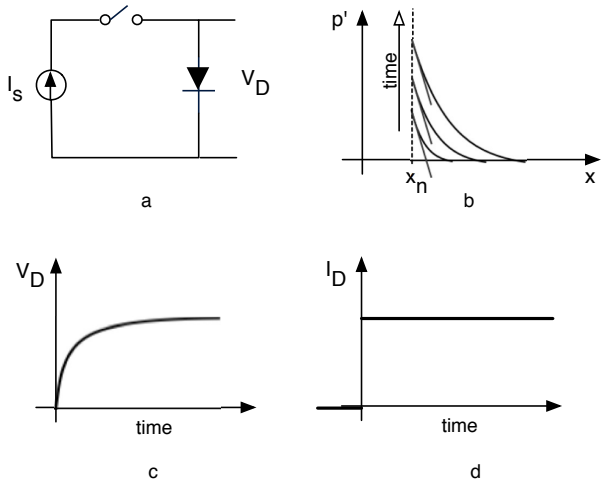


Fig. 5.18. *a: current source bias circuit; b: behaviour of the excess charges in the neutral zone. In figure the holes in the N-type zone are shown. Different curves corresponds to the different times after the switch is turned on. The gradient at x_n is shown, note that that the gradient is constant over the time. As the voltage across the diode increases, x_n is shifted towards left this effect is not represented in figure; c: time behaviour of the voltage across the diode; d: diode current vs time.*

in reverse bias and this condition is kept for $t > 0$.
To switch off a diode means to deplete the neutral zones of the accumulated charges until no excess charges are found in the bulk of the material. The excess charges profiles is due to two phenomena: the injection of minority charges across the depletion layer and their recombination in the neutral zone.
The voltage is applied only to the depletion layer, so when the voltage sign is changed the depletion layer is immediately in reverse bias and then the majority charges of each regions are no more transferred to the complementary zone. Furthermore, the circuit requires that the current flows in the opposite direction with respect to the forward biased diode. The injection of charges in the neutral zone suddenly stops and the amount of charges are the edge of the space charge region decreases. On the other hand, the recombination requires time to consume the excess charges. The decrease at the interface is faster than the recombination, then except for a small region close to the interface the charges profile changes slowly. The decrease of charges at the interface gives rise to an inversion of the gradient and then to the inversion of the current direction. As a consequence, immediately after the switch the current changes sign and maintain a large value until the interface is depleted of excess charges. After that, the recombination eliminates the charges in the bulk and the forward current can decrease towards zero and the only circulating current is the inverse current required by the reserve bias. Figure 16b shows the variable profile of excess charges while in figure 16c the behaviour of the current is shown.

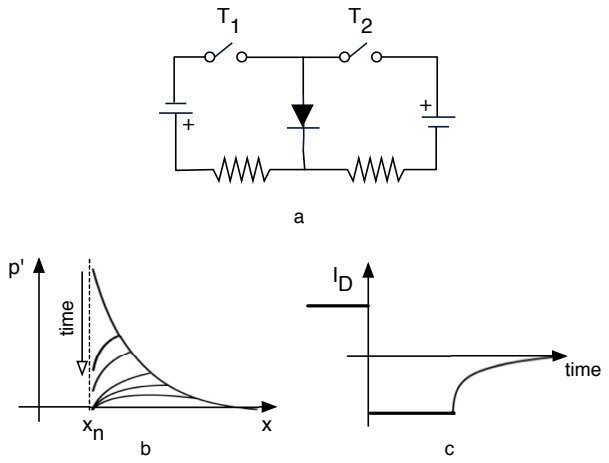


Fig. 5.19. *a: Circuit used to discuss the transition from forward to direct bias. b: behaviour of the excess charge after the transition from forward to reverse condition. Note that the depletion layer actually short twords right. c: behaviour of the ideal current. Consider that this is superimposed to the reverse current that is due to the generation phenomena in the space charge region.*

5.5 Breakdown phenomena

The exponential relationship between the current and the voltage strongly limits the voltage drop in a forward biased diode. In practice, in a silicon PN junction the forward voltage is considered constant at around 0.7 V. Larger voltages results in a divergent value of the current that bring to the thermal breakdown of the device. On the contrary, the reverse current is very small and only weakly variable with the voltage, then the reverse voltage is virtually unlimited. However, since the voltage drops in the narrow region of the depletion layer, the electric field in the space charge region may become very large. This leads to non linear phenomena that are manifested as a sharp increase of the reverse current. These phenomena are generally called breakdown and their origin is due to two distinct causes: the avalanche effect and the Zener effect.
The avalanche effect is a classical phenomena and it happens when the electrons in the depletion are accelerated to a kinetic energy that is sufficient to ionize the atoms to which the electrons impact. The Zener effect is based on the quantum tunnel effect.
Both the effects take place around the interface between the P and N materials where the electric field reaches its maximum value (see figure 4).

5.5.1 Avalanche effect

As previously mentioned, the avalanche effect can be efficiently explained in the frame of the classical physics and it depends on the amount of kinetic energy that an electron acquires under the influence of an electric field. This energy is given by:

$$\Delta E = q \int_0^l \mathcal{E} dx \tag{5.84}$$

where l is the free mean path of the electron, namely the average distance travelled by an electron between two consecutive impacts.
The energy necessary to ionize the impacted atom (E_{gap}) can be roughly calculated considering a free electron impacting with a velocity (v_0) with a lattice atom and producing a electron-hole pair. Let us assume that the three particles (two electrons and one hole) have the same mass, and after the impact they have the same kinetic energy. The from conservation of energy and conservation of momentum we have:

$$\begin{aligned} \frac{1}{2} m v_0^2 &= E_{gap} + \frac{3}{2} m v_f^2 \\ m v_0 &= 3 m v_f \end{aligned}$$

Then the threshold energy to ionize the impacted atom is $\Delta E = \frac{1}{2} m v_0^2 = \frac{3}{2} E_{gap}$. In spite of the very simple model, this is a reasonable value of the expected ionization kinetic energy.
As shown in figure 5.18, the avalanche effect can take place is a restricted region around the interface. The consequence of the ionisation, is that for each impacted electrons two electrons emerge from the impact, these electrons can then be accelerated again to liberate other free electrons giving rise to a sudden and large increase of the reverse current.
The free mean path is an essential quantity to activate the avalanche, indeed larger is the free mean path lager is the energy acquired between two impacts and then more probable is the ionisation of the impacted atom.
The free mean path is strictly related to the mobility and it decreases with the doping concentration. Then in heavily doped devices, the free mean path becomes short and in order to obtain the avalanche is necessary a larger applied voltage. In silicon, when N_D and N_A are of the order of $10^{18} cm^{-3}$ the electric field required for the avalanche is comparable with the saturation field and the probability of the avalanche effect becomes small.

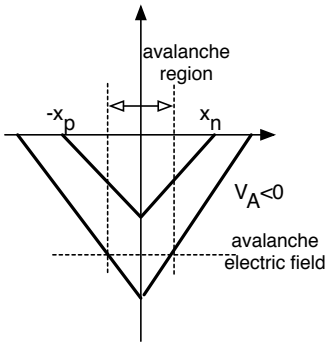


Fig. 5.20. *Behaviour of the electric field at the equilibrium and under reverse bias. The electric field necessary to activate the avalanche is achieved in a region across the interface.*

5.5.2 Zener effect

The Zener effect is the other breakdown phenomena leading to an abrupt increase of the reverse current. Its manifestation is similar to the avalanche effect but it is due to a typical quantum phenomena: the tunnel effect.
The probability of tunnelling becomes relevant when the de Broglie wavelength of the electrons is on the order of the potential barrier width. In heavily doped material ($N_D \approx N_A \approx 10^{18} cm^{-3}$) the depletion layer is of the order of few nanometers. In these conditions, under reverse bias the wavelength of electrons can become smaller enough to be comparable with that of the depletion layer. As a consequence, the transfer of electrons from the valence band of the p-type material to the conduction band of the N-type material becomes possible leading to a large increase of the reverse current. The shape of the barrier is triangular and the probability of transmission of electrons across the barrier is: $T \approx exp(\int_0^l \Delta E)$.
Then, the current depend exponentially on the applied voltage (V_A) then the Zener effect results in a large reverse current that can stabilise the value of the voltage drop across the diode. As shown in figure 5.19, although the depletion layer is larger under reverse bias, the large shift of energy of the electrons in the P-type material makes the barrier shorter favouring the tunnelling of electrons. Furthermore, the electric field in the depletion layer accelerates the electrons increasing their energy and then increasing the transmission probabilility.

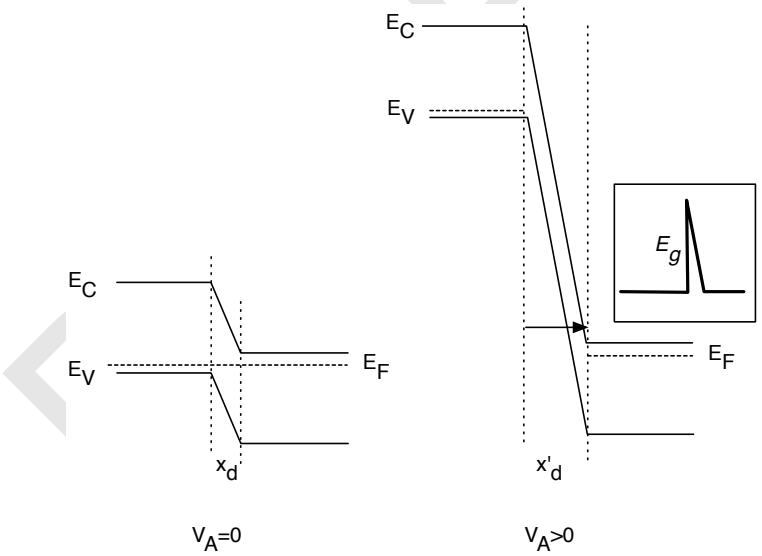


Fig. 5.21. *Band diagram of a PN junction made by heavily doped semiconductors. At the equilibrium ($V_A = 0$) although the depletion layer may be narrow, a null density of states corresponds to the electrons of the valence band. Under a large reverse bias, the density of states accessible to the valence band electrons becomes large and a tunnel current of electrons can flow from the P-type to the N-type material. Due to the large increase of energy in the valence band, in spite of the enlargement of the depletion layer the barrier width becomes smaller than the equilibrium. In the inset, the shape of the triangular barrier is shown. Electrons are accelerated in the depletion layer before to reach the barrier*

The breakdown phenomena are used to design devices that can stabilise the voltage across a passive element. These devices are called Zener diodes even if both the avalanche and the Zener might coexist. It is important to the note that the applied voltage is also fixed in a forward biased diode, but the voltage drop in the forward bias is fixed at around 0.7 V (in case of silicon) while the breakdown voltage in a Zener diode is of the order to Volts. This makes extremely more flexible the use of Zener diode as voltage regulators in practical circuits.
Avalanche and Zener effects show a different behaviour with the temperature. In particular the breakdown voltage at which the Zener effect takes place (BV_{zener}) decreases with the temperature. This can be attributed to the fact that the band gap (E_{gap}) is slightly dependent on the temperature ($\frac{\Delta E_{gap}}{\Delta T} \approx 0.3\frac{mV}{K}$), then the barrier height becomes smaller as the temperature increases and a smaller reverse bias is necessary to activate the breakdown.
On the other hand, the voltage necessary to ignite the avalanche effect ($BV_{avalanche}$) increases with the temperature. This can be understood considering that the mobility decreases with the temperature then also the free mean path decreases. As a consequence, a larger reverse bias is necessary to activate the avalanche.
In silicon devices with $BV \approx 5 \div 6V$ both the effects are possible, as a consequence, a stabilisation of the breakdown voltage with respect to the temperature is observed.

5.6 Numerical example

Let us consider a PN junction formed by a P-type and N-type silicon equally doped with a concentration of $10^{17}\ cm^{-3}$ of acceptors and donors respectively.
Figure 5.21 shows the band diagram at the equilibrium, since the concentration of dopants is the same in both the sides of the junction the depletion layer is equally distributed.

Figure 5.22 shows the results of the Poisson equation in case of the deep depletion hypothesis. The total width of the depletion layer is about 147 nm equally distributed between the two sides of the junction. The maximum electric field at $x = 0$ is $5.68 \cdot 10^4 V/cm$ and the built-in potential is 0.83 V.

Figure 5.23 shows the effects of the applied voltage in terms of depletion layer size, current, and capacitance. Ideal diode current and generation-recombination current are calculated and summed. At $V_A > 0.5\ V$ the ideal current dominates the total current of the device.
The total capacitance is calculated as the sum of the depletion layer capacitance and the diffusion capacitances due to electrons and holes. At $V_A < 0.6\ V$, and in particular in reverse bias, the total capacitance is dominated by the depletion layer contribution.

Figures 5.24, 5.25, and 5.26 shows the band diagram, the equilibrium electrostatic quantities, and the I/V and C/V curves in case of a asymmetric PN junction where $N_D = 5 \cdot 10^{17}\ cm^{-3}$ and $N_A = 10^{16}\ cm^{-3}$. The asymmetric doping results in a asymmetric junction making the PN junction rather close to the metal-semiconductor case.
It is important to observe that in the less doped region, p-type in this example, the intersection between the intrinsic Fermi level and the Fermi level does not occur as the interface (as in the equally doped diode) but inside the depletion layer of the p-type region. This means that there is a region inside the depletion where the p-type material is characterized by a dominance of electrons. This effect is called inversion and it is fully exploited in the metal-oxide-semiconductor devices. The intrinsic condition occurs where the potential is null.

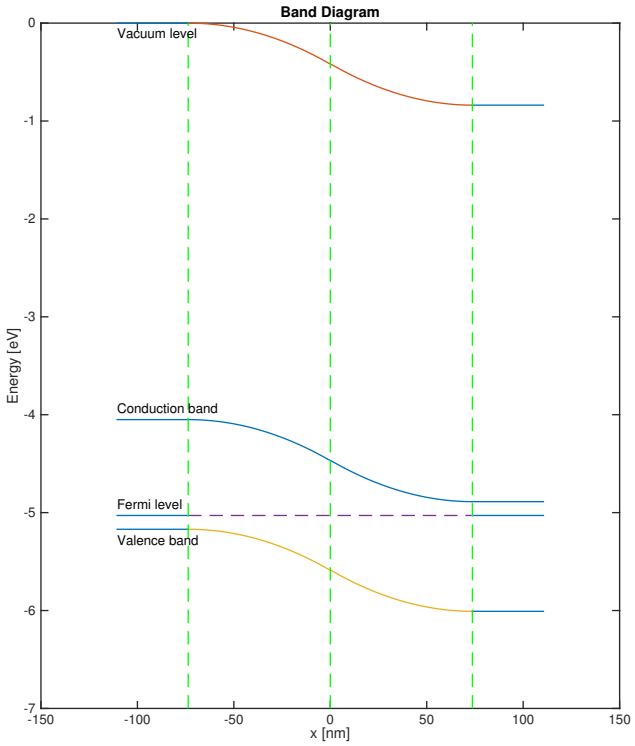


Fig. 5.22. *Equilibrium band diagram of a PN junction equally doped with* 10^{-17} cm^{-3} *dopant atoms.*

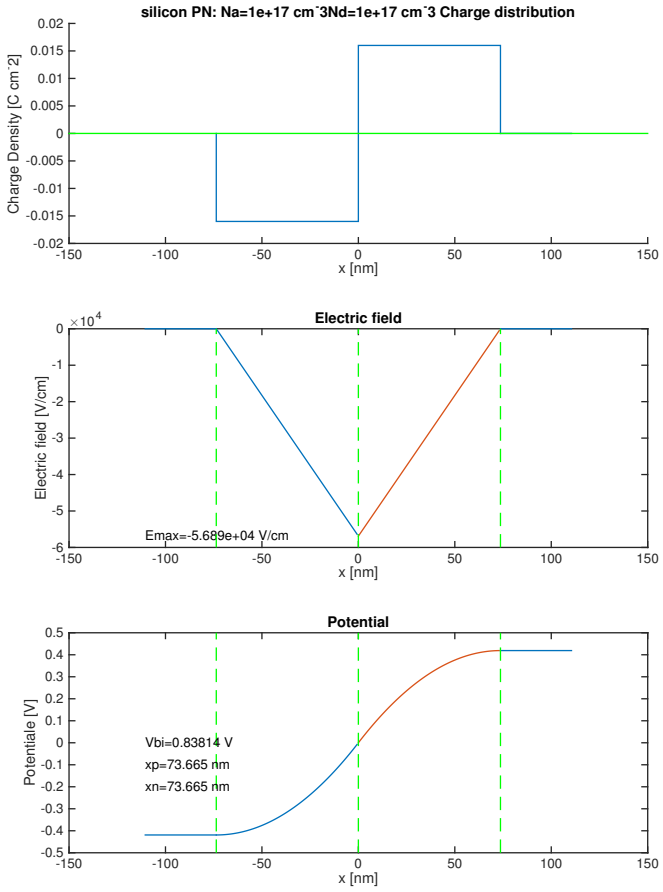


Fig. 5.23. *Charge density, electric field and potential of a PN junction equally doped with* 10^{-17} cm^{-3} *dopant atoms.*

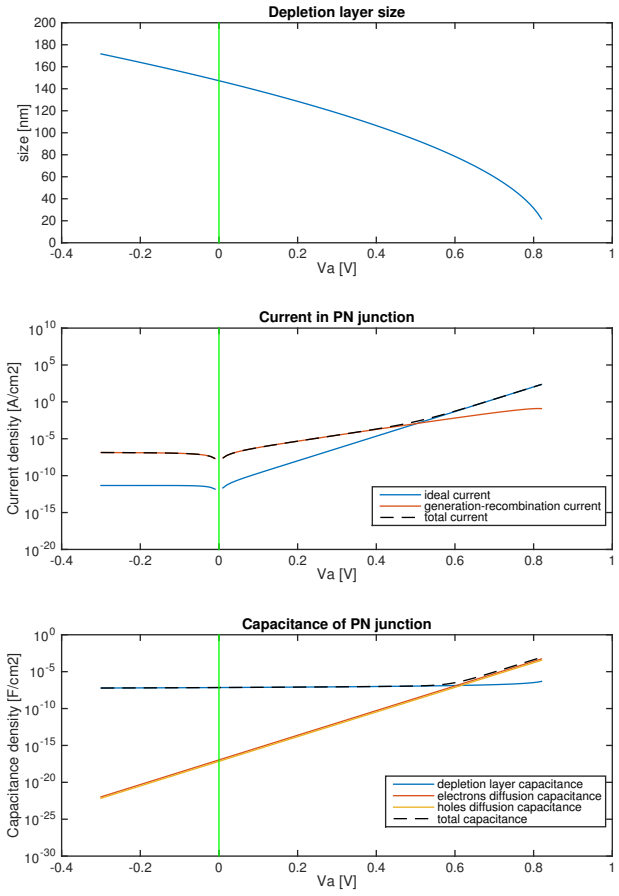


Fig. 5.24. *Effects of the applied voltage in terms of depletion layer size, currents and capacitances of a PN junction equally doped with* 10^{-17} cm^{-3} *dopant atoms.*

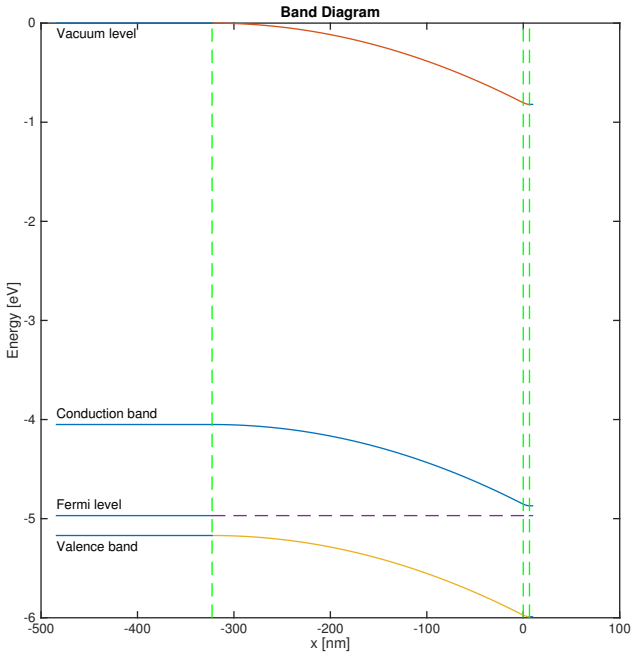


Fig. 5.25. *Equilibrium band diagram of a PN junction asymmetrically doped with* $N_D = 5 \cdot 10^{17}\ cm^{-3}$ *and* $N_A = 10^{16}\ cm^{-3}$.

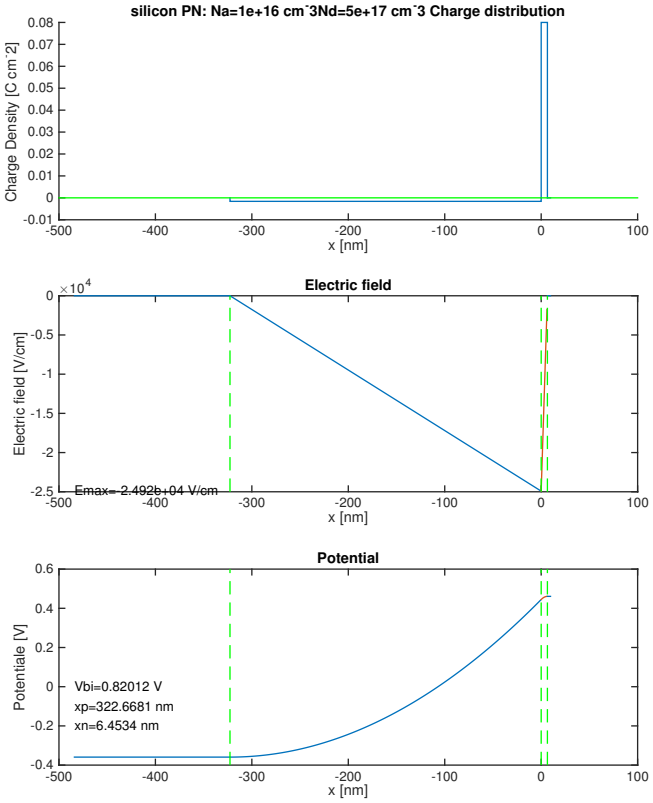


Fig. 5.26. *Charge density, electric field and potential of a PN junction asymmetrically doped with* $N_D = 5 \cdot 10^{17}\ cm^{-3}$ *and* $N_A = 10^{16}\ cm^{-3}$.

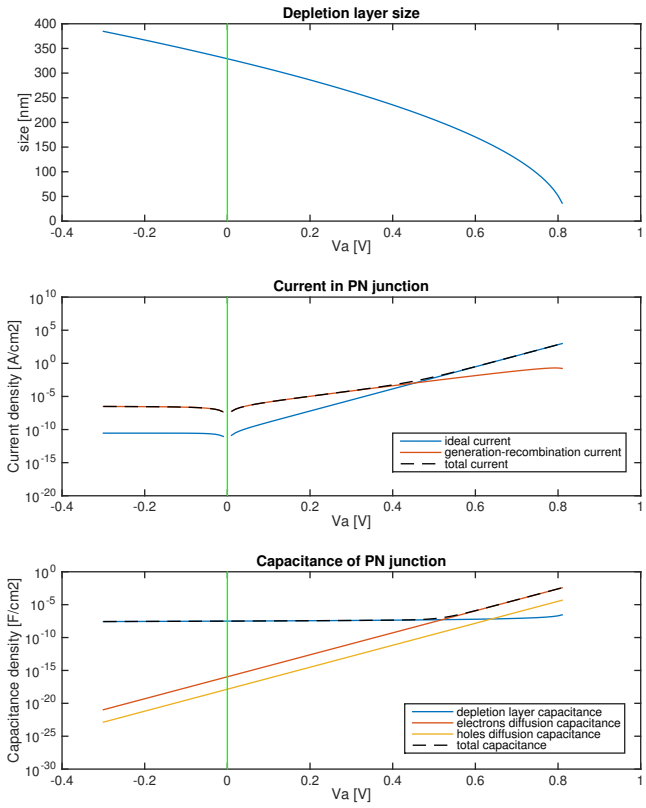


Fig. 5.27. *Effects of the applied voltage in terms of depletion layer size, currents and capacitances of a PN junction asymmetrically doped with* $N_D = 5 \cdot 10^{17}\ cm^{-3}$ *and* $N_A = 10^{16}\ cm^{-3}$.