6

Negative Differential Resistance Effects

6.1 Introduction

NEGATIVE differential resistance (NDR) is a condition found in some devices and materials under proper circumstances. In these components, at a given interval of applied voltage, the slope of the I/V curve is negative. It is straightforward that the definition of negative resistance is only related to the differential resistance ($\frac{dV}{dI}$) but the value of the resistance ($\frac{V}{I}$) is always positive.
The behavior of a device characterized by a NDR can be appraised considering the intersection with a load curve. As shown in figure 1, in the case of NDR the presence of more than one quiescent point leads to an unstable condition in the circuit that may give rise to an oscillatory pattern of the voltage and the current.
The NDR behavior can be found in a number of different devices. Here, the examples of NDR devices are illustrated considering a device such as the *Tunnel diode* and the semiconductors (like the GaAs) where the conduction band exhibits two relative minima.

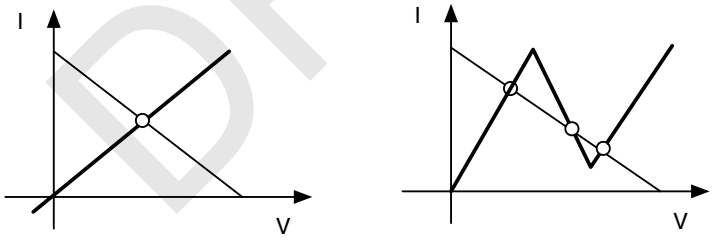


Fig. 6.1. *Example of positive and negative differential devices and their quiescent points when they are loaded by a load curve.*

6.2 Tunnel Diode

The tunnel diode is a PN junction made by heavily doped semiconductors. The doping in this device is large enough to displace the Fermi level into the conduction and the valence bands as shown in figure 6.2. Such materials are called degenerate, and in practice they behaves like a metal but with the great difference that the charge carriers are still electrons and holes.
In these conditions, the Boltzmann approximation of the Fermi-Dirac function is no more valid and a thorough description of the device requires the use of the complete equation.
Due to the heavy doping, the space charge regions are very narrow and the tunnel effect drives the crossing of the barriers.
However, since also the tunnel current needs available free arrival states, the current depends on the correspondence between filled and empty states across the junction. Such a device shows the same I/V curve of a standard diode but with the important addition of an extra current due to the tunnel effect.

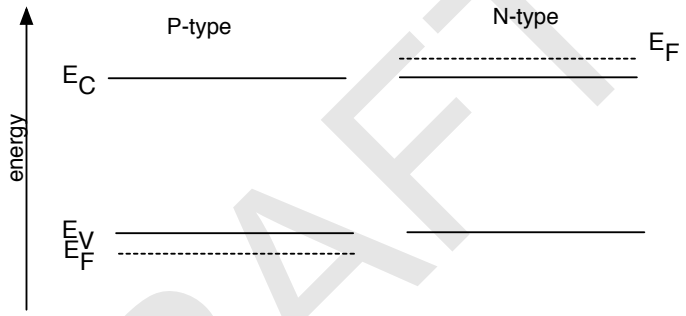


Fig. 6.2. *Band diagram of heavily doped till degenerate P and N type semiconductor.*

The behavior of the tunnel current and the origin of the NDR can be illustrated by a simple qualitatively discussion. For the scope, let us consider the tunnel diode in the hypothetical condition of T=0 K. In this situation, only the states lying below the Fermi level are filled.
At the equilibrium (fig. 6.3a), filled states correspond to the electrons in the valence and conduction bands of P and N materials respectively. Then, even if the built-in potential cannot stop the tunnel current, the current is actually null. In reverse bias the energy of electrons in the P-type material is shifted upward (fig. 6.3b) and the electrons in the valence band see empty states in the conduction band of the N-type semiconductor. This allows for a tunnel current from the P-type to the N-type semiconductor. This is a reverse current for the diode. Since at growing negative voltage the amount of empty density of states that can be reached by tunnelling increases, also the current increases with the applied voltage as found in the Zener effect.
On the other hand, under forward bias, the energy of the electrons in the P-type material is shifted downward. This makes a number of states in the valence band available to the electrons of the conductance band of the n-type side (figure 6.3c). The increase of current is proportional to the available density of states that reaches the maximum value when the valence band of the P-type material is aligned with the Fermi level of the N-type semiconductor. It is easy to observe that this

current has a opposite direction with respect to the reverse current, than it is a forward current. The forward current grows until the top of the valence band is aligned with the Fermi level of the electrons in the N-type material ($V_A = V_P$). Beyond this point, the accessible density of states in the valence band of the P-type semiconductor decreases and so the current, until $V_A = V_0$ where the bottom of the conductance band aligns with the top of the valence band and the tunnel current becomes definitely zero.

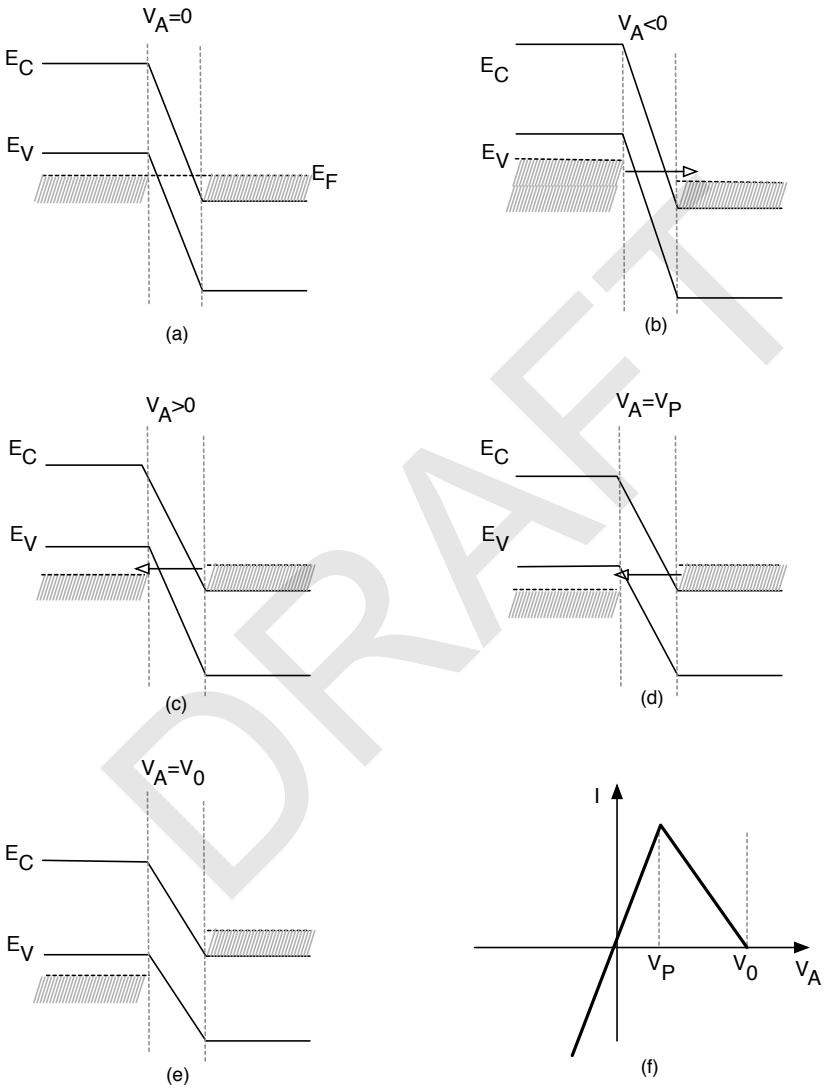


Fig. 6.3. *The band diagram of a tunnel diode under different applied voltages. In the last figure the contribution of the tunnel current to the I/V curve of the device is shown.*

The tunnel current is additive respect to the current of the standard diode, and the total I/V curve of the device follows the shape shown in figure 4. It is important to consider that since the built-in potential is much larger than in a normal diode ($\phi_i > E_{gap}$), the ideal current overcomes the tunnel current well beyond V_0 so that the tunnel current can dominate at small voltages. Typical values of V_0 are of the order of hundreds of mV. On the other hand, due to the very narrow depletion layer the generation recombination current is negligible and the reverse current is made by the tunnel current such as a Zener effect with a null breakdown voltage.

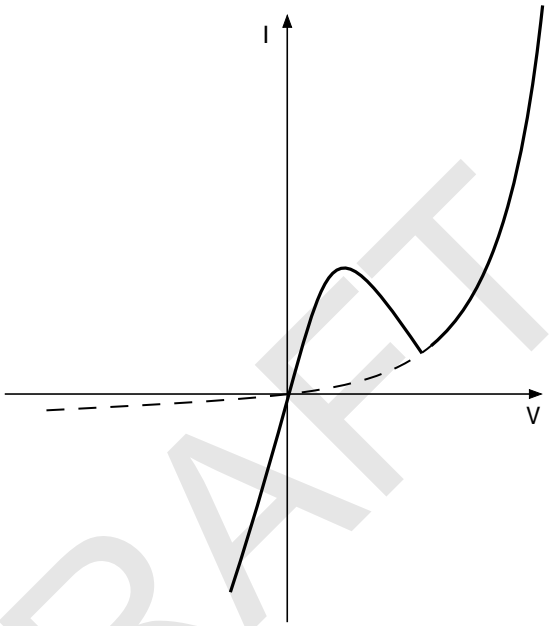


Fig. 6.4. *Complete I/V curve of a tunnel diode.*

6.3 NDR behavior in GaAs

The NDR behavior is spontaneously observed in homogeneous semiconductors made of atoms of III and V group of the periodic table of the elements such as GaAs and InP.
These materials are characterized by a peculiar shape of the bands in the k space where the conduction band exhibits a structure with a double minima. In GaAs, the first minimum occurs at 1.42 eV above the top of the valence band with $\Delta k = 0$ (direct band-gap), and the second minimum is displaced in k and it is located at about 0.3 eV above the first minimum. The curvature of the conduction band in the neighbor of the second minimum is smaller than the curvature of the main minimum, then the effective mass is greater and the mobility is smaller.

The transition from the two minima requires a change of the momentum, that can be achieved when the electrons are accelerated by a large electric field. In practice, as the electric field grows,

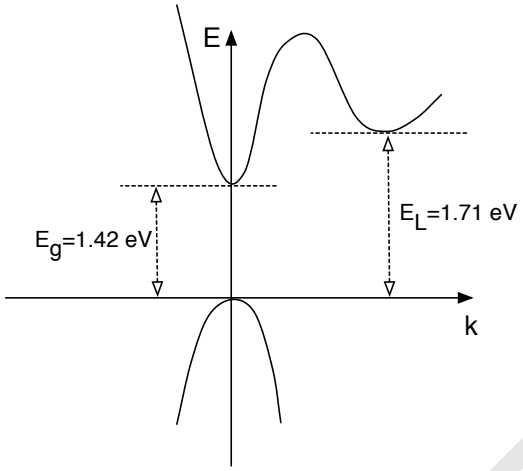


Fig. 6.5. *Simplified band diagram of GaAs. The conduction is characterized by two valleys with different curvatures. As a consequence the electron effective mass in the two valleys is different as well the mobility.*

the concentration of electrons in the second valley becomes numerically important, and part of the conduction electrons move with a smaller mobility.
The electric field necessary to populate the second valley is large enough that the drift velocity is comparable with the thermal velocity. In other words, the "temperature" of the electrons becomes larger than the temperature of the lattice. Electrons in this conditions are called *hot electrons*.

$$T = \frac{1}{2} m (v_{th}^2 + v_d^2) \tag{6.1}$$

In order to calculate the relationship between the average velocity and the electric field, let us assume that there is a concentration n_1 of electrons in the deepest valley while the concentration in the second valley is n_2. The curvatures (the second derivative of the band profile) of the band in the two valleys are different, and so the electrons in the two valleys experience a different effective masses. Eventually, the two populations of electrons move with different mobilities μ_1 and μ_2.
The current is given by the sum of the contributions of the two groups of electrons:

$$J = q(n_1\mu_1 + n_2\mu_2)\mathcal{E} = q(n_1 + n_2)v \tag{6.2}$$

from which

$$v = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}\mathcal{E} = \frac{\mu_1(n_1\mu_1 + \frac{\mu_2}{\mu_1}n_2)}{n_1 + n_2}\mathcal{E} \tag{6.3}$$

Let us assume that $\mu_2 \gg \mu_1$ then

$$v = \frac{\mu_1}{1 + \frac{n_2}{n_1}}\mathcal{E} \tag{6.4}$$

Since the interval of energy under consideration are well above the Fermi level, the ratio between the concentration of electrons in the two valleys can be calculated with the Boltzmann relationship,

$$\frac{n_2}{n_1} = R\ exp\left(-\frac{\Delta E}{kT_e}\right) \tag{6.5}$$

Where R is the ratio between the densities of the states in the two valleys and T_e is the effective temperature of the electrons. Replacing the last equation in eq. 6.4 we get:

$$v = \mu_1 \mathcal{E}\left[1 + R\ exp\left(-\frac{\Delta E}{kT_e}\right)\right]^{-1} \tag{6.6}$$

ΔE is the energy lost by the electrons in the scattering with the lattice atoms. This can be written in terms of the electrons effective temperature (T_e) as: $\Delta E = \frac{3}{2}kT_e$ Then, recalling the scattering processes outlined in chapter 1 we can write:

$$\Delta E = \frac{3}{2}kT_e = q\mathcal{E}l_c \tag{6.7}$$

where $l_c = v\tau_c$ is the mean free path.
Finally, the effective temperature can be calculated:

$$T_e = \frac{2q\tau_c}{3k}\mathcal{E}^2\mu_1\left[1 + R\ exp\left(-\frac{\Delta E}{kT_e}\right)\right]^{-1} \tag{6.8}$$

Equations 6.6 and 6.8 form a couple of parametric equations. In practice for each value of T_e the corresponding $\mathcal{E}$ and v are calculated and a curve of the velocity vs. the electric field can be drawn. The result is shown in figure 5.
The curve is characterized by a NDR tract occurring beyond the peak. In the case of gallium arsenide the peak occurs at $\mathcal{E} \approx 4 \cdot 10^4 \frac{V}{cm^2}$, and the corresponding velocity is $v_{peak} = 2 \cdot 10^7 \frac{cm}{s}$. The NDR tract converges towards a saturation velocity $v_{sat} \approx 10^7 \frac{cm}{s}$.

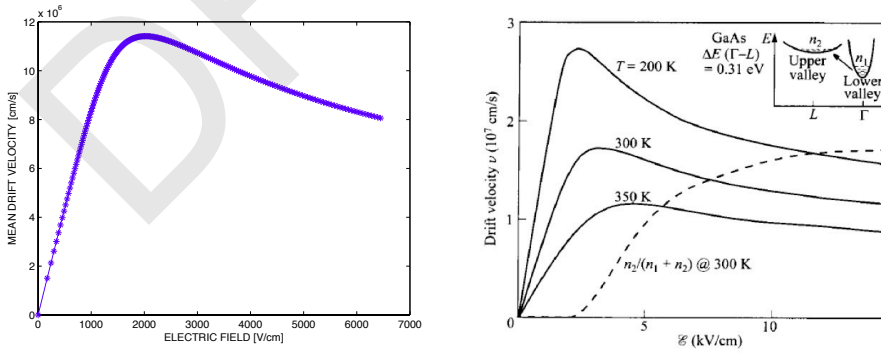


Fig. 6.6. *Left: calculated velocity vs. electric field. The calculus has been done considering T_e in the range 300 K - 700 K, $\tau_c = 10^{-12}s$, $R = 10^3$, and $\mu = 8800cm^2V^{-1}s^{-1}$. Right: velocity vs. electric field in GaAs at different temperature.*

The saturation velocity is more than an order of magnitude larger than in the silicon, this provides the basis for the largest frequency of operation of the devices based on III-V semiconductors. This NDR tract in III-V semiconductors, and in particular in GaAs, is at the basis of a complex phenomenon called Gunn oscillation that is exploited to generate high frequency signals.

6.4 Gunn oscillations

The Gunn oscillations appears in a homogeneous resistor made of a III-V semiconductor (such as GaAs) biased with a voltage such that the electric field exceeds the peak value in the velocity/electric field relationship.
In this condition the material becomes unstable, namely small fluctuations in the supposed uniform concentration of electrons are not canceled but they rather grows and propagate until to reach the electrode where they give rise to high frequency current components. A simple explanation of the phenomenon can be obtained considering the effects of electron concentration fluctuations on the electric field distribution and then on the velocity of the charge carriers.
The static situation is depicted in figure 6.7. The concentration of electrons is uniform and then also the electric field is uniform and its value lies beyond the peak of the velocity/electric field relationship. Namely, the material is in the negative differential resistance regime.

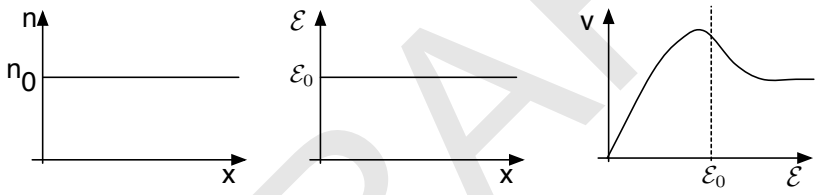


Fig. 6.7. *Initial conditions for Gunn oscillations.*

Let us suppose that a small fluctuation of concentration occurs in a position close to the negative electrode (see figure 6.8). The charge fluctuation corresponds to a double layer of charges that generates a fluctuation in the electric field that is summed to the electric field due to the applied voltage. Since the applied voltage is constant, the integral of the electric field along the material is also constant, then a local increase of the electric field is balanced by a decrease of the electric field in the rest of the semiconductor. In other words, the electric field decreases everywhere except in the small region interested by the fluctuation where it increases. The material is biased in the NDR condition, then the electrons interested in the fluctuation decrease their velocity while the velocity of the rest of the electrons increases.

This condition leads to a progressive enforcement of the density of electrons interested by the fluctuations. The fluctuation tends to diverge as the electric fields of perturbed and non perturbed regions becomes more different one each other. In particular, the smallest electric field may beecome smaller than the peak value and then the electrons outside the perturbation begin to behave according to a positive differential resistance losing their velocity. Eventually, the velocity of all the electrons (perturbed and non perturbed) equalize and the density of perturbed electrons stops to

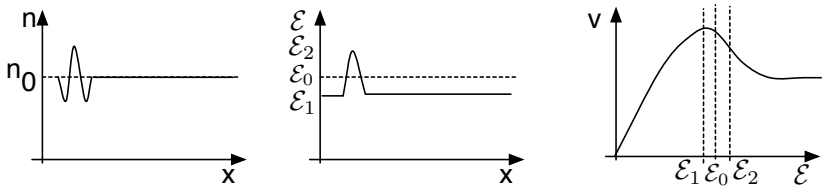


Fig. 6.8. *Perturbed electrons are slowed down while the rest of the electrons accelerate .*

grow and a stable dominion is formed (figure 9).

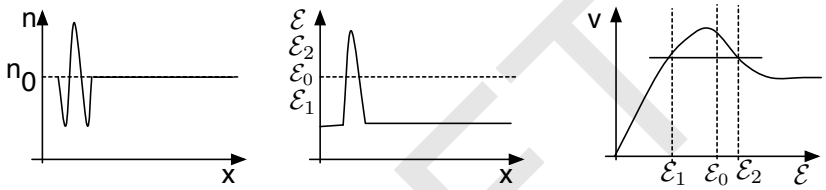


Fig. 6.9. *The perturbation converges to a stable condition when the velocity of the electrons becomes the same.*

The dominion moves towards the positive electrode where it elicits a temporary increase of current. Fluctuation appears spontaneously in the material then a train of current spikes is observed. Eventually, although the resistor is biased with a d.c. voltage the current is characterized by a stable a.c. component which occurs in the microwave spectral region.
Gunn oscillations are exploited in a device called Gunn diode that is used as a microwave signals generator.

7

Bipolar Junction Transistor

7.1 Introduction

The Bipolar Junction Transistor (BJT) is a direct consequence of the properties of the PN junction. It is made by a cascade of two PN junctions with three contacts. The main application of the device is the current amplification, but it is also used as a switch, commuting from high to low current states.
The concept of the BJT stems from the observation that in a PN junction the doping of the two regions modulates the relative intensity of the currents carried by electrons and holes. In practice, changing the doping it is possible to make the current dominated either by electrons or by holes. However, the PN junction has two terminals, then, since the collected current is the sum of the currents carried by electrons and holes, which of the charge carriers actually dominates remains invisible to the external circuit.
The ideal structure of a BJT is shown in figure 1. It is a sequence of NPN (or PNP) materials defining two junctions. Each portion of the device has its own contact then the two junctions can be separately biased. The three regions are called emitter, base, and collector and the two junctions are the base-emitter and the base-collector junctions. V_{BE} and V_{BC} are the corresponding voltage drops across each junction respectively.
The central part of the device is the base, this region exchanges charges through three contacts: two junctions and a metallic terminal. Let us consider that one of the junctions defines a normal diode, then the charge injected into the base region can be extracted at two contacts. One is the metal electrode than accepts both electrons and holes, and the other PN junction that accepts holes or electrons (namely base minority or majority charges) according to the sign of the applied voltage. When this junction is reversely biased it drags the minority charges of the base across the junction. In this condition, the two contacts of the base are selective: minority charges are collected at the junction and majority charges are collected by the metal contact.
The inverse current in a PN junction is orders of magnitude smaller than the forward current. Anyway there are some conditions when this current may become large such as a photodiode where the generated electron-hole pairs are drifted away from the electric field in the space charge region and then contribute to the inverse current. In the BJT, a forward biased junction injects in the base a large excess of minority charges that can be collected at the other junction if this is reversely biased. The majority charges are of course collected at the metallic contact. This behavior occurs if the base is short enough to avoid the recombination of the injected minority charges allowing

the electrons to reach the opposite junction. We will see that the geometric width of the base is a fundamental parameter for the device performance.
In the following of this chapter, the discussion about the BJT properties is carried out for a NPN configuration. A completely symmetric behavior is obtained with the PNP sequence of materials. This complementarity leads to an important degree of freedom in circuits design.

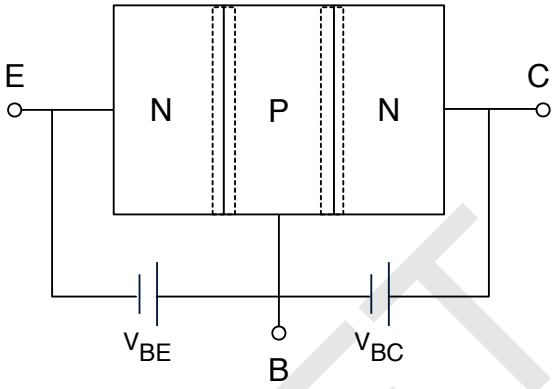


Fig. 7.1. *Scheme of principle of a BJT.*

The geometrical arrangement of the BJT makes the holes contribution to the current crossing the device from the emitter to the collector negligible with respect to the current carried by the electrons under any combination of V_{BE} and V_{BC} . Of course, the role of electrons and holes is exchanged in a PNP device. Then in a BJT the majority charges of emitter and collector regions are the dominant carriers of the device.
Figure 2 shows the holes current in the three possible bias configurations where the junctions are biased either forward or reverse.

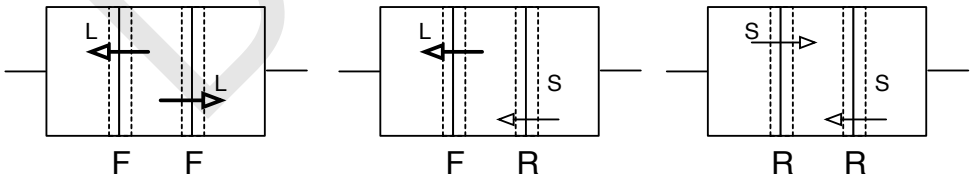


Fig. 7.2. *Holes contribution to the current flowing from the emitter to the collector in the three junctions bias conditions. In the first case two large holes currents are oppositely injected from the base to the two adjacent regions. In the second case one of the current is large and the other is small then the small current dominates. Finally in the third region a small holes current is injected from the two adjacent regions towards the base. In all cases the total holes current is small and then negligible.*

7.2 The ideal transistor

The current flowing from the emitter to the collector can be studied on the basis of the models developed for the PN junction. For this scope let us consider the ideal structure shown in figure 1. Emitter and collector regions are more doped with respect to the base ($N_D > N_A$) and the width of the base is much shorter than the recombination length of the electrons in the base. Under these conditions, the base region can be treated as a short base diode and the recombination of the charges injected from forward biased junctions is a rather unlikely event. Noteworthy, the ideal device is completely symmetric, namely the emitter and the collector can be exchanged. Later we will see some reasons leading to a non symmetrical device where the collector properties are different from those of the emitter.
For sake of simplicity let us consider $N_D = 10^{16} cm{-}3$ and $N_A = 10^{14} cm{-}3$.
Figure 3 shows the equilibrium band diagram and the concentration profile of electrons.

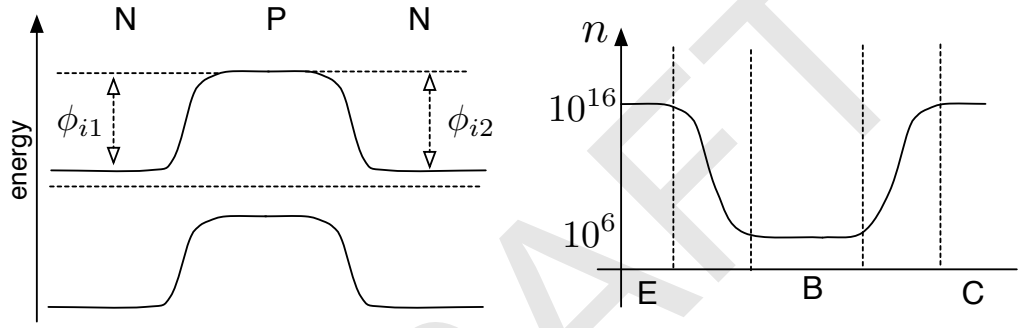


Fig. 7.3. *Left: equilibrium band diagram. Right: electrons concentration profile. In base the electrons concentration is given by the mass action law:* $10^{20}/10^{14}$.

Figure 4 shows the band diagram and the electrons concentration profile in case of reverse-reverse, forward-forward, and forward-reverse bias of the emitter-base and base-collector junctions respectively. As usual, the applied voltage is only distributed across the space charge regions, then the bulk of the semiconductors are neutral zones where the electric field is null.
In case of reverse-reverse bias, the base is depleted of electrons and a small electrons current flows between emitter and collector. In the case of forward-forward bias the base is over populated by electrons. In this condition a large current of electrons can be expected. Finally, in the case of forward-reverse bias the largest gradient of electrons concentration is found in base.
This last case is named active zone. The electrons injected from the emitter-base junction that is forward biased are completely collected at the base-collector junction which is under reverse bias. The collector junction plays the role of a electric contact but with the important difference that it can drag only electrons. The active zone is the configuration where the BJT behaves as a current amplifier.

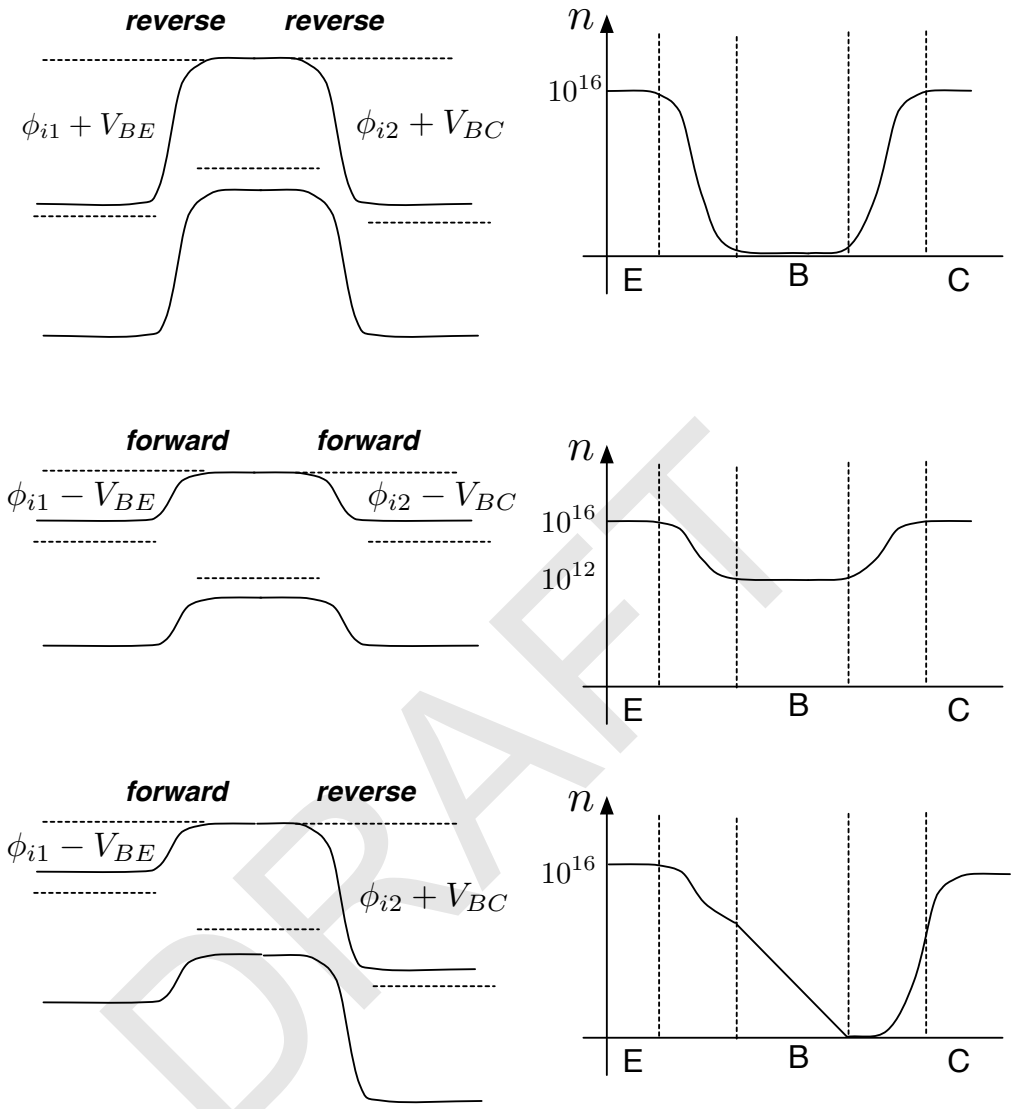


Fig. 7.4. *Band diagram and electrons concentration profile under the three different bias scheme that can be applied to the prototype BJT.*

7.2.1 Electrons current in active zone

The active zone is a bias scheme where the two junctions are oppositely biased. The junction base-emitter (BE) is forward biased and the junction base-collector (BC) is reverse biased. Then electrons from the emitter, where they are majority charges, are injected into the base where they are minority charges. Reverse current flows in the base-collector junction then the electrons injected in the base are collected by the BC junction and injected into the collector. This is the basic transistor mechanism: the direct current of a diode is collected by a reverse biased diode. The reverse biased BC junction transfers the electrons from the base to the collector and the holes from the collectors to the base. Then from the point of view of the base, the collector in active zone can only accept electrons. In practice, the bipolar conductivity of the diode is broken into a unipolar current. In the NPN device, the current flowing from collector towards emitter is mainly due to electrons. Later we will discuss the holes contribution, for the moment let us calculate the electrons current in active zone. The capability of the reverse biased BC junction to collect electrons depends on the width of the base. Indeed in case of a long base, all the injected excess electrons would recombine in the base and the electrons current is converted into a holes current. The BC junction being reversely biased cannot provide a large current of holes so the forward current injected from the BE junction is not collected at the collector terminal. Then in order to collect current at the collector it is necessary that the base is short with respect to the recombination length of the electrons in base. Then the first requirement for the BJT is that the base region has to be short, then the electrons injected from the emitter experience the short-base diode condition. The BC junction gathers the injected excess electrons and then for the electrons, but non for the holes, it is equivalent to a metal contact. The concentration of excess injected electrons is shown in fig. 5.

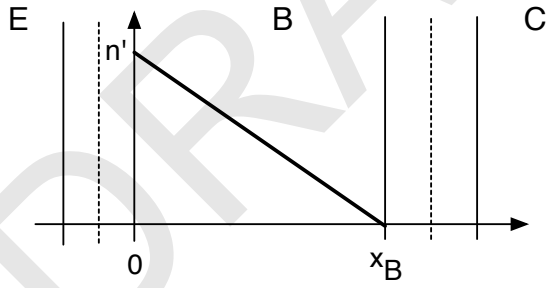


Fig. 7.5. *Profile in base of the concentration of the excess electrons injected from the emitter. The coordinate starts from the edge of the depletion layer of the BE junction (x=0) to the edge of the depletion layer of the BC junction (x_B).*

The equations calculated for the short base diode are still valid and the profile of injected electrons is linear

$$n'(x) = n_{p0}\left[exp\left(\frac{qV_{BE}}{kT}\right) - 1\right]\left(1 - \frac{x}{x_B}\right) \tag{7.1}$$

under forward bias $qV_{BE} \gg kT$ then

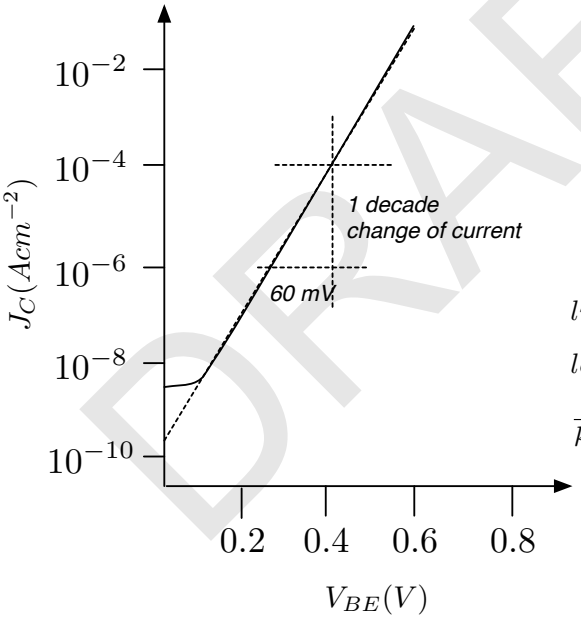
$$n'(x) = n_{p0} exp\left(\frac{qV_{BE}}{kT}\right)\left(1 - \frac{x}{x_B}\right) \tag{7.2}$$

This distribution produces a diffusion current of electrons:

$$J_n = qD_n \frac{dn'}{dx} = -qD_n \frac{n_i^2}{N_A}\frac{1}{x_B} exp\left(\frac{qV_{BE}}{kT}\right) \tag{7.3}$$

Then: $J_n = -J_S exp(\frac{qV_{BE}}{kT})$ with $J_P \approx 0$.
The current at the collector terminal is opposite to the electrons current: $J_C = -J_n$. Then the current collected at the collector terminal is controlled by the voltage applied to the BE junction. Figure 6 shows an example of J_C as a function of V_{BE}, the exponential behaviour is valid for about 7 decades of current density values.
The theoretical value of the current at $V_{BE} = 0$ gives the experimental estimation of J_S.

$$J_S = qD_n \frac{n_i^2}{N_A x_B} \tag{7.4}$$



$$ln\ J_C - ln\ J_S = \frac{qV_{BE}}{kT}$$
$$log\ J_C = log\ J_S + \frac{q}{kT} ln\ 10\ V_{BE}$$
$$\frac{q}{kT} ln\ 10 = 60\ mV\ at\ T = 300\ K$$

Fig. 7.6. *Example of the comparison between experimental and theoretical behaviour of the collector current versus the base emitter voltage (temperature T=300 K). J_S is the theoretical value of the current at $V_{BE} = 0$ the actual value is approximately one order of magnitude larger.*

An important quantity in the BJT is the total current stored in the base region Q_{B0} in case of homogenous doping this is simply given by: $Q_{B0} = qN_A x_B$ in general the doping is not constant, then the charge in the base is given by:

$$Q_{B0} = q \int_0^{X_B} p dx \tag{7.5}$$

If the doping is not constant also the diffusion constant is not a constant, in order to take into account this condition the diffusion constant is replaced by a average quantity indicated as $\tilde{D}_n$. Using the total charge and average diffusion constant, the term J_S can be written as:

$$J_S = \frac{q^2 \tilde{D}_n n_i^2}{Q_{B0}} \tag{7.6}$$

Q_{B0} is a characteristics parameter of the BJT, it is established during the physical fabrication of the device and it is given by the total amount of holes in the base. The total charge in one of the region of the BJT is called Gummel Number (GN). The Gummel number of the base is given by:

$$GN_B = \int_0^{X_B} N_A(x) dx = \frac{Q_{B0}}{q} = \frac{q^2 \tilde{D}_n n_i^2}{J_S} \tag{7.7}$$

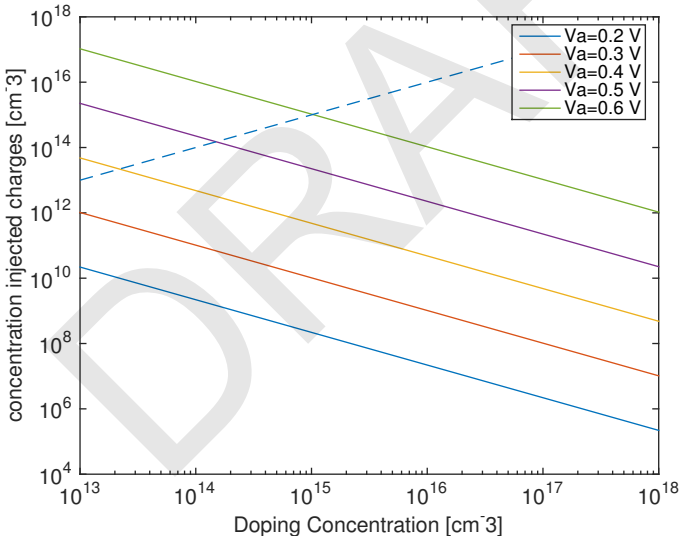


Fig. 7.7. *Charges injected at the emitter interface as a function of the doping of the base. The quantity is calculated for different applied voltages. The calculation is strictly valid only when the concentration of charges is smaller than the doping, namely in the low injection conditions. However it gives the idea that the low injection condition is not valid a low doped base.*

The base Gummel number (GN_B) is a quality factor of the BJT that defines the current of the device. A small base Gummel number is necessary for a large current.

A small GN_B can be obtained in two ways. The more immediate method suggests to use a small doping N_A, however this method may jeopardise the validity of the low-injection current. Figure 7 shows the behavior of the ratio of injected charges at the interface with respect the doping as a function of the doping. The plot is only quantitative because when the low injection is violated the model is not valid, however it provides a evidence that the low injection limit is violated at low base doping concentration. The second method is based on a variable doping concentration profile, larger in the region close to the emitter and smaller elsewhere. The low injection condition is critical at the interface where the electrons are injected from the emitter while the Gummel number is the integral of the profile. Then with a variable profile it is possible to maintain a small integral of the doping atoms and at the same time a large concentration of doping at the border with the emitter. It is has to be noted that a variable doping profile leads to a non zero electric field in the base. This electric field is neglected here but it will be considered later.
It is worth to mention that the low-injection condition ensures the validity of these calculations. Actually, in the short base diode model the recombination is neglected, however the condition for which the short-base diode approximation applies depends on the recombination length and then on the model used to evaluate the recombination processes. Besides these evident model simplification issues, in the high injection regime the performance of the devices are deteriorated then the low-injection condition preserve also the optimal working condition of the devices.

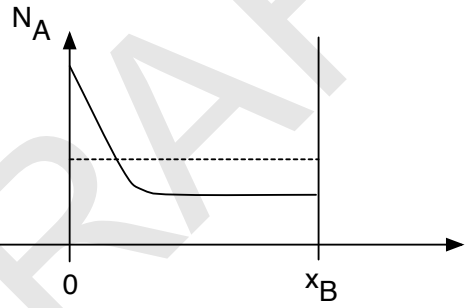


Fig. 7.8. *The same Gummel number can be obtained with a constant base doping or a variable profile with a large concentration close to the emitter where the amount of injected electrons is larger (compare with figure 5).*

A non uniform profile of acceptors in base makes possibile to maintain the low injection limit with a small Gummel number. However, a non uniform density of dopants gives rise a non uniform distribution of holes. At the equilibrium, an electric field is necessary to compensate the diffusion current due to the non zero gradient of holes. This electric field keeps at zero the holes current in base. It is interesting to note that current of holes in always negligile between the emitter and the collector, and then the same electric field that keeps at zero the holes current may act on the excess electrons to create the electron current in base.
The electric field in base necessary to satisfy the condition $J_P = 0$ can be calculated from the equilibrium between drift and diffusion current:

$$q\mu_p p\mathcal{E}_x - qD_p\frac{dp}{x} \;\rightarrow\; \mathcal{E}_x = \frac{kT}{q}\frac{1}{p}\frac{dp}{x} \tag{7.8}$$

The same electric field acts on the electrons and gives rise to the electrons current:

$$J_n = q\mu_n n\mathcal{E}_x + qD_n\frac{dn}{x} = q\mu_n n\left(\frac{kT}{q}\frac{1}{p}\frac{dp}{x}\right) + qD_n\frac{dn}{x} = qD_n\frac{n}{p}\frac{dp}{dx} + qD_n\frac{dn}{dx} \tag{7.9}$$

This can be written as:

$$J_n = q\frac{D_n}{p}\left(n\frac{dp}{dx} + p\frac{dn}{dx}\right) = q\frac{D_n}{p}\frac{d}{dx}(np) \tag{7.10}$$

The current is then calculated by an integration in base from x=0 (at the interface with the emitter) to x_B (at the interface with the collector).

$$J_n\int_0^{x_B}\frac{p}{qD_n}dx = \int_0^{x_B} d(np) = n(x_B)p(x_B) - n(0)p(0) \tag{7.11}$$

the product of the concentration of holes and electrons at the edge of the depletion layer has been calculated in chapter 5 (see eq. 5.59 and 5.60): $np = n_i^2 exp(qV_A/kT)$. Replacing the voltage with V_{BE} and V_{BC}, the final expression for the current is calculated:

$$J_n = \frac{qn_i^2}{\int_0^{x_B}\frac{p}{D_n}dx}\left[exp\left(\frac{qV_{BC}}{kT}\right) - exp\left(\frac{qV_{BE}}{kT}\right)\right] \tag{7.12}$$

As shown in section 2.5, the profile of holes at the equilibrium is equivalent to the profile of acceptors (N_A). The uneven concentration of acceptors makes variable the diffusion constant which is replaced by an equivalent term $\tilde{D_n}$. Then the integral at the denominator is the total mobile charge in the base at the equilibrium (Q_{Bo}) and the current can be written as:

$$J_n = \frac{qn_i^2\tilde{D_n}}{Q_{Bo}}\left[exp\left(\frac{qV_{BC}}{kT}\right) - \left(\frac{qV_{BC}}{kT}\right)\right] \tag{7.13}$$

In case of active bias $V_{BE} > 0$ and $V_{BC} < 0$ and the equation 7.13 is equivalent to equation 7.3 where the total charge in base replaces the product $N_A x_B$. The two equations were calculated in one case in the hypothesis of $X_B \ll L_n$, namely a short base region, and in the other case in the hypothesis of $J_p = 0$. Obviously in a short base the current in active zone is totally due to the excess electrons being the recombination current negligible and then also the current of holes is negligible.
Equation 7.13 is sometimes called the equation of the transistor because it establishes the current with any value of V_{BE} and $V_{BC}0$.

7.3 Current gain

In the previous section it has been stated that the current due to holes is always negligible. This is particularly true when the device is biased in active zone. In this condition the base-collector

junction is reversely biased, and the minority charges of the base are dragged towards the collector terminal. However, since the base-emitter junction is forwardly biased, the electrons, majority charges in the emitter, are injected in the base. then the concentration of the minority charges in the base exceeds of several orders of magnitude the concentration of minority charges in the collector. Eventually, the forward current of the BE junction is injected into the BC junction. It has been calculated in the previous section that the current collected at the collector depends on the bias applied to the BE junction.
The current at the collector is numerically dominated by electrons , however, holes exists in the device and an important property of the semiconductor devices is that all the quantities in the structure are in equilibrium one each other. This means that even if the holes are numerically negligible, their current is always proportional to the electrons current.
The device is built to take advantage of this relationship, and then a ohmic contact is applied to the base not only to bias the BE and BC junctions but also to inject and extract currents. In active bias the base, with respect to the BE junction, is endowed with two terminals: the ohmic contact and the BC reversely biased junction, and the two contacts separate the charge carriers in the base. As previously discussed the BC junction collects electrons and then the ohmic contact collects the holes.
In the device it is possible to identify at least five different holes currents, as illustrated in figure 8.

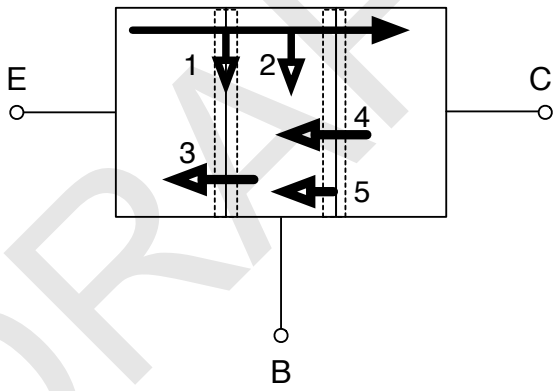


Fig. 7.9. *The five holes current sources in the BJT in active zone: 1 recombination current in the BE junction; 2 recombination current in the base; 3 forward holes current injected in the emitter; 4 reverse diffusion current from the collector; 5 reverse generation-recombination current in the BC junction.*

The components 4 and 5 in figure 9 are reverse currents while the others are direct currents. Then the holes currents contributing to the reverse current of the BC junction are always negligible with respect to the direct current contributions. The recombination current in the BE junction is given by eq. 5.65. It depends from $exp(\frac{qV_{BE}}{2KT})$ while the diffusion current depends on $exp(\frac{qV_{BE}}{KT})$. The ratio between the two exponentials is negligible when V_{BE} is greater than few hundreds of millivolts. For instance the ratio between $exp(\frac{qV_{BE}}{KT})$ and $exp(\frac{qV_{BE}}{2KT})$ is about 50 for $V_{BE} \approx 200mV$. Eventually the current of holes collected at the base terminal is made by the recombination of electrons in base and the direct injection of holes in the emitter. Since the current of holes is in

equilibrium with the current of electrons the ratio between the current collected at the collector and the current injected in the base is in equilibrium too. Then a slight increase of the base current results in an increase of the collector current. The ratio between these two quantities defines the current gain.

7.3.1 Recombination current in base

The recombination in the neutral zones has been studied in chapter 4. The same equations are still valid in the base of the BJT. In chapter 4 we observed that the magnitude of the recombination phenomena in the base depends on the length of the base region, or in better words, on the base length with respect to the recombination length. Then, shortest is the base smaller is the recombination current.
The calculation of the recombination current in base requires the exact solution of the continuity equation (eq. 5.42). Here, we calculate the recombination current in two steps. In the first the profile of excess electrons is calculated as in the short-base diode linearizing the eq. 5.42 and then the contribution to the current of the recombination is calculated applying the generation-recombination function to this excess of charge.
To calculate the holes current let us consider the continuity equation for the stationary condition

$$\frac{\partial p}{\partial t} = 0 = -\frac{1}{p}\frac{\partial J_p}{\partial x} - U \tag{7.14}$$

And the current is the integral of the recombination function

$$J_p = -q\int_0^{x_B} U dx \tag{7.15}$$

Under the hypothesis of low-injection and using the approximations detailed in chapter 4, the recombination function is the ratio between the excess charges and the recombination time. In order to describe the actual current collected at the terminals it is necessary to introduce the area through which the current flows, here the relevant area is the area of the emitter contact through which the direct current, producing the charge excess, originates.
Then the current at the base terminal is:

$$I_{rb} = -J_p A_E = qA_E \int_0^{x_B} \frac{n'(x)}{\tau_n} dx \tag{7.16}$$

Replacing the formula of the charge excess in a short diode we have:

$$I_{rb} = \frac{qA_E}{\tau_n}\frac{n_i^2}{N_A}\frac{x_B}{2}\left(exp\frac{qV_{be}}{kT} - 1\right) \tag{7.17}$$

The recombination current describes the loss of electrons traveling from the emitter to the collector. The loss of electrons is equivalent to the holes injected from the base contact. This is measured by the transport factor in base (α_T) that is defined as the ratio between the current leaving the emitter and the current at the base contact.

$$\alpha_T = \frac{|I_{ne} - I_{rb}|}{|I_{ne}|} = 1 - \left|\frac{I_{rb}}{I_{ne}}\right| \tag{7.18}$$

The current from the emitter is the short diode forward current:

$$I_{ne} = \frac{qA_E}{\tau_n}\frac{n_i^2}{N_A}\frac{D_n}{x_B}\left(exp\frac{qV_{be}}{kT} - 1\right) \tag{7.19}$$

The transport factor is easily calculated replacing the eq. 7.11 and 7.13 in eq. 7.12:

$$\alpha_T = 1 - \frac{x_B^2}{2L_n^2} \tag{7.20}$$

Where the definition of $L_n = \sqrt{D_n\tau_n}$ has been used.
This result is quite predictable because the loss of electrons in base, and then the holes current due to recombination, depends on the ratio between the physical length of the base and the recombination length.
Thus, the transport factor depends on the recombination length and base region size. The recombination length is the square root of the product of the diffusion constant and the recombination time, then it depends both on the concentrations of the dopant atoms and the recombination centers. On the other hand, x_b is of course limited by the technological capabilities of the device fabrication.
It is interesting to note that $\frac{x_B^2}{2L_n^2}$ corresponds to the ratio between the transit time in base (see eq. 5.81) and the recombination time. Indeed, an electron injected in the base is either recombined inside the base by a hole or it reaches the interface with collector; thus, the ratio between the typical time constants of these two events is a measure of the efficiency of the transfer of electrons across the base.
It is interesting to note that in a diode it is impossible to discriminate, observing the current, between short or long bases. But in a BJT since the contacts of the base are selective for the charge carriers, the two currents (electrons and holes) are separated in distinct networks.
Quantitatively, α_T is very close to one. In a typical case where $L_n = 10\mu m$ and $x_B = 0.3\mu m$, the transport factor is $\alpha_T = 0.9996.$, thus $I_{ne} \approx 2500 I_{rb}$. This means that one holes injected in the base controls the transfer of 2500 electrons from the emitter to the base. This is due to the fact that the current of holes is a recombination current while current of electrons current is a diffusion gradient. In the short space of the base the subtraction, by recombination, of few electrons may give rise to a large change of the gradient of the density of electrons and then to a large increase of the current of electrons. In a BJT made with a long base, also the current of electrons is a recombination current and each hole injected in the base recombines one electron and an additional electron, of the same amount, is required from the emitter. In other words, due to the complete recombination, the current injected from the emitter is totally converted in the holes current collected at the base terminal, namely $|I_{ne}| = |I_{rb}|$ and $\alpha_T = 0$. Such a device in active zone would merely behave as a diode between the base and the emitter contacts with the additional contact of collector where only a negligible reverse current could be collected.

7.3.2 Forward holes current in the emitter

In active zone, the BE junction is a forward biased PN junction, then the flow of electrons injected from the emitter to the base is complemented by the flow of holes from the base to the emitter. The holes current obviously originates from the base contact, then it is an additional contribution to the recombination current calculated in the previous section.

The current of holes can be calculated from the theory of the PN junction, then the equations 5.47 and 5.52 hold if the emitter is long or short respectively. As usual, long or short is related to the recombination length of the holes in the emitter neutral region. Typically, the transistor dimensions are such that also the emitter is a short region.
The relationship between the currents of electrons and holes is described by the emitter efficiency γ that is defined as the ratio between the electrons current and the total current at the emitter contact.

$$\gamma = \frac{|I_{ne}|}{I_E} = \frac{|I_{ne}|}{|I_{ne} + I_{pe}|} = \frac{1}{1 + \left|\frac{I_{pe}}{I_{ne}}\right|} \tag{7.21}$$

In the case of a short emitter region, the current of holes is:

$$I_{pe} = qA_E D_{pe} \frac{n_i^2}{N_D x_E} \left[exp\left(\frac{qV_A}{kT}\right) - 1 \right] \tag{7.22}$$

the current of electrons has been previously calculated (eq. 7.13) and then from the two currents the emitter efficiency is:

$$\gamma = \frac{1}{1 + \frac{x_B N_A D_{pe}}{x_E N_D D_{nb}}} \tag{7.23}$$

Thus, the emitter efficiency depends on the doping of the base with respect to the doping of the emitter. The previous equation has been calculated for homogeneous doping of the emitter and the base. The same equation can be extended to the case of variable doping profiles using the concept of Gummel number. In case of a uniform doping, the Gummel number of emitter and base are: $GN_B = x_B \cdot N_A$; and $GN_E = x_E \cdot N_D$. Then the emitter efficiency takes the following general form:

$$\gamma = \frac{1}{1 + \frac{GN_B D_{pe}}{GN_E D_{nb}}} \tag{7.24}$$

In order to maximize the emitter efficiency it is necessary to make the Gummel number of the emitter larger than the Gummel number of the base. Since both the regions are short, this condition is met when the doping of the emitter is much larger than the doping of the base. It is worth to mention that, due to the difference of doping concentrations, the diffusion constant in the base is larger than the diffusion constant of the emitter.

7.3.3 Numerical comparison of α_T and γ

The transport factor and the emitter efficiency are two figures of merit describing the ratio between the current due to the electrons and the current due to the holes. Since electrons are collected at the collector and the holes are collected at the base, these quantities actually describe the ratio between the currents at the contacts of the device.
The transport factor and the emitter efficiency are numerically slightly different. This can be appreciated in a practical case. Let us consider a BJT made in silicon with $N_{De} = 10^{17}\ cm^{-3}$ and $N_{Ab} = 10^{15}\ cm^{-3}$, namely the doping in base is 100 times smaller than the doping in the emitter. In these condition at room temperature the recombination lengths are: $L_{pe} = 9\ \mu m$ and $L_{nb} = 14\ \mu m$,

and the diffusion coefficients are $D_{pe} = 8.25\ cm^2 s^{-1}$ and $D_{nb} = 18.90\ cm^2 s^{-1}$. In order to fulfil the short neutral zone condition, the neutral zones are chose about 30 times smaller than the respective recombination lengths, then: $x_E = 0.30\ \mu m$ and $x_B = 0.45\ \mu m$.
With these values the transport factor is $\alpha_T = 0.9997$.
The Gummel numbers of the emitter and the base are: $GN_e = x_E N_{De} = 3.02 \cdot 10^{12}\ cm^{-2}$ and $GN_b = x_B N_{Ab} = 4.58 \cdot 10^{10}\ cm^{-2}$. Then the emitter efficiency is $\gamma = 0.9934$.
Eventually, both the figures of merit are very close to one, but γ is less close to one being its third decimal figure different from zero.

7.3.4 Total current gain

In order to calculate the total current gain, let us consider Figure 10 where the relationship between the external and the internal currents is shown.
According to the previous definitions, the current of electrons from the emitter is γ times the emitter current: $I_{ne} = \gamma I_E$ and the emitter current is the sum of the electrons and holes currents: $I_E = I_{ne} + I_{pe}$.
On the other hand the current collected at the collector is α_T times the current of electrons injected in the base (note the minus sign due to the conventional verse of currents) $I_C = -\alpha_T I_E$.
Then, combining the previous definitions we have:

$$I_C = -\alpha_T I_{ne} = -\alpha_T \gamma I_E = -\alpha_F I_E \tag{7.25}$$

where $-\alpha_F = -\alpha_T \gamma$ describes the current loss between the emitter and the collector contacts.
The current gain defined as the ratio between the collector and base currents is obtained considering the current node in the BJT: $I_C + I_E + I_B = 0$, from which we can write:

$$I_C - \frac{I_C}{\alpha_F} + I_B = 0 \tag{7.26}$$

from which:

$$I_C = \frac{\alpha_F}{1 - \alpha_F} I_B = \beta_F I_B \tag{7.27}$$

where $\beta_F = \frac{\alpha_F}{1-\alpha_F}$ is the d.c. amplification of the BJT. Numerically, considering the above calculated numbers $\alpha_T = 0.9997$ and $\gamma = 0.9934$, we have $\alpha_F = 0.9931$ and $\beta_F = 144$.
Both α_T and γ contribute to β_F even if with different magnitudes. In order to evaluate the contribution of each term let us calculate β_F for each factor keeping at 1 the other, namely considering the other as ideal. This gives the following results:
$1 \cdot \alpha_T \rightarrow \beta_F = 3460$
$\gamma \cdot 1 \rightarrow \beta_F = 151$
$\gamma \cdot \alpha_T \rightarrow \beta_F = 144$
Clearly, the d.c. amplification factor is mostly based on the emitter efficiency, then it is the difference of doping that builds the amplification factor.
The reasons for the current amplification are found in the equilibrium between J_n and J_p and in the fact that different doping and narrow base can greatly change the ratio of the current, but they are still connected one each other. The relationship between currents balance and doping are yet present in the PN junction, the great advance in BJT is the double contact of the base that makes possible the separation of electrons from the holes.

7.4 BJT operative conditions

The behavior of a BJT depends on the voltages applied to the two junctions. According to the sign of the applied voltages there are four different operative conditions that are usually called: cut-off, saturation, active zone, and inverse active zone. Until now we have not found any reason to make the collector different from the emitter, so in principle the active zone can be obtained either with the junction BE forward and BC reverse or vice-versa. Later, we will see that in order to ensure a better behavior it is necessary to make the doping of the collector different from that of the emitter.

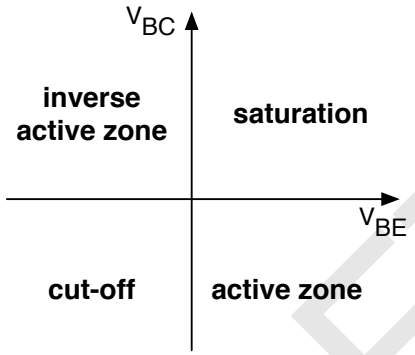


Fig. 7.10. *The four different operative conditions of the BJT are defined by the voltage applied to the junctions.*

In the cut-off condition, both the junctions are reverse biased, then the base is depleted of electrons and any current flowing in the device is small.
In saturation condition both the junctions are forward biased. In this state, there is an injection of electrons into the base from both sides. Since the base is narrow, the electrons concentration profile is linear and the total current is proportional to the slope of the profile. The situation is depicted in figure 10. If the applied voltage is exactly the same then the total current is zero, and the current reaches its maximum value when one of the two voltages (V_{BC} in particular) is null or negative (active zone). In general, the current is given by the differences of the current injected by each junction

$$J_n = J_0 exp\left(\frac{qV_{be}}{kT} - \frac{qV_{bc}}{kT}\right) \tag{7.28}$$

In active zone, the current still depends on the voltages applied to the two junctions but also on the current injected in the base contact. To describe the device let us consider the so-called common emitter configuration where the device is described by two currents (I_C and I_B) and two voltages (V_{BE} and V_{CE}). V_{CE}=V_{BE}-V_{BC} is the difference between the voltages applied to the two junctions.

Each quantity is a function of the other threes, such as $I_C = f(I_B, V_{BE}, V_{CE})$. Instead of considering a multidimensional function is more useful to split the representation in two separate characteristics that are called the input characteristic ($I_B = f(V_{BE})$) and the output characteristic

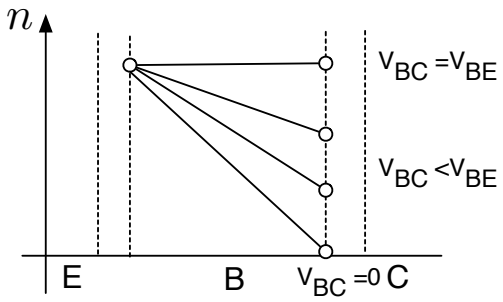


Fig. 7.11. *The current in saturation depends on the relative magnitude of V_{BC} with respect toV_{BE} .*

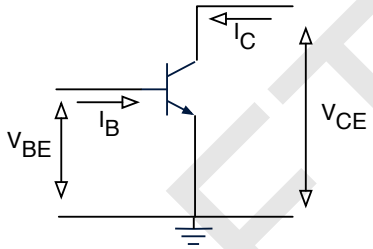


Fig. 7.12. *Input and output quantities of the common emitter configuration.*

($I_C = f(V_{CE})$) with reference to the two separated input and output networks of the common emitter configuration.
The input characteristic has the analytical form of a forward biased diode but since the current flowing from the base to the emitter is only made by holes , the magnitude of this current is at least two orders of magnitudes smaller with respect to the normal forward current of PN junction. For the output characteristics, when $V_{CE} = 0$ then $V_{BE} = V_{BC}$, the quantities of electrons injected from the two sides are the same and the current is zero regardless the value of V_{BE}. At $V_{CE} > 0$ we have $V_{BE} > V_{BC}$ then the current grows as V_{CE} increases. The growth stops when $V_{CE} = V_{BE}$ namely when $V_{BC} = 0$. Beyond this value the BJT works in the active zone, the collector current remains constant, an any further increase of V_{CE} makes larger the reverse bias of the base-collector junction.
The value at which the active zone is reached depends on the base current (I_B). Indeed, considering $qV_{BE} \gg kT$ the following condition holds:

$$I_C = \beta_F I_B = \beta_F I_{B0} exp\left(\frac{qV_{be}}{kT}\right) \tag{7.29}$$

The functional behavior of the output characteristics is described by the reverse bias of the base-collector junction with the very important difference that the current is the current of electrons injected by the emitter region (and controlled by the base current) and then it is quantitatively large. Large values of V_{CE} can prompt the phenomena of junction breakdown that have been described

in section 4. Actually, since the base is poorly doped, the breakdown in BJT is dominated by the avalanche effect.
It is worth to remind that input and output quantities are defined by the circuital configurations, for instance in the common base configuration, the base is the common terminal and the quantities of the input network are I_E and V_{BE} and the quantities of the output network are I_C and V_{CB}, and obviously in such a configuration the current gain is little less than one.

7.5 Non ideal behaviors

The BJT model, developed so far, is a plain derivation from the ideal current of the PN junction. The main characteristics of the derived model is the independence of the device properties (in particular the current gain) from the operative conditions. Of course there is a number of minor effects that has to be considered to improve the description of the device.

7.5.1 Early effect

The simple model of BJT predicts that in active zone the device behaves as a ideal current source. Namely, the current is independent from the voltage and the output characteristics is flat.
It is known that ideal current source violates the fundamental principles of electric networks; in particular, the fact that different BJTs connected in series are forced to provide the same current even if each device is biased with a different V_{BE}.
As a consequence, I_C has to show some dependence on V_{CE}. The simplest evidence of this dependency is offered by the Early effect. This is a consequence of the reverse bias of the BC junction. In active zone, since V_{BE} is locked at the forward bias value (typically less than 0.75 V for silicon devices), any positive increase of V_{CE} results in an increase of the absolute value of V_{BC}, and then in a more deep reverse bias. Under reverse bias the space charge region becomes wider and the position along the coordinate x where the excess injected charge is zero moves towards the emitter. Namely, the slope of the concentration of excess electrons in base increases and the current increases.
The Early effect consists in a slight increase of the current with V_{CE}, making the output characteristics in active zone far to be flat.
The magnitude of the Early effect is described by a voltage (V_A) called Early voltage. It is defined by the direct calculation of the slope of the output characteristics. The slope is the derivative of I_C with respect to V_{CE} that corresponds to the derivative with respect to V_{BC}. Since the current changes because of the variation of x_B, it is convenient to write:

$$\frac{\partial I_C}{\partial V_{CE}} = \frac{\partial I_C}{\partial V_{BC}} = \frac{\partial I_C}{\partial x_B}\frac{\partial x_B}{\partial V_{BC}} \tag{7.30}$$

Replacing the collector current (eq. 7.23) and differentiating with respect to x_B we have:

$$\frac{\partial I_C}{\partial V_{BC}} = -\frac{I_C}{x_B}\frac{\partial x_B}{\partial V_{BC}} \tag{7.31}$$

And finally the Early tension can be defined as:

$$V_a = -\frac{x_B}{\frac{\partial x_B}{\partial V_{BC}}} \tag{7.32}$$

Graphically, the Early voltage on the VC_{CE} axis of the output characteristics corresponds to the convergence point of the $I_C = f(V_{CE})$ curves.
The base region is narrow, and the space charge region of the BC junction extends into the base. In extreme cases, the space charge region can occupy all the neutral zone in the base touching the base-emitter junction. This condition is said punch-through and corresponds to a shunt of the base. In order to mitigate the Early effect it is necessary to limit the expansion of the space charge region, under reverse bias, towards the base. For this scope, the collector region needs to be less doped than the base. This condition breaks the symmetry in the device and makes the inverse active zone different than the active zone.
The base is poorly doped in order to maintain large the current gain, and the collector is still less doped in order to mitigate the Early effect. Clearly, the doping of the collector cannot be too small in order to guarantee that the low-injection condition holds.

7.5.2 Emitter band-gap narrowing

The emitter efficiency depends on the ratio between the doping of the emitter and the base, thus it is straightforward to deduce that increasing the doping of the emitter leads to an increased ideality of the emitter efficiency. However, the increase of the emitter efficiency is limited by an additional phenomenon called *band-gap narrowing*. As the doping concentration increases the dopant atoms becomes to feel the presence of each other, then due to the Pauli principle, the donor level splits into a band. At a doping concentration larger than $10^{19}\ cm^{-3}$ the donor levels begin to merge with the conduction band and the whole band gap of the semiconductor is narrowed. Eventually, the intrinsic concentration (n_i) in the emitter becomes larger, and the holes current injected in the emitter increases and the emitter efficiency is reduced ($n_i = N_c N_v vexp(-\Delta E_g/kT)$.
The emitter efficiency (eq, 7.23) can be rewritten including the intrinsic concentration of the base (n_{ib}) and the emitter (n_{ie}) as:

$$\gamma = \frac{1}{1 + \frac{n_{ie}^2 x_B N_A D_{pe}}{n_{ib}^2 x_E N_D D_{nb}}} \tag{7.33}$$

Figure 13 shows an example of the relationships between the doping concentration and the intrinsic concentration and the emitter efficiency. Clearly, the increase of doping beyond $10^{19}\ cm^{-3}$ results in a deterioration of the emitter efficiency.

7.5.3 Small base current

The voltage drop across the base-emitter junction is

$$V_{BE} = \frac{kT}{q} ln \frac{I_b}{I_{b0}} \tag{7.34}$$

Then it is straightforward that if I_B is small also V_{BE} is small. As discussed in section 5.3.4 at small direct bias voltages the recombination current in the depletion layer is not negligible. The recombination current, labeled as 1 in figure 7.9, has been ignored so far because we supposed the base-emitter junction fully directly biased. Actually when the exponential term of the ideal current

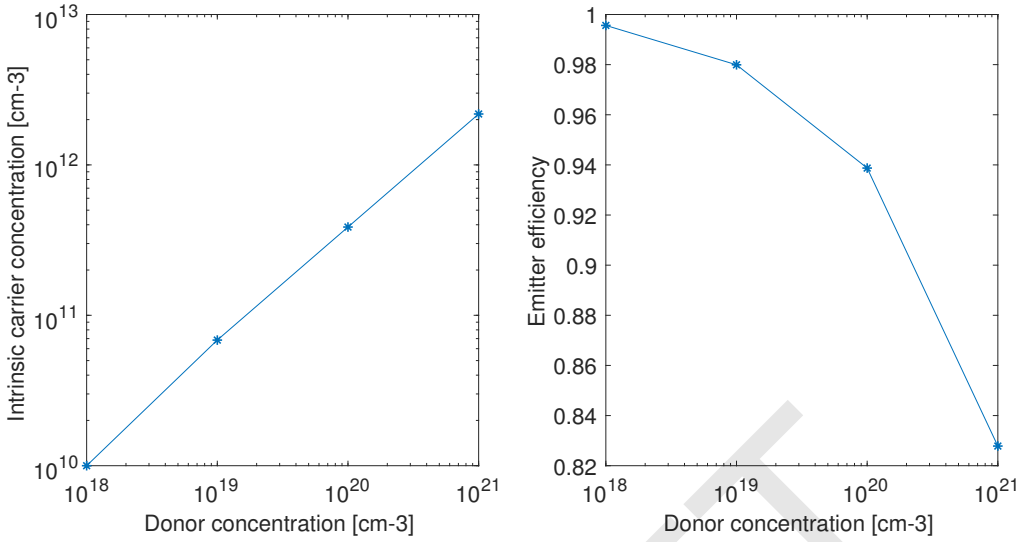


Fig. 7.13. *Emitter band-gap narrowing effect. Left: intrinsic concentration vs. doping, Right: emitter efficiency vs. doping. The plots are a direct application of equation 7.33 with* $N_A = 10^{16}\ cm^{-3}$. *The band-gap narrowing is negligible when* $N_D = 10^{18}\ cm^{-3}$ *and it is 50 meV, 95 meV, and 150 meV at* N_D *equal to* 10^{19} , 10^{20} *and* $10^{21}\ cm^{-3}$ *respectively. Data on band-gap narrowing from C. Hu, Modern Semiconductor devices for integrated circuits, J. Wiley.*

is comparable with the exponential term of the recombination current; namely when $exp(\frac{qV_{BE}}{kT}) \approx exp(\frac{qV_{BE}}{2kT})$. In this condition, an additional quantity of electrons is lost travelling from the emitter to the collector. It corresponds to a decrease of α_F, since β_F is very sensitive to small variations of α_F the recombination current in the depletion layer of the base-emitter junction results in a sensitive reduction of the current gain factor.

7.5.4 High injection effects

High injection jeopardizes the validity of the model developed so far, in particular the generation recombination function cannot be approximated with eq. 4.27. Furthermore, high injection is also detrimental to the performance of the device. The effects of high injections can be observed at both base-emitter and base-collector junctions.

High injection effects at the base-emitter junction
The high injection at the base-emitter junction results in a decrease of the relationship between V_{be} and the injected electrons $n'(0)$. This can be easily understood considering the product between electrons and holes at the interface between the base and the depletion layer of the base-emitter junction. This quantity has been calculated in chapter 5 (eq. 5.60). In high injection $n'(0) \approx N_A$, and then $n \approx p$. The concentration of injected electrons can be directly calculated from eq. 5.60

$$n'(0) \cdot p = n_i^2 exp(\frac{qV_{be}}{kT}) \rightarrow n'(0) = \sqrt{n'(0) \cdot p} = n_i exp(\frac{qV_{be}}{2kT}) \tag{7.35}$$

In practice, the efficiency of V_{be} to control the electrons current is reduced.

High injection effects at the base-collector junction
The consequence of the high injection at the base-collector junction is complex to be described and it is usually called Kirk effect. Without going into the details it is possible to explain the decrease of performance with a simple consideration about the charges in the depletion layer of the BC region. When the collector current increases, a large amount of electrons transits through the depletion layer of the BC junction. The doping concentrations of the base and the collectors are typically not very large; the first to get the current gain and the second to restrain the Early effect. Then, the concentration of electrons may become comparable with the concentration of the dopant atoms . This affects the total charge in the depletion layer. In the side of the base: $Q = -N_A - n$ and in the collector side: $Q = +N_D - n$. Then, in one case the total charge increases and in the other case it decreases.

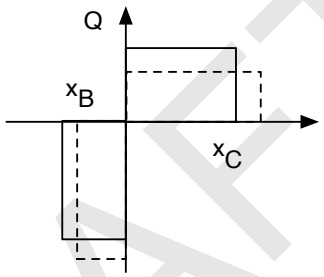


Fig. 7.14. *Change of charge profile in the base-collector depletion of layer. solid line is the equilibrium condition and dashed line is the situation in the high injection. Note that the collector is less doped than the base to mitigate the Early effect.*

The charge in the depletion layer affects the electric field. In particular the electric field in the base-side of the depletion layer increases and the electric field in the collector-side decreases:

$$\mathcal{E}_B = \frac{q(-N_A - n)}{\epsilon_S}(x + x_B) \tag{7.36}$$

$$\mathcal{E}_C = \frac{q(+N_D - n)}{\epsilon_S}(x - x_C) \tag{7.37}$$

The voltage drop across the junction is due to the applied V_{CE} and in presence of a d.c. bias is practically constant ($V_{BC} = V_{CE} - V_{BE}$). The relationship between the voltage drop and the electric field is:

$$V_{BC} = -\int_{x_B}^{x_C} \mathcal{E} dx \tag{7.38}$$

Then, since the electric field changes, in order to keep constant the voltage drop it is necessary that the dimensions of the depletion layer change. Due to the different sign of the charges variation,

the depletion layer in the base region shrinks and in the collector region expands. As in the Early effect, this has a consequence on the current: indeed the neutral zone in the base region increases and then the current decreases. Thus the Kirk effect has an opposite behaviour with respect to the Early effect.
Again, at large base currents the total current gain decreases, the qualitative behavior of β_f vs. V_{bE} is shown in figure 13. Reassuming, Figure 15 shows a qualitative behaviour of the output characteristics of a generic BJT.

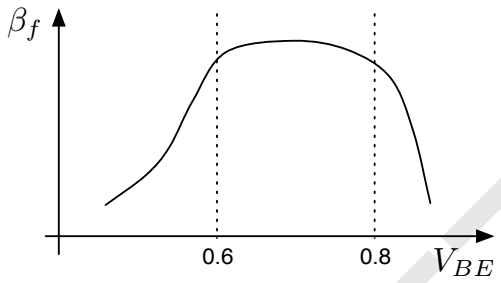


Fig. 7.15. *Low injection and high injection both results in a degradation of the current gain.*

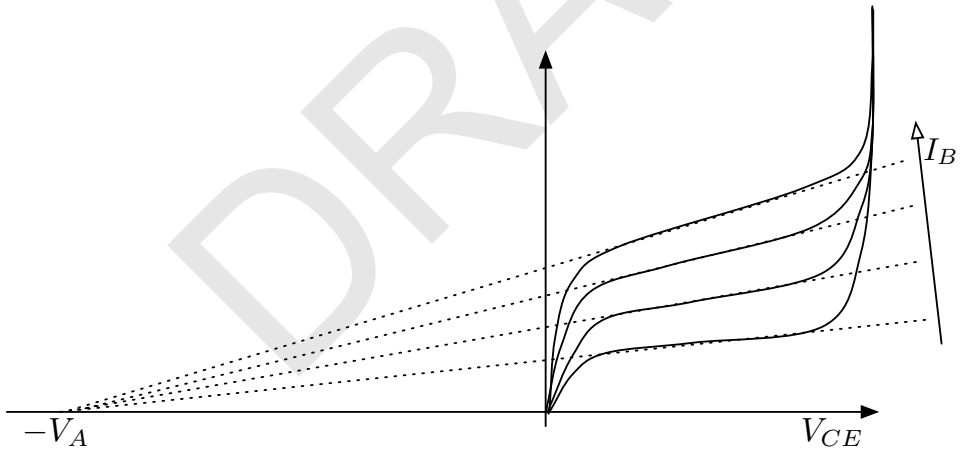


Fig. 7.16. *Typical behaviour of the output characteristics of a BJT. dotted lines indicate the Early effect characterised by the Early voltage ($-V_A$). Breakdown effects at the base-collector junction appears at large V_{CE}.*

7.6 Physical effects in real BJT

Figure 15 shows a practical implementation of the BJT in planar technology. In practice, the design is even more complex, but this scheme gives an idea of the geometrical arrangement of the device. The transistor is grown on a p-type substrate and it is separated from the rest of the wafer by deep oxide trenches. The core of the BJT is the n+ doped emitter, the narrow p-type base and the n doped collector. In order to separate the base and collector contacts, the collector contact is made behind a shallow oxide trench that splits the collector in two regions connected by a thin n+ layer. This layer further insulate the device creating a deep depletion in the substrate and then limiting the current leakage through the substrate itself.
The base contact is annular around the emitter contact. The emitter area (A_E) is defined by the area below the emitter contact.

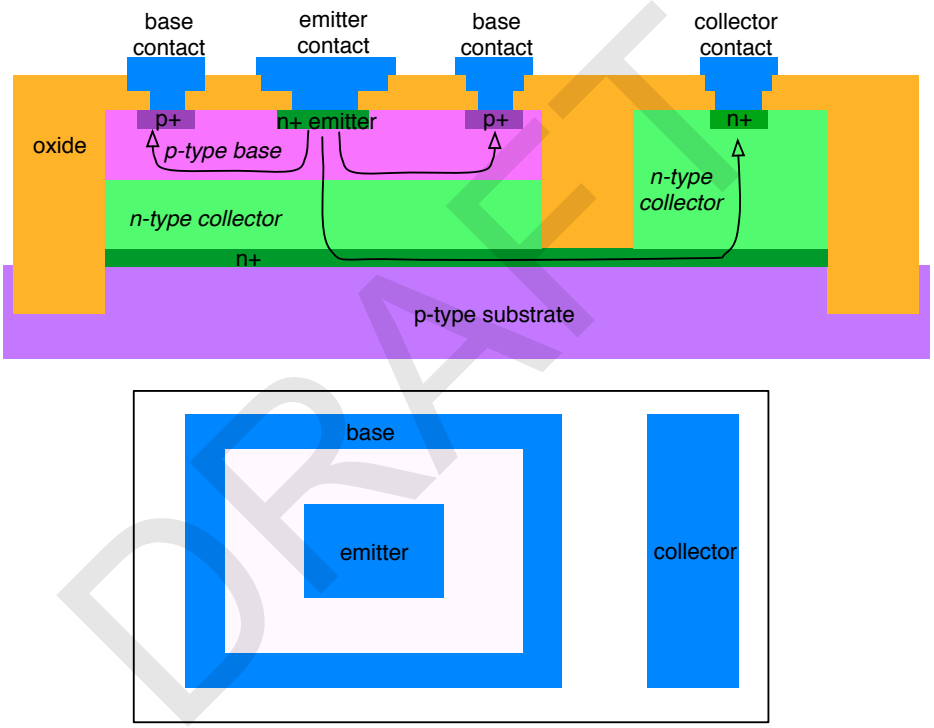


Fig. 7.17. *Schematic arrangement of an integrated npn BJT on a p-type substrate. The path of base and collector currents are indicated by arrows. The upper view show the concentric emitter and base contacts and the lateral collector contact. The figure shows a principle implementation and drawing is not in scale.*

The particular configuration of the base contact gives rise to the so-called base current-crowding effect. The base is poorly doped then its resistivity is high. As a consequence, the current from the base to the emitter, tends to follow the shortest path accumulating at the border of the emitter region. Then, even if the total base current is modest, the density of current at the edges of the

emitter tends to become large, the high injection regime in this region can be activated and a asymmetric heating may happen. The current crowding can be avoided reducing the distance between the emitter and the base contact, but this solution in the configuration of Fig. 7.16 requires a decrease of the emitter area and then of the current.
An advantageous solution to maintain a large emitter area and a short inter-contacts distance is offered by the interdigitated contacts as shown in Fig. 16. This geometry allows a reduction of the distance between the base contact and the base-emitter junction maintaining the same emitter area. In practice it corresponds to a parallel connection of many base-emitter junctions.
Finally, it has to be mentioned that the distance between contacts also affects the actual V_{BE} across the junction which is given by: $V'_{BE} = V_{BE} - R_B I_B$. Where R_B is the total base resistance.

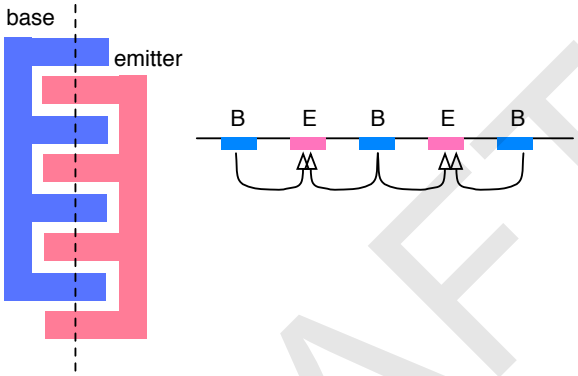


Fig. 7.18. *Interdigitated base-emitter contacts. A large emitter area can be obtained reducing the distance between the contacts, and then the base resistance. It corresponds to a number of parallel base-emitter junctions.*

7.7 Dynamic response

The dynamic response of the BJT depends on three major factors: The capacitance of the base-collector junction (C_{jbc}), the diffusion capacitance of the base-emitter junction (C_{dbe}), and the base resistance. C_{jbc} is the junction capacitance associated to the depletion layer of the reverse biased base-collector junction, while C_{jbe} is the diffusion capacitance due to the accumulation of minority charges in the forward biased base-emitter junction.
An equivalent circuit of the BJT for the common emitter configuration is shown in figure 16. The resistance of the base is split in two parts related to the effective resistance of the base-emitter junction (R_{BE}), and an additional resistance due to the distance between the base contact and the base-emitter junction (R'_{BB}). The diffusion capacitance C_{dbe} is in parallel to (R_{BE}) while C_{jbc} connects the base and the collector contact.

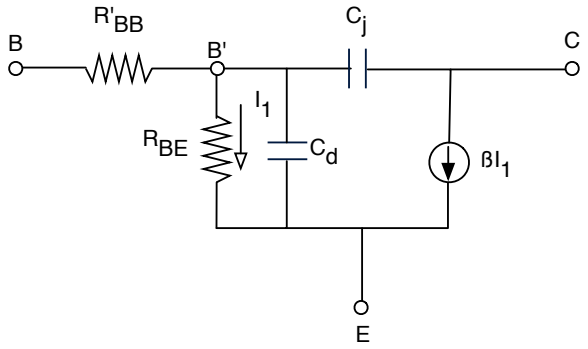


Fig. 7.19. *BJT equivalent circuit. B indicates the physical contact of the base, while B' is the internal contact of the base-emitter junction.*

7.7.1 Junction and diffusion capacitances

The junction capacitance between the base and the collector is simply the depletion layer capacitance previously calculated both in Schottky and PN diodes:

$$C_j = \frac{\epsilon_s}{x_d} \tag{7.39}$$

where x_d is the depth of the depletion layer $x_d = \sqrt{\frac{2\epsilon_s \phi_i}{q}(\frac{1}{N_A} + \frac{1}{N_D})}$.

The diffusion capacitance of a forward bias PN junction has been calculated in chapter 5. Here we recall that the diffusion capacitance is the derivative of the total charge of the injected minority charges (electrons and holes in the respective parts of the junction) with respect to the junction voltage. The capacitance in the p-type base is due to the injected electrons.

$$C_d = \frac{dQ_{nb}}{dV_{BE}} \tag{7.40}$$

The relationship between Q_{nb} and the diffusion current generated by this charge is

$$dQ_{nb} = \frac{x_B^2}{2D_n} J_n = \tau_{tr} J_n \tag{7.41}$$

where τ_{tr} is the time of transit of the charges in the base region. Eventually, the diffusion capacitance in the base is:

$$C_d = \tau_{tr} \frac{dJ_n}{dV_{BE}} \tag{7.42}$$

Then it depends on the transit time and on the quiescent point of the device.
The transit time can be made small if the base region (x_B) is small and the diffusion constant is large, this last condition means the base is poorly doped.
To provide an example of transit time, let us consider a p-type base where $N_A = 10^{15} cm^{-3}$, $L_{nb} = 14\mu m$ and $D_{nb} = 18.9 cm^2/s$. For a base region depth $x_B = 0.45\mu m$ the transit time is about

50 ps.
The transit time can be further reduced if the transport of electrons in base is facilitated by an electric field. The presence of an electric field provides indeed an additional drift velocity to the diffusing charges ($v = \mu\mathcal{E}$).
The transit time due to drift is:

$$\tau_{tr}^{drift} = \frac{x_B}{v} = \frac{x_B}{\mu_n \mathcal{E}} \tag{7.43}$$

The electric field is proportional to the voltage drop across the base region: $\mathcal{E} = \Delta V / x_B$, then

$$\tau_{tr}^{drift} = \frac{x_B^2}{\mu_m \Delta V} \tag{7.44}$$

The application of an electric field is of benefit if the transit time due to drift exceeds the transit time due to diffusion. Considering the ratio between the two transit times a condition for the voltage drop across the base necessary to improve the transit time can be obtained.

$$\frac{\tau_{tr}^{drift}}{\tau_{tr}^{diff}} = \frac{x_B^2}{\mu_m \Delta V} \cdot \frac{2D_n}{x_B^2} = \frac{kT}{q\Delta V} \tag{7.45}$$

Where the Einstein relationship has been used to replace the diffusion constant with the mobility. Then, the transit time is dominated by drift if the voltage drop is at least twice the thermal voltage. At room temperature we have:

$$\Delta V \geq 2\frac{kT}{q} \simeq 50mV \tag{7.46}$$

This voltage drop can be produced by a graded doping profile of the base. Figure 18 shows that in case of a graded doping, the equilibrium (Fermi level constant) is maintained by a voltage drop across the material. Let N_{AE} and N_{AC} be the doping concentration at the borders of the emission and collector regions. The voltage drop at the equilibrium is given by:

$$\Delta V = \frac{kT}{q} ln(\frac{N_{AE}}{N_{AC}}) \tag{7.47}$$

Then the condition $\Delta V = 2\frac{kT}{q}$ is obtained when

$$\frac{N_{AE}}{N_{AC}} = e^2 \simeq 7 \tag{7.48}$$

However, the excursion of the doping concentration has to fulfill the amplification of current, the Early effect and the low-injection condition.
The diffusion capacitance due to the holes in the emitter region is given by:

$$C_{dp} = \frac{dQ_{pe}}{dV_{be}} \tag{7.49}$$

where $Q_{pe} = \tau_{trE} J_{ne}$. Where τ_{trE} is the transit time of holes in the emitter region. The current defining the holes injected in the emitter is the current of current flowing to the base. Actually, the concentration of injected holes and electrons is rather different due to the amplification, but disrespectful of their own value their ratio is always constant; namely the equilibrium is always

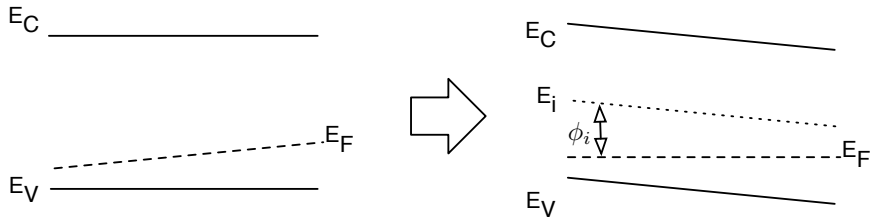


Fig. 7.20. *Band diagram of a non uniformly doped p-type region. left: non equilibrium condition, right: equilibrium condition. The equilibrium condition is maintained by a built-in voltage drop.*

maintained. Then the time necessary to accumulate the injected holes is the same time necessary to accumulate the injected electrons.
The concentration of injected holes is:

$$Q_{pe} = \int_0^{x_E} qp'(x)dx = \int_0^{x_E} \frac{n_i^2}{N_D} exp(\frac{qV_{be}}{kT})(1 - \frac{x}{x_E})dx = q\frac{n_i^2}{N_D} exp(\frac{qV_{be}}{kT})\frac{x_E}{2} \tag{7.50}$$

Reminding the relationship between electrons and holes current in the forward biased BE junction and the expression for the holes current (equations 16 and 18), the transit time of holes in the emitter can be written as:

$$\tau_{trE} = \frac{x_E^2}{2D_{nb}} \frac{G_{nb}}{G_{ne}} \tag{7.51}$$

Where G_{nb} and G_{ne} are the Gummel numbers of base and emitter respectively. The diffusion capacitance due to the injection of holes in the emitter is then given by:

$$C_{dE} = \tau_{trE} \frac{dJ_n}{dV_{BE}} = \frac{x_E^2}{2D_{nb}} \frac{G_{nb}}{G_{ne}} \frac{dJ_n}{dV_{BE}} G_{ne} \tag{7.52}$$

Then, with respect to the diffusion capacitance due to the injected electrons, the capacitance due to the injection of holes is depressed of a factor equal to the ratio of the Gummel numbers. Namely, as much as the BJT amplifies as much this capacitance is small. Since the two diffusion capacitances are in parallel, the only meaningful capacitance of the base-emitter junction is the diffusion capacitance of the electrons injected in the base.

7.8 Conclusions

The BJT properties derive from a balance between the three doping concentrations and the dimensions of the three neutral regions. Different objectives may require different and sometimes opposite conditions. In particular, the conditions leading to a large current gain are in contrast with the conditions required for a high frequency operation. In the next chapter we will see that heterostructures, namely junctions formed by materials differing not only for their doping but also for band gap and affinity, offer an optimal solution for the design of BJTs where large gain is not opposite to the large bandwidth.

Table 7.1. Doping - performance relationship

Current gain	$N_D^E > N_A^B$ and short base
Early effect counteraction	$N_A^B > N_D^C$
Kirk effect counteraction	large N_A^B and $N_D^C 4$
Small transit time in the base	small N_A^B or graded $N_A^B(x)$
Small Junction capacitance	large N_A^B and N_D^C then narrow BC junction depletion layer
Small base resistance	large N_A^B and/or short base region

8

Heterojunctions

8.1 Introduction

The cases studied in the previous chapters were related to devices made by one semiconductor even if characterized by different dopings. This means that in the band diagrams drawn so far, the affinity and the energy gaps of the semiconductors were the same and the materials differed only for their work functions that, of course, depend on dopings.
Actually, the large variety of available semiconductors may give rise to junctions between different materials. The most important semiconductors are the elements of the IV group of the periodic table (silicon and germanium among the others) and their combination such as the silicon carbide (SiC). Other important semiconductors are the compounds obtained combining elements of the groups III and V. Among them, gallium arsenide (GaAs) and indium phosphate (InP) are endowed with very interesting properties making them appealing for high frequency and optoelectronic applications. One of such features is the negative differential resistance that has been discussed in chapter 5. Even ternary mixtures may result in semiconducting materials such as aluminium gallium arsenide (AlGaAs).
The semiconducting character is also exhibited by the oxides of the transition metals. Some of these materials are gaining importance for their applications in optoelectronic and piezoelectric devices such as the zinc oxide (ZnO) and the titanium dioxide (TiO_2). More recently, semiconductor materials based on organic molecules and polymers are also emerging giving rise to the so-called organic electronics.
With all these materials, it is reasonable to conceive devices where more materials can coexist forming junctions between different semiconductors. Such junctions are called heterojunctions while all the junctions considered so far can be called homojunctions.
In the frame of the theory developed in the past chapters, a heterojunction is defined as the junction between two materials where, besides work function, either the affinity or the energy gap or even both are different. We will see later that the properties of heterojunctions are peculiar respect to homojunctions, in particular, we will see the benefits in a BJT where the base-emitter junction is formed by two different materials. These properties where theoretically understood since the beginning of the semiconducting era, so that the idea of heterojunction BJT is contained in the first patent of the bipolar transistor technology released by W. Shockley.
However, the fabrication of heterojunctions has not been possible until the beginning of the seventies. The main technological problem in making heterojunctions is the amount of defects at the

interface. Chapter 2 idescribes the interface defects arising in a metal-semiconductor junction, the surface defects are such to change completely the electrons energy and the device properties. This defects are much more evident in a junction between two different semiconductors where the materials should preserve their crystalline structure.
The most evident problem comes from the fact that since the crystalline periodicity is different and the materials have to interact each other, at the interface there is a strain of the crystals. This gives rise to interface defects that modify the work function (as seen in chapter 2), the mobility and the processes of generation and recombination. In practice, for many years the concentration of interface defects have been so high that the devices built with heterojunctions could not work properly.
A mitigation of the crystalline mismatch occurs in the alloys of semiconductors where the transition between one material to the other can be obtained with a gradual addition of one element. Examples of these materials are the silicon-germanium ($Si_{1-x}Ge_x$) and the aluminum-gallium-arsenide ($Al_xGa_{1-x}As$). The previous composition formulas are characterised by a variable x that is a measure of the alloy composition. For instance, in case of silicon-germanium, x=0 indicates pure silicon and x=1 is the case of pure germanium. Thus, as a function of x, the properties of the material change. The crucial quantity that is altered by the alloy composition is the energy gap. For the silicon-germanium case, the energy gap is between 1.12 eV (pure silicon) and 0.67 eV (pure germanium). For small values of x (of the order of x¡0.4) the energy gap of the alloy is given by a simple linear relationship.

$$E_G = 1.12 - 0.45 \cdot x \ eV \tag{8.1}$$

In case of $Al_xGa_{1-x}As$ the energy gap is in the interval from 1.14 eV (GaAs) and 2.16 eV (AlAs).
The lattice constant, namely the distance among atoms, is slightly different between the two pure materials, $a_{Si} = 5.43\mathring{A}$ and $a_{Ge} = 5.65\mathring{A}$. This small difference is enough to produce strains. Again, the lattice constant, for small values of x, is linearly variable with x.

$$a_{SiGe} = 5.43 + 0.04 \cdot x \ \ \mathring{A} \tag{8.2}$$

The fabrication of heterojunctions requires a technology able to grow a new material over a pre-existing one. These kind of growth is called epitaxial and in order to reduce the concentration of defects at the interface the rate of production of the new material has to be slow and controlled. Molecular Beam Epitaxy (MBE), developed at the end of the sixties, is a largely utilised technology for the fabrication of heterojunctions . Figure 8.1 shows a over simplified scheme of principle of a MBE apparatus.

8.2 Band diagram

Also for heterojunctions, the band diagram confirms to be a fundamental tool to study the property of devices.
Since none of the three features of the semiconductors are in principle the same, the relative positions of the bands of the materials before the formation of the junction can be anything. However, most of cases belong to two main configurations that are called staggered gap and straddled gap.
The staggered gap occurs when the conduction band of one material falls in the band gap of the

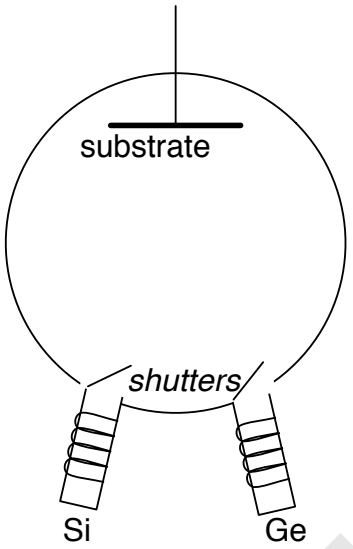


Fig. 8.1. *Scheme of principle of a Molecular Beam Epitaxy . The machine is a ultra high vacuum chamber ($P \approx 10^{-8}$ Pa) where the substrate is placed. The materials necessary to form the epitaxial film (silicon and germanium in the example) are placed in effusion cells. The effusion cells are heated above the sublimation point of the materials). The flow of evaporated atoms is controlled by the temperature and by mechanical shutters. In ultra high vacuum condition the free mean path of the evaporated atoms in gas phase is of the order of kilometers. Then the only impact that they have is with the surfaces and in particular the substrate. These machines are usually complemented with instruments to measure the thickness of the film and its composition.*

other, while in the straddled gap case the conduction and the valence bands on of one materials are contained in the band gap of the other material.
In the following, the equilibrium band diagrams of both cases are discussed.

8.2.1 Staggered bandgaps

Figure 8.2 shows the bands before the formation of the junction in the case of staggered band gaps. Let us consider a case where $q\chi_1 > q\chi_2$ of course, band gaps and work functions are also different. In this example the material with the largest band gap is n-type and the other is p-type.

The steps for the correct drawing of the equilibrium band diagram are outlined in table 1 of Chapter 2. Figure 8.3 shows the band alignment procedure applied to the caseof staggered band gaps.

As a consequence of the differences of affinities and energy gaps, the conduction and the valence bands are not continuos. The differences at the interface of conduction band and valence band energies are given by:

$$\Delta E_C = q(\chi_1 - \chi_2) = q\Delta\chi; \quad \Delta E_V = (q\chi_1 + E_{gap1}) - (q\chi_2 + E_{gap2}) = \Delta E_g + q\Delta\chi \tag{8.3}$$

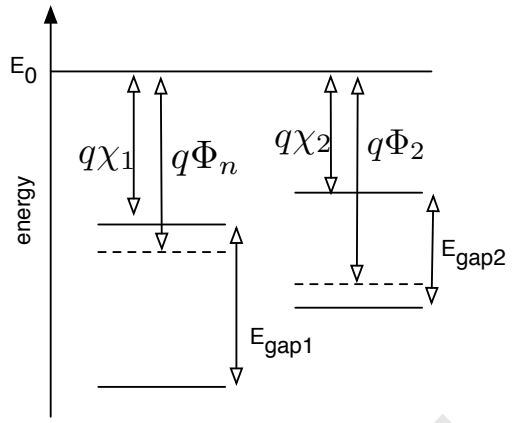


Fig. 8.2. *Band diagrams of staggered band gaps materials. In this example* $q\chi_1 > q\chi_2$, $q\phi_1 < q\phi_2$ *and* $E_{gap1} > E_{gap2}$. *Furthermore the material 1 is n-type and the material 2 is p-type.*

8.2.2 Straddled bandgaps

This case is typical of alloys, for instance Si vs. SiGe. In this arrangement, shown in figure 8.4, the band diagram of the material with the smallest band gap is contained in the band gap of the material with the largest band gap.

The application of the procedure outlined in figure 2 leads to the equilibrium band diagram in figure 8.5. The main difference between the two cases is the shape of the discontinuity of the conduction band. In the case of the straddled band gap the discontinuity has the shape of a spike. The main consequence of the spike is that the concentration of electrons in the band gap is not gradually decreasing from the n-type material to the p-type material but at the interface there is a small region where the concentration of electrons increases.
In normal conditions, this increase is modest, and the interface region is still depleted of electrons. However, under proper conditions, they spike may give rise to a narrow region where the concentration of electrons is very high concentration. This case is discussed in the last chapter of this textbook.

The difference of conduction and valence bands at the interface are still given by the same formulas of the staggered case:

$$\Delta E_C = q(\chi_1 - \chi_2) \quad \Delta E_V = \Delta E_g + q\Delta\chi \tag{8.4}$$

However, due to the relative magnitude of the affinities, ΔE_C is negative.
In figure 8.5 $q\Phi_{bi}$ is the contact potential difference that is equal to the difference of the work function. This quantity can only be measured at the surface of the materials and it is different from the built-in potential for the electrons ($q\Phi_{bn}$) and the built-in potential for the holes ($q\Phi_{bp}$). Thus, one of the most characteristic feature of heterojunctions, regardless the mutual position of the bands, is the different potential barriers acting on electrons and holes. This difference is at the basis of the behaviours that will be discussed later.
In figures 8.3 and 8.5 the curves connecting the non perturbed band diagram have been only

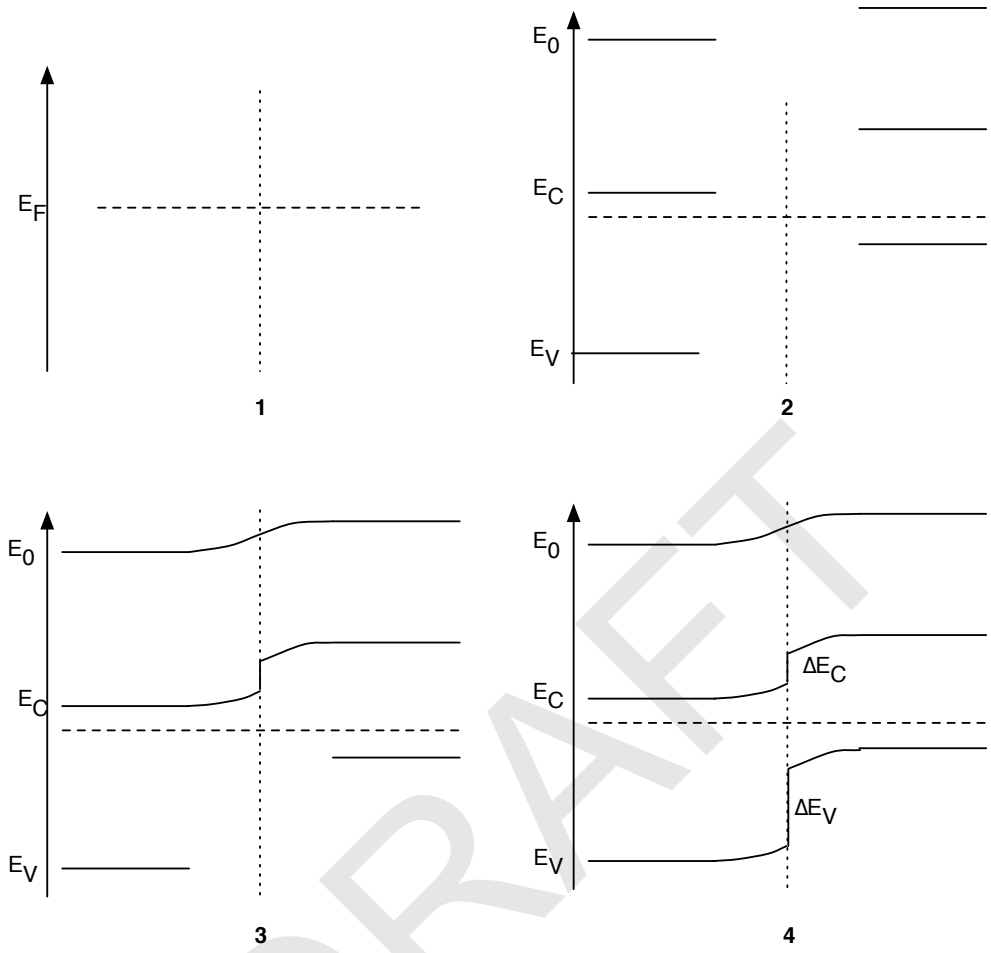


Fig. 8.3. *Steps for the construction of the equilibrium band diagram in the case of staggered bands.* **1** *At the equilibrium the Fermi level is constant.* **2** *The bulk of the material is not modified by the junction.* **3** *The vacuum level is continuous and the affinity is constant, then, in each material, the conduction band is parallel to the vacuum level. Since the affinities are different, at the interface the conduction band is not continuous.* **4** *The energy gap is constant, then the valence bands are parallel to the conduction bands. Since the band gaps are different, the valence band is not continuous at the interface.*

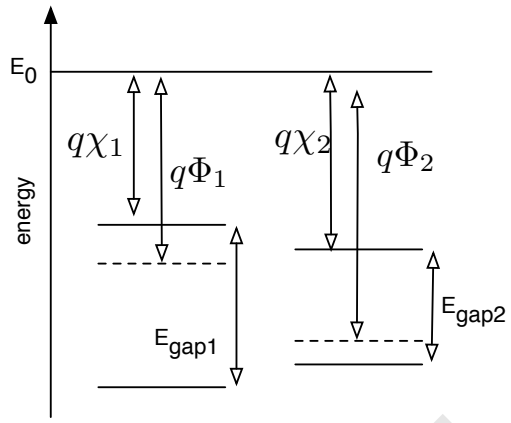


Fig. 8.4. *Band diagrams of straddled band gaps materials. In this example* $q\chi_1 < q\chi_2$, $q\phi_1 < q\phi_2$ *and* $E_{gap1} > E_{gap2}$. *Furthermore the material 1 is n-type and the material 2 is p-type.*

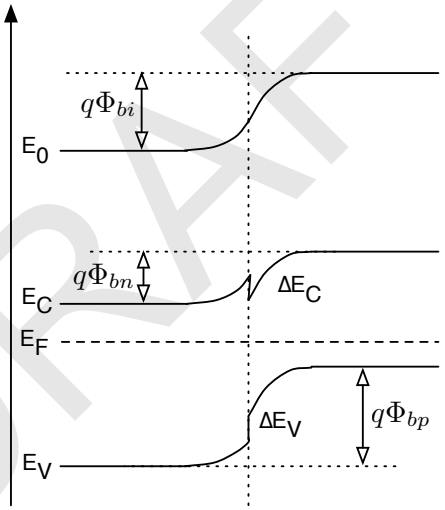


Fig. 8.5. *Equilibrium band diagram of the straddled band gap of figure 4.*

qualitatively drawn. The exact behaviour is the result of the Poisson equation. Since the straddled case is more frequent and more interesting, the electrostatic quantities are here calculated for this case. However, the extension of the calculus to the straddled case is totally straightforward.

8.3 Electric field and built-in potential

To solve the Poisson equation let us follow the same procedure used for the PN homojunction. An important difference here is the different dielectric constants in the two regions; then, since the amount of charges in the two regions has to be same (charge neutrality condition) the electric fields at the interface are different.
The main assumption for the solution of the Poisson equation concerns the distribution of the dopant atoms, and then the distribution of the total charge. Here we continue to apply the step junction and the deep depletion hypotheses. Namely, the dopants are uniformly distributed until the interface and the space charge region is completely depleted of the majority charges until the border of the neutral zones. The distribution of charge is shown in figure 8.6.

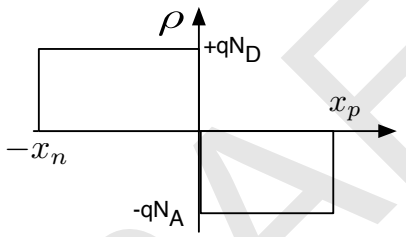


Fig. 8.6. *Charge density distribution according to the step junction and deep depletion hypothesis.*

The electric field in the two regions are calculated integrating the charge density distributions:

$$\mathcal{E}_n(x) = \frac{qN_D}{\epsilon_1}(x + x_n) \tag{8.5}$$

$$\mathcal{E}_p(x) = \frac{qN_A}{\epsilon_2}(x_p - x) \tag{8.6}$$

Since $\epsilon_1 \neq \epsilon_2$, the electric field is not continuous at the interface. Actually, the continuous quantity is the electric displacement D, whose relation with the electric field is: $D = \epsilon\mathcal{E}$. Then at the interface the continuity of the electric displacement $D_1(0) = D_2(0)$ provides:

$$\epsilon_1 \frac{qN_D}{\epsilon_1} x_n = \epsilon_2 \frac{qN_A}{\epsilon_2} x_p \tag{8.7}$$

Which still corresponds to the charge neutrality condition: $qN_D x_n = qN_a x_p$.
In the case of $Si - Si_{1-x}Ge_x$ junction, since $\epsilon_{Si} = 11.7$ and $\epsilon_{Ge} = 16$, we have $\epsilon_1 < \epsilon_2$ and the electric field in silicon is greater respect to the germanium side.

The potential is calculated integrating each electric field in its own region.

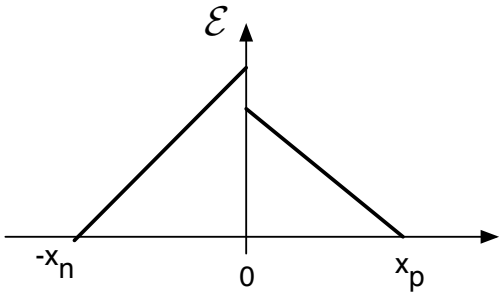


Fig. 8.7. *Electric field in case* $\epsilon_1 > \epsilon_2$.

$$\phi_1(x) = \phi(-x_n) - \frac{qN_D}{2\epsilon_1}(x + x_n)^2 \tag{8.8}$$

$$\phi_2(x) = \phi(x_p) + \frac{qN_A}{2\epsilon_2}(x_p - x)^2 \tag{8.9}$$

The behaviour of the potential is shown in figure 8.8.

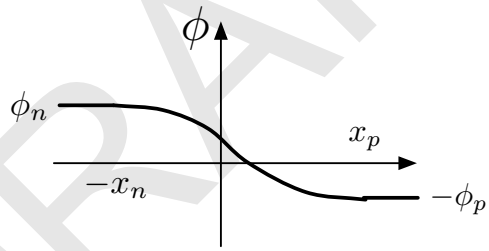


Fig. 8.8. *Electric potential in case* $\epsilon_1 > \epsilon_2$.

The potential, of course is continuous at the interface $\phi_1(0) = \phi_2(0)$.

$$\phi(-x_n) - \frac{qN_D}{2\epsilon_1}x_n^2 = \Phi(x_p) + \frac{qN_A}{2\epsilon_2}x_p^2 \tag{8.10}$$

Thus the built-in potential is:

$$\phi_{bi} = \phi(-x_n) - \phi(x_p) = \frac{qN_D}{2\epsilon_1}x_n^2 - \frac{qN_A}{2\epsilon_2}x_p^2 \tag{8.11}$$

It is important to remind that the built-in potential of an heterojunction appears only between the surfaces of the materials. As it will see in the next section, it has not a direct influence on the behaviour of the device but rather it allows for the measurement of the size of the depletion layer.

Indeed, following the case of PN homojunction, combining the charge neutrality condition with the built-in potential we can find the relationship between the depletion layer size and the materials parameters.

$$x_p = \sqrt{\frac{2\epsilon_1\epsilon_2 N_D \phi_{bi}}{qN_A(\epsilon_1 N_D + \epsilon_2 N_a)}} \quad x_n = \sqrt{\frac{2\epsilon_1\epsilon_2 N_A \phi_{bi}}{qN_D(\epsilon_1 N_D + \epsilon_2 N_a)}} \tag{8.12}$$

Note that, as found in the PN homojunction, the size of the depletion layer in each region is proportional to the doping in the opposite region.

8.4 Current-voltage relationship

The main consequence of the heterojunction is the different barriers that keep in equilibrium electrons and holes. We will see that this difference also implies that the currents of electrons and holes are different.
The calculus of the current/voltage relationship follows the same steps of the PN homojunction case. In particular, the ideal current of the PN junction is evaluated. The calculation moves from evaluating the concentration of charges at the equilibrium at the edge of the neutral zones. For this scope, it is necessary to evaluate the barriers for electrons and holes. Let us consider the equilibrium band diagram in figure 8.5. The concentration of the charges at the borders of the depletion layer is.

$$\begin{aligned} n_{n0} &= N_D \\ p_{p0} &= N_A \\ n_{p0} &= N_D exp\left(-\frac{q\phi_{bn}}{kT}\right) = \frac{n_{ip}^2}{N_A} \\ p_{n0} &= N_A exp\left(-\frac{q\phi_{bp}}{kT}\right) = \frac{n_{in}^2}{N_D} \end{aligned}$$

The concentrations of the minority charges at the equilibrium depends on the barrier heights but through the mass action law they depend also on the concentration of the intrinsic charge carriers. The quantity n_i^2 is a function of the energy gap: $n_i^2 = N_C N_V exp(-E_G/kT)$ then it is different in the two materials. The square of the intrinsic concentrations can be very different; considering the extreme case of silicon and germanium we have: $n_{iSi}^2 = 10^{20} cm^{-6}$ and $n_{iGe}^2 = 10^{26} cm^{-6}$.
This difference makes the concentration of the minority charges very different in the two semiconductors.
From the equation of the minority charges we have the relationship between the potential barriers and the energy gaps:

$$n_{p0} = N_D exp(-\frac{q\phi_{bn}}{kT}) = \frac{N_{Cp}N_{Vp}exp(-E_{Gp}/kT)}{N_A}$$
$$p_{n0} = N_A exp(-\frac{q\phi_{bp}}{kT}) = \frac{N_{Cn}N_{Vn}exp(-E_{Gn}/kT)}{N_D}$$

From where

$$q\phi_{bn} = E_{Gp} - kTln\left(\frac{N_{Cp}N_{Vp}exp(-E_{Gp}/kT)}{N_A N_D}\right)$$
$$q\phi_{bp} = E_{Gn} - kTln\left(\frac{N_{Cn}N_{Vn}exp(-E_{Gp}/kT)}{N_A N_D}\right)$$

the difference between the two barriers is:

$$q\phi_{bp} - q\phi_{bn} = E_{Gn} - E_{Gp} - kTln\left(\frac{N_{Cn}N_{Vn}}{N_{Cp}N_{Vp}}\right) \tag{8.13}$$

In chapter 1 the expressions of the density of the states was calculated:

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

The density of states depends on the effective mass of electrons and holes which are determined by the shape of the bands (the curvature). Thus, the effective mass is a peculiar characteristics of each material. However, the differences are modest and the density of states are barely similar in different materials.
In table 1 the effective masses and the density of states for silicon and germanium are listed. To appreciate the difference due to the density of state let us consider a case of $N_D = N_A = 10^{18}\ cm^{-3}$ in this case considering the extreme case of N-type silicon and P-type germanium, at room temperature, $kTln\left(\frac{N_{Cn}N_{Vn}}{N_{Cp}N_{Vp}}\right) \approx 40\ meV$. This small quantity is small with respect to the energy gaps difference and then the potential barrier applied to the holes can be simply written as:

$$q\phi_{bp} = q\phi_{bn} + \Delta(E_G) \tag{8.14}$$

where $\Delta(E_G)$ is the difference between the two band gaps. It is clear that the different energy gap makes the intrinsic concentrations different and, if the two materials are doped with the same amount of doping, the concentration of the minority charges might also be very different.

Under external bias, both the barriers change: $\phi'_{bn} = \phi_{bn} - V_A$ and $\phi'_{bp} = \phi_{bp} - V_A$. Then while the density of the majority charges remain constant the density of the minority charges increase

Table 8.1. Density of states

	Silicon	Germanium
N_C	$2.8 \cdot 10^{19}$	$1.04 \cdot 10^{19}$
N_V	$1.04 \cdot 10^{19}$	$0.60 \cdot 10^{19}$
m_n^*	$1.08 \cdot m_0$	$0.55 \cdot m_0$
m_p^*	$0.81 \cdot m_0$	$0.30 \cdot m_0$

$$n_p = N_D exp\left(-\frac{q\phi_{bn} - qV_A}{kT}\right)$$
$$p_n = N_A exp\left(-\frac{q\phi_{bn} - qV_A}{kT}\right) exp\left(-\frac{\Delta E_G}{kT}\right)$$

Which gives rise to the excess minority charges: $n' = n - n_0$ and $p' = p - p_0$.

$$n'_p = N_D exp\left(-\frac{q\phi_{bn}}{kT}\right)\left[exp(\frac{qV_A}{kT}) - 1\right]$$
$$p'_n = N_A exp\left(-\frac{q\phi_{bn}}{kT}\right) exp\left(-\frac{\Delta E_G}{kT}\right)\left[exp(\frac{qV_A}{kT}) - 1\right]$$

As seen in the discussion of the homojunction, the charges in excess at the border of the neutral zone diffuse and during the diffusion are recombined. Under a stationary bias, this results in a steady distribution of minority charges that gives rise to a diffusion current. The expression of the current depends on the size of the neutral zone. In case of a long-base diode, namely when the distance between the depletion layer and the electrode is much larger than the recombination length, the recombination consumes the majority charges, a diffusion current of majority charges is required to maintain the steady state condition. The majority charges current has the same magnitude but opposite direction. Eventually, electrons and holes produce a net current:

$$J_n = q\frac{D_n}{L_n} N_D exp\left(-\frac{q\phi_{bn}}{kT}\right)\left[exp(\frac{qV_A}{kT}) - 1\right] \tag{8.15}$$

$$J_p = q\frac{D_n}{L_n} N_D exp\left(-\frac{q\phi_{bn}}{kT}\right) exp\left(-\frac{\Delta E_G}{kT}\right)\left[exp(\frac{qV_A}{kT}) - 1\right] \tag{8.16}$$

On the other hand,in a short-base diode L_n is replaced by W_B and L_p by W_E.
The ratio between the two currents is:

$$\frac{J_n}{J_p} = \frac{D_n}{L_n}\frac{L_p}{D_p}\frac{N_D}{N_A} exp\left(\frac{\Delta E_G}{kT}\right) \tag{8.17}$$

Except the last term, this is the same equation found in the case of the PN homojunction. In that case the ratio between the current was different by one only if the dopings were different. In

the case of heterojunction, the ratio can be very different even in case of equal doping. This result evidences that the current in the junction is dominated by the minority charges.
For instance in case of a junction between Si and $Si_{1-x}Ge_x$ with x=0.3 the ratio $\Delta E_G/kT = 5.4$ whose exponential is about 220.

8.4.1 Thermionic current

In case of straddled band gap, the discontinuity of the conduction band at the interface takes the shape of a spike. The spike means that the concentration of electrons inside the depletion layer abruptly increases immediately beyond the interface. If the bottom of the spike is sufficiently above the Fermi level the increase of concentration does not interfere with the depletion layer hypothesis, and the concentration of electrons still remains negligible.
The main hypothesis in the calculation of the current is that the concentration of minority charges derive from the concentration of the majority charges in the opposite region weighted by the potential barrier that is the difference on the conduction band at the border of the space charge region. This hypothesis is valid if the top of the spike stays below the conduction band in the neutral region. As the applied voltage increases, it may happen that the the top of the spike exceeds the conductance band, as shown in figure 9.
In this condition, the actual barrier for the electrons is $E_s - E_{C1}$ where E_s is the energy at the top of the spike. Since V_A is applied across the whole depletion layer, the voltage that drops between the top of the spike and the conduction band is lower that V_A. Thus there is a smaller efficiency in the current.

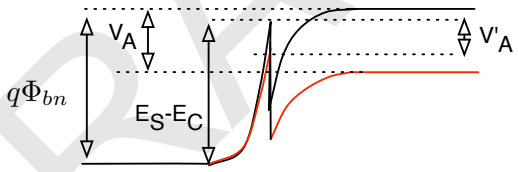


Fig. 8.9. *Conduction band of straddled band gap case before and after the applied voltage.*

In this condition, the current is coincident with the flow of electrons that overpass the barrier $E_s - E_C$. In practice, it is similar to the thermoionic current of the Schottky diode. Eventually, a decrease of the differential conductance is observed for large applied voltage.

Finally, it has to be remarked that, even in case of a graded alloy, in a heterojunction the interface region is much more defected with respect to a homojunction. Then, a more efficient recombination in the depletion layer is expected with respect to a normal homojunction. This will increase the magnitude of the recombination current in the depletion layer under direct bias. In comparison with a homojunction, the effect of the recombination current persists for a larger value of the applied voltage.

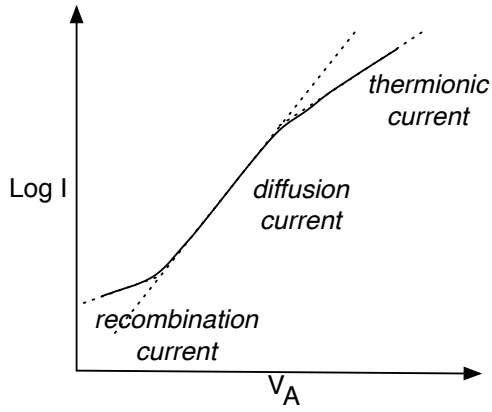


Fig. 8.10. *Qualitative current-voltage relationship of a heterojunction.*

8.5 Heterojunction Bipolar Transistor

One of the most immediate applications of heterojunctions is the Heterojunction Bipolar Transistor (HBT). The advantages offered by heterojunctions were theoretically clear since the first development of BJT, and the concept of HBT was contained in the original patent released by the Bell Telephone in 1948. The patent contains a design of HBT where the base-emitter junction is an heterojunction and the base-collector junction is a homojunction.
In a n-p-n device, the material of the emitter has a larger band gap with respect to the material where base and collector are made. The majority charges of the large band gap material (electrons in a n-p-n device) defines the main current of the BJT.
Figure 11 shows the equilibrium band diagram of such a device

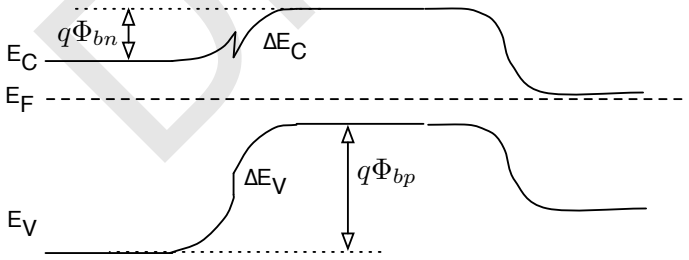


Fig. 8.11. *Equilibrium band diagram of a n-p-n HBT.*

The most important consequence of the heterojunction is the asimmetry between the currents of electrons and holes across the base-emitter junction. The ratio between the currents is proportional

to the exponential of the difference between the energy gaps. The ratio between the currents is important for the emitter efficiency that is the most important element of the current gain

$$\gamma = \frac{1}{1 + \frac{J_p e}{J_n e}} \tag{8.18}$$

Using the same nomenclature used in chapter 6 to describe the BJT in the equation 18, the emitter efficiency can be written as:

$$\gamma = \frac{1}{1 + \frac{x_B N_A D_{pe}}{x_E N_D D_{nb}} exp(-\frac{\Delta E_G}{kT})}) \tag{8.19}$$

The above equation holds for uniform doping distributions, in case of variable doping profiles the Gummel numbers replaces the product of doping and neutral regions size:

$$\gamma = \frac{1}{1 + \frac{G N_B D_{pe}}{G N_E D_{nb}} exp(-\frac{\Delta E_G}{kT})}) \tag{8.20}$$

Thanks to the exponential of the difference of energy gaps, the emitter efficiency can be very close to one even if the condition $GN_B \ll GN_E$ is not fulfilled. Then with the HBT in order to obtain a large current gain is not necessary a small doping of the base, but the gain is actually achieved even if the base if highly doped.

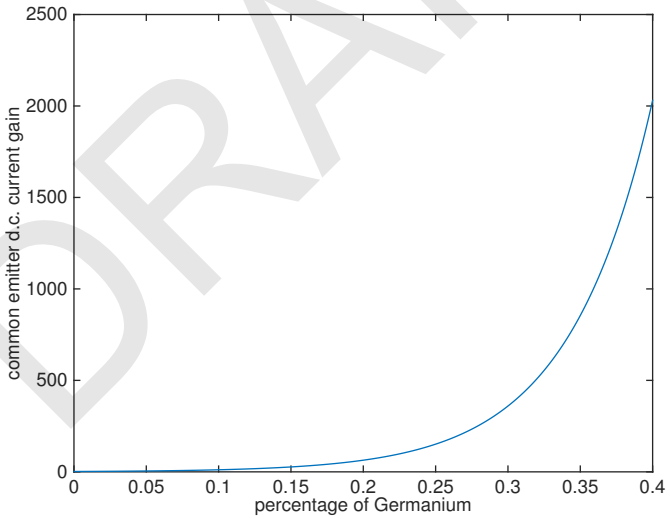


Fig. 8.12. *β_f as a function of the x percentage of Germanium in a $Si - Si_{1-x}Gex$ base-emitter junction.*

A large doping of the base is an efficient counteraction against the Kirk effect (extending the low injection limit of the device) and the Early effect improving the current source characteristics at the output of the device.

Actually the amplification involves not only the emitter efficiency but also the transport factor in base:

$$\alpha_T = 1 - \frac{x_B^2}{2L_n^2} \tag{8.21}$$

In BJT the size of the base cannot be narrowed too much because of the Early effect and the consequent punchthrough of the base. In HBT the large doping of the base, strongly reduces the Early effect and the risk of punchthrough. So, since the base length is immune from the bias, the base can be made very narrow and α_T is practically equal to one.
Eventually, the current amplification factor β_F depends only on the emitter efficiency.

$$\beta_f = \frac{\gamma\alpha_T}{1-\gamma\alpha_T} \approx \frac{\gamma}{1-\gamma} \tag{8.22}$$

Due to the influence of the recombination current and the thermionic current, the current gain is expected to be at the maximum level for a shorter interval of V_{BE} with respect to a homojunction BJT.
Another important advantage of HBT is concerned with the dynamic response. Since the base may be highly doped the base resistance may be made very small. In this way the base is immune to

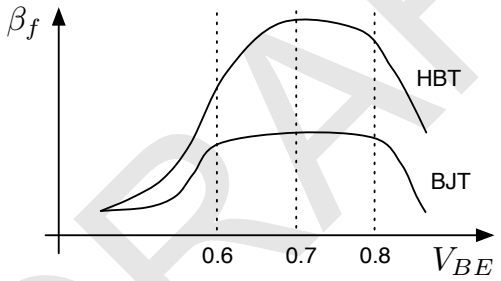


Fig. 8.13. *Qualitative behavior of β_f vs. V_{BE} for HBT and BJT. Typically, HBT exhibits a larger β_f but for a shorter range of V_{BE}.*

Figure 14 shows a simplified scheme of a HBT made in GaAs technology. The emitter is made in n-AlGaAs and the base is a subtle layer of $p^+ - GaAs$. Note that due to the low conductivity of the intrinsic GaAs ($n_i \approx 10^6\ cm^{-3}$ the device is insulated from the substrate.

8.5.1 Graded band gap

The technology of band gap engineering allows for the preparation of materials where the band gap can be graded by a continuous variation of the x parameter. Materials with a graded band gap show an interesting effect about their built-in electric field. Indeed, in materials characterized by a constant band gap, any applied potential, either built-in or external, always bends the bands in a parallel way. This means that electrons in conduction band and holes in valence band react

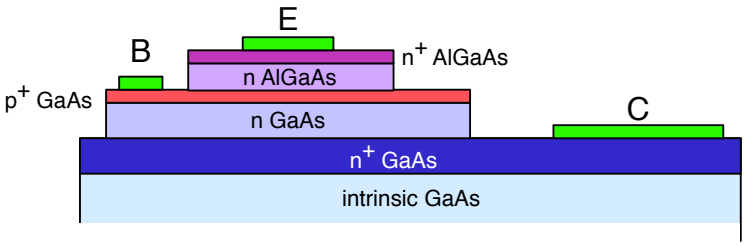


Fig. 8.14. *Simplified scheme of a epitaxial HBT in GaAs technology.*

to the same electric field moving in opposite directions. On the other hand, in a graded band gap semiconductor, even intrinsic, since n_i^2 is different at each section of the material, the gradient of the concentration of charges are parallel. This means that the bands are not parallel but convergent and the electric fields acting on electrons and holes point to different directions.
The conduction band is the potential energy of the electrons; thus, its derivative is the force acting on the electrons. The derivative of the conduction band in case of graded band gap has to include also the variability of the affinity ($q\chi$), then the total force for electrons is:

$$F_n = -\frac{dE_c}{dx} = q\frac{d\phi}{dx} + q\frac{d\chi}{dx} \tag{8.23}$$

similarly, the force applied to the holes is:

$$F_p = +\frac{dE_v}{dx} = q\frac{d\phi}{dx} - q\frac{d(\chi + E_g)}{dx} \tag{8.24}$$

The forces applied to electrons and holes are not iequal neither in magnitude nor in direction, as it should be expect for forces due to real electric fields. Actually, the electric field is only one component of these forces. Since it is usual to think of electric fields producing forces on charges, the fields acting on the charges are called *quasi-electric fields*.
In a doped material, let us suppose P-type, the quasi electric field tend to separate the holes (mobile) from the acceptors (fixed). This displacement gives rise to a built-in potential that cancel the quasi electric field of the holes. As we have seen in section 2.5, a non uniform distribution of charges evolves to the equilibrium creating a built-in potential, and the displacement of charges necessary for the equilibrium is negligible with respect to the initial distribution of charges. Then, if the material is P-type doped by N_A acceptors, the concentration of holes at the equilibrium is N_A everywhere and the valence band is flat at a constant distance from the Fermi level.
However, the built-in electric field that cancelled the quasi-electric field of the holes, acts on the electrons. Then, the electrons are subjected to an electric field while the holes are kept at a null electric field.

Graded band gap materials may help to improve some features of electronic devices. For instance, the response time of bipolar transistors is limited by the transit time in base. It was discussed in chapter 6 that a method to reduce the transit time is to create an electric field in the base region

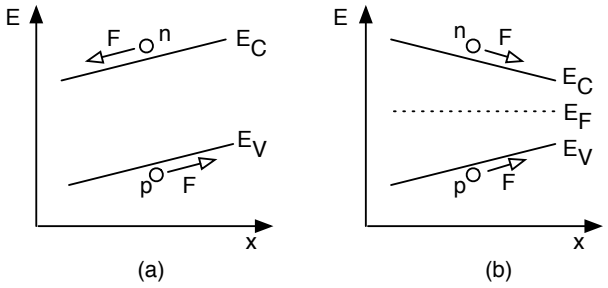


Fig. 8.15. *(a) In a material with constant band gap, an external voltage bends the bands in the same direction, and electrons and holes move in opposite directions. (b) The built-in potential in a graded band gap material forces the charges to move in the same direction.*

by a graded doping.
The technology of band gap engineering offers a more efficient solution to the problem avoiding the reduction of doping at the base-collector interface that can counteract both the Kirk and the Early effects. In practice the base is fabricated with a material with a graded decrease of the band gap and doped with a uniform concentration of acceptors. The band diagrams before the equilibrium and at the equilibrium are shown in figure 12.

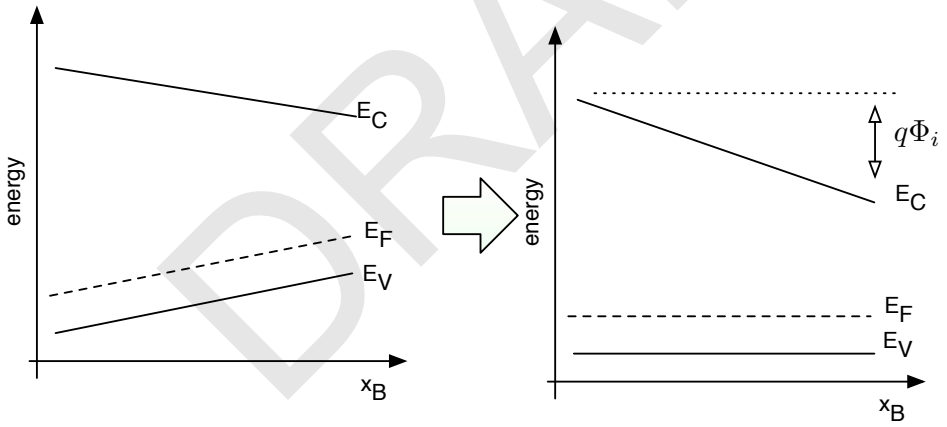


Fig. 8.16. *Bands diagram of a graded band gap material uniformly p-type doped*

The potential applied to the electrons is:

$$\phi_i = \frac{kT}{q} ln \frac{n_c}{n_e} \tag{8.25}$$

where n_e and n_c are the concentrations of electrons at the interface with the emitter and the collector respectively.
These concentrations can be calculated from the mass-action law

$$n_e = \frac{n_{i1}^2}{N_A} = \frac{N_C N_V}{N_A} exp\left(-\frac{\Delta E_{Ge}}{kT}\right)$$
$$n_c = \frac{n_{i2}^2}{N_A} = \frac{N_C N_V}{N_A} exp\left(-\frac{\Delta E_{Gc}}{kT}\right)$$

Then the potential is given by:

$$\phi_i = \frac{kT}{q} ln\left[exp\left(-\frac{\Delta E_{Gc} - \Delta E_{Ge}}{kT}\right)\right] = \frac{\Delta E_G}{q} \tag{8.26}$$

With such a potential, the transit time due to drift is:

$$\tau_{tr}^{drift} = \frac{x_B^2}{\mu_n \Delta V} = \frac{q x_B^2}{\mu_n \Delta E_G} \tag{8.27}$$

It is important also to point out that in a graded doping material, the graded doping gives rise to an electric field that accelerate both electrons and holes. Then as the transit time of electrons decreases also the holes current slightly increases. On the other hand, in a graded band gap material, only the electrons are accelerated.
In order to be efficient, the transit time due to drift has to be greater than the diffusion transit time: $\tau_{tr}^{diff} = \frac{x_B^2}{2D_n}$. To compare the two transit times it is then convenient to express the mobility through the diffusion coefficient using the Einstein relation. Then the ratio between the two transit times can be written as:

$$\frac{\tau_{tr}^{drift}}{\tau_{tr}^{diff}} = \frac{2kT}{\Delta E_G} \tag{8.28}$$

The ratio is smaller than one if $\Delta E_G > 2(kT/q) \approx 52meV$.
in $Si_{1-x}Ge_x$, this condition is fulfilled with an increase of the germanium of $\Delta x = 0.1$. Than a little increment of germanium is enough to improve the transit time of the electrons in the base and then to reduce the diffusion capacitance and to extend the bandwidth of the device. In terms of frequency range, Si/SiGe devices can work up to 300 GHz, the largest operating frequency up to 500 GHz has been achieved by a InP/InGaAs device.
The technique of graded band gap can also be used at the emitter-base junction to smooth the electrons concentration spike at the interface and then to reduce the influence of the thermionic current.