

## 5. The bipolar junction transistor

### 5.1. Introduction

The bipolar junction transistor was the first solid-state amplifier element and started the solid-state electronics revolution. It was invented by Bardeen, Brittain and Shockley at the Bell Laboratories in 1948 as part of a post-war effort to replace vacuum tubes with solid state elements. Their work led them first to the point-contact transistor and then to the junction transistor. They used germanium as the semiconductor of choice because it was possible to obtain high purity material. The extraordinary large diffusion length of minority carriers in germanium provided functional structures despite the large dimensions of the early devices.

Since then, the technology has progressed rapidly. The development of a planar process yielded the first circuits on a chip and for a decade, bipolar transistor operational amplifiers, like the 741, and digital TTL circuits were the workhorse of any circuit designer.

The spectacular rise of the MOSFET market share during the last decade has completely removed the bipolar transistor from center stage. Logic circuits, microprocessor and memory chips contain exclusively MOSFETs.

Bipolar transistors remain important devices for ultra-high-speed discrete logic circuits such as ECL, power-switching applications and in microwave power amplifiers.

In this chapter we first present the structure of the bipolar transistor and show how a three layer structure with alternating n-type and p-type regions can provide current and voltage amplification. We then present the ideal transistor model and derive an expression for the current gain in the forward active mode of operation. Next, we discuss the non-ideal effects, the modulation of the base width, recombination in the depletion region of the base-emitter junction and high injection effects. This discussion is followed with a description of the different circuit models and the bipolar transistor fabrication technology.

### 5.2. Structure and principle of operation

A bipolar junction transistor consists of two back-to-back p-n junctions, which share a thin common region with width,  $w_B$ . Contacts are made to all three regions, the two outer regions called the emitter and collector and the middle region called the base. The structure of an NPN bipolar transistor is shown in Figure 5.1 (a). The device is called “bipolar” since its operation involves both types of mobile carriers, electrons and holes.

## Bipolar junction transistors

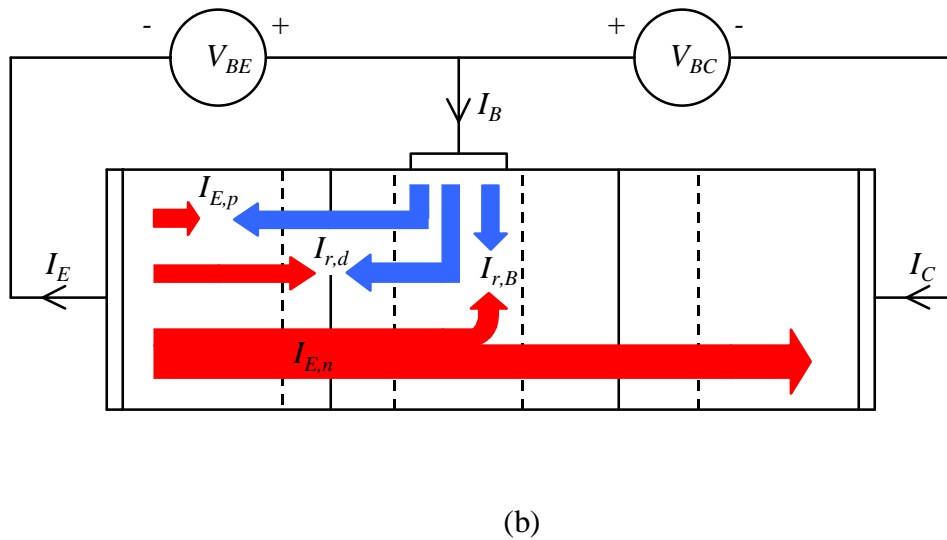
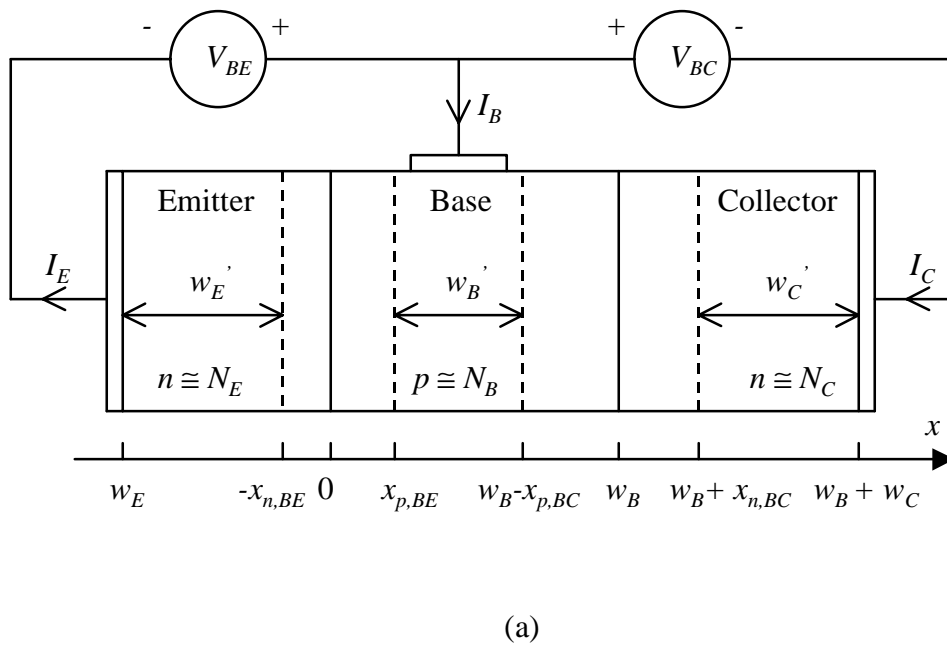


Figure 5.1. (a) Structure and sign convention of a NPN bipolar junction transistor. (b) Electron and hole flow under forward active bias,  $V_{BE} > 0$  and  $V_{BC} = 0$ .

Since the device consists of two back-to-back diodes, there are depletion regions between the quasi-neutral regions. The width of the quasi neutral regions in the emitter, base and collector are indicated with the symbols  $w_E'$ ,  $w_B'$  and  $w_C'$  and are calculated from

$$w_E' = w_E - x_{n,BE}$$

$$w_B' = w_B - x_{p,BE} - x_{p,BC}$$

$$w_C' = w_C - x_{n,BC}$$

where the depletion region widths are given by:

$$x_{n,BE} = \sqrt{\frac{2\mathcal{E}_s(\phi_i - V_{BE})}{q} \frac{N_B}{N_E} \left( \frac{1}{N_B + N_E} \right)}$$

$$x_{p,BE} = \sqrt{\frac{2\mathcal{E}_s(\phi_i - V_{BE})}{q} \frac{N_E}{N_B} \left( \frac{1}{N_B + N_E} \right)}$$

$$x_{p,BC} = \sqrt{\frac{2\mathcal{E}_s(\phi_i - V_{BC})}{q} \frac{N_C}{N_B} \left( \frac{1}{N_B + N_C} \right)}$$

$$x_{n,BC} = \sqrt{\frac{2\mathcal{E}_s(\phi_i - V_{BC})}{q} \frac{N_B}{N_C} \left( \frac{1}{N_B + N_C} \right)}$$

The sign convention of the currents and voltage is indicated on Figure 5.1(a). The base and collector current are positive if a positive current goes into the base or collector contact. The emitter current is positive for a current coming out of the emitter contact. This also implies that:

$$(5.1) \quad I_E = I_C + I_B$$

The base-emitter voltage and the base-collector voltage are positive if a more positive voltage is applied to the base contact.

The operation of the device is illustrated with Figure 5.1 (b). We consider the forward active bias mode of operation, obtained by forward biasing the base-emitter junction and reverse biasing the base-collector junction. To simplify the discussion further, we also set  $V_{CE} = 0$ . Electrons diffuse from the emitter into the base and holes diffuse from the base into the emitter. This carrier diffusion is identical to that in a p-n junction. However, what is different is that the electrons can diffuse as minority carriers through the quasi-neutral region in the base. Once the electrons arrive at the base-collector depletion region, they are swept through the depletion layer due to the electric field. These electrons contribute to the collector current. In addition, there are two more currents, the base recombination current and the depletion layer recombination.

The total emitter current is therefore the sum of the electron diffusion current,  $I_{E,n}$ , the hole diffusion current,  $I_{E,p}$  and the depletion layer recombination current,  $I_{r,d}$ .

$$(5.2) \quad I_E = I_{E,n} + I_{E,p} + I_{r,d}$$

The total collector current is the electron diffusion current,  $I_{E,n}$ , minus the base recombination

current,  $I_{r,B}$ .

$$(5.3) \quad I_C = I_{E,n} - I_{r,B}$$

The base current is the sum of the hole diffusion current,  $I_{E,p}$ , the base recombination current,  $I_{r,B}$  and the depletion layer recombination current,  $I_{r,d}$ .

$$I_B = I_{E,p} + I_{r,B} + I_{r,d}$$

The transport factor is defined as the ratio of the collector and emitter current:

$$(5.4) \quad \alpha = \frac{I_C}{I_E}$$

Using Kirchoff's current law and the sign convention shown in Figure 5.1(a), we find that the base current equals the difference between the emitter and collector current. The current gain is defined as the ratio of the collector and base current and equals:

$$(5.5) \quad \beta = \frac{I_C}{I_B} = \frac{\alpha}{1-\alpha}$$

This explains how a bipolar junction transistor can provide current amplification. If the collector current is almost equal to the emitter current, the transport factor,  $\alpha$ , approaches one. The current gain,  $\beta$ , can therefore become much larger than one.

To facilitate further analysis, we now rewrite the transport factor,  $\alpha$ , as the product of the emitter efficiency,  $\gamma_E$ , the base transport factor,  $\alpha_T$ , and the depletion layer recombination factor,  $\delta_r$ .

$$(5.6) \quad \alpha = \alpha_T \gamma_E \delta_r$$

The emitter efficiency,  $\gamma_E$ , is defined as the ratio of the electron current in the emitter,  $I_{E,n}$ , to the sum of the electron and hole current diffusing across the base-emitter junction,  $I_{E,n} + I_{E,p}$ .

$$(5.7) \quad \gamma_E = \frac{I_{E,n}}{I_{E,n} + I_{E,p}}$$

The base transport factor,  $\alpha_T$ , equals the ratio of the current due to electrons injected in the collector, to the current due to electrons injected in the base.

$$(5.8) \quad \alpha_T = \frac{I_{E,n} - I_{r,B}}{I_{E,n}}$$

Recombination in the depletion-region of the base-emitter junction further reduces the current gain, as it increases the emitter current without increasing the collector current. The depletion

layer recombination factor,  $\delta_r$ , equals the ratio of the current due to electron and hole diffusion across the base-emitter junction to the total emitter current:

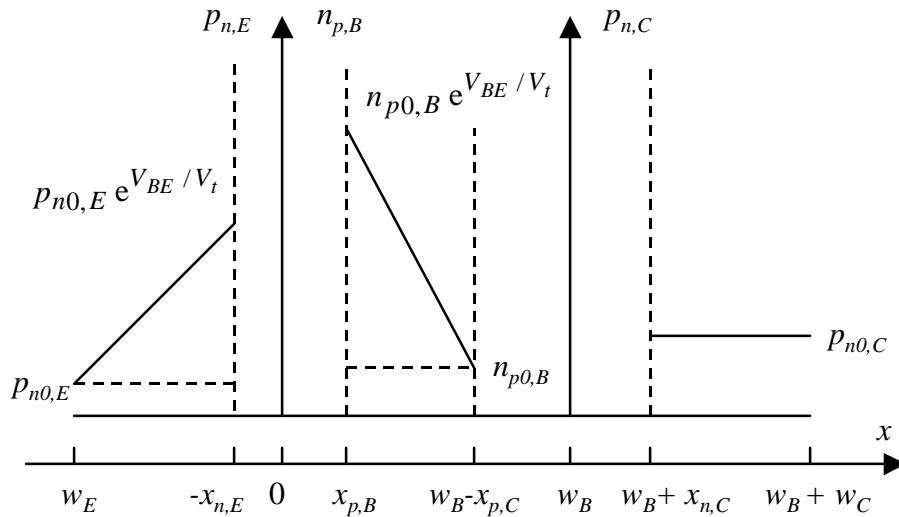
$$(5.9) \quad \delta_r = \frac{I_E - I_{r,d}}{I_E}$$

### 5.3. Ideal transistor model

The ideal transistor model is based on the ideal p-n diode model. We will assume that all quasi-neutral regions in the device are smaller than the minority-carrier diffusion lengths in these regions, so that the "short" diode expressions apply. We will ignore the recombination within the depletion region. A discussion of this recombination current can be found in section 5.3.2.

#### 5.3.1. Forward active mode of operation

The forward active bias is obtained by forward biasing the base-emitter junction. In addition we reverse the base-collector junction and will ignore the base-collector junction current by setting  $V_{BC} = 0$ . The minority-carrier distribution in the quasi-neutral regions of the bipolar transistor, as shown in Figure 5.2, is used to analyze this situation in more detail.



(a)

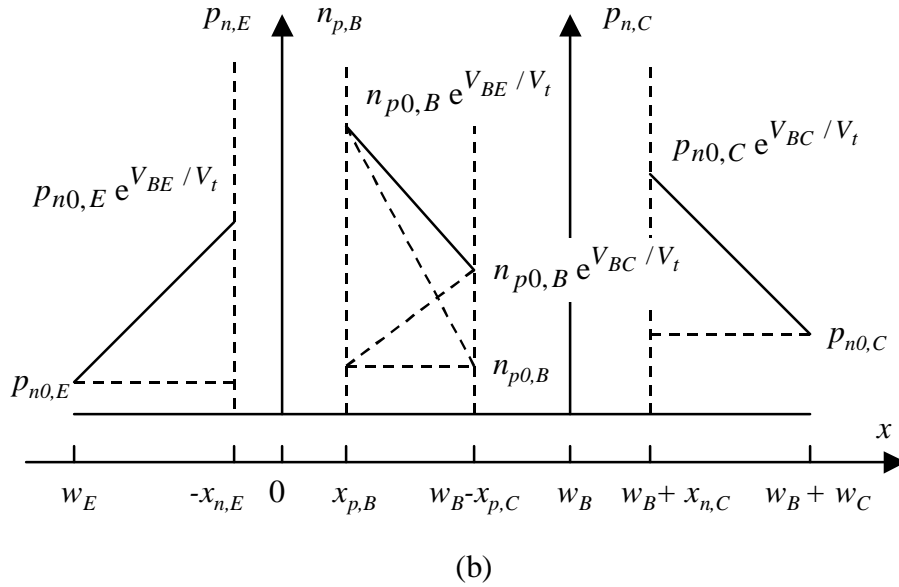


Figure 5.2. Minority-carrier distribution in the quasi-neutral regions of a bipolar transistor (a) Forward active bias mode. (b) Saturation mode.

The values of the minority carrier densities at the edges of the depletion regions are indicated on the figure. The carrier densities vary linearly between the boundary values as expected when using the assumption that no significant recombination takes place in the quasi-neutral regions. The minority carrier densities on both sides of the base-collector depletion region equal the thermal equilibrium values since  $V_{BC}$  was set to zero. While this boundary condition is mathematically equivalent to that of an ideal contact, there is an important difference. The minority carriers arriving at  $x = x_B - x_{p,C}$  do not recombine. Instead, they drift through the base-collector depletion region and become majority carriers in the collector region.

The emitter current due to electrons and holes are obtained using the "short" diode expressions, yielding:

$$(5.10) \quad I_{E,n} = qn_i^2 A_E \left( \frac{D_{n,B}}{N_B w_B} \right) \left( \exp\left(\frac{V_{BE}}{V_t}\right) - 1 \right)$$

and

$$(5.11) \quad I_{E,p} = qn_i^2 A_E \left( \frac{D_{p,E}}{N_E w_E} \right) \left( \exp\left(\frac{V_{BE}}{V_t}\right) - 1 \right)$$

It is convenient to rewrite the emitter current due to electrons,  $I_{E,n}$ , as a function of the total excess minority charge in the base,  $\Delta Q_{n,B}$ . This charge is proportional to the triangular area in the quasi-neutral base as shown in Figure 5.2 and is calculated from:

$$(5.12) \quad \Delta Q_{n,B} = q A_E \int_{x_{p,E}}^{w_B - x_{p,C}} n_p(x) - n_{p0} dx$$

which for a "short" diode becomes:

$$(5.13) \quad \Delta Q_{n,B} = q A_E \frac{n_i^2}{N_B} \left( \exp\left(\frac{V_{BE}}{V_t}\right) - 1 \right) \frac{w_B'}{2}$$

And the emitter current due to electrons,  $I_{E,n}$ , simplifies to:

$$(5.14) \quad I_{E,n} = \frac{\Delta Q_{n,B}}{t_r}$$

where  $t_r$  is the average time the minority carriers spend in the base layer, i.e. the transit time. A comparison between equations (5.10), (5.13) and (5.8) yields the transit time as a function of the quasi-neutral layer width,  $w_B'$ , and the electron diffusion constant in the base,  $D_{n,B}$ .

$$(5.15) \quad t_r = \frac{w_B'^2}{2D_{n,B}}$$

We now turn our attention to the recombination current and obtain it from the continuity equation:

$$(5.16) \quad \frac{\partial n_p(x)}{\partial t} = \frac{1}{q} \frac{\partial J_n(x)}{\partial x} - \frac{n_p(x) - n_{p0}}{\tau_n}$$

In steady state and applied to the quasi-neutral region in the base, the continuity equation yields the base recombination current,  $I_{r,B}$ :

$$(5.17) \quad I_{r,B} = q A_E \int_{x_{p,BE}}^{w_B - x_{p,BC}} \frac{n_p(x) - n_{p0}}{\tau_n} dx$$

which in turn can be written as a function of the excess minority carrier charge,  $\Delta Q_{n,B}$ , using equation (5.12).

$$(5.18) \quad I_{r,B} = \frac{\Delta Q_{n,B}}{\tau_n}$$

The emitter efficiency, defined by equation (5.7), becomes:

$$(5.19) \quad \gamma_E = \frac{1}{1 + \frac{D_{p,E} N_B w_B}{D_{n,B} N_E w_E}}$$

It is typically the emitter efficiency, which limits the current gain in transistors made of silicon or germanium. The long minority-carrier lifetime and the long diffusion lengths in those materials justify the exclusion of recombination in the base or the depletion layer. The resulting current gain, under such conditions, is:

$$(5.20) \quad \beta \cong \frac{D_{n,B} N_E w_E}{D_{p,E} N_B w_B}, \text{ if } \alpha \cong \gamma_E$$

From this equation, we conclude that the current gain can be larger than one if the emitter doping is much larger than the base doping. A typical current gain for a silicon bipolar transistor is 50 - 150.

The base transport factor, as defined in equation (5.8), equals:

$$(5.21) \quad \alpha_T = 1 - \frac{t_r}{\tau_n} = 1 - \frac{w_B^2}{2D_{n,B}\tau_n}$$

Keep in mind that this expression is only valid if the base transport factor is very close to one.

### 5.3.2. General bias modes of a bipolar transistor

While the forward active mode of operation is the most useful when using a bipolar junction transistor as an amplifier, one can not ignore the other bias modes. All possible bias modes are illustrated with Figure 5.3. They are the forward active mode of operation, the reverse active mode of operation, the saturation mode and the cut-off mode.

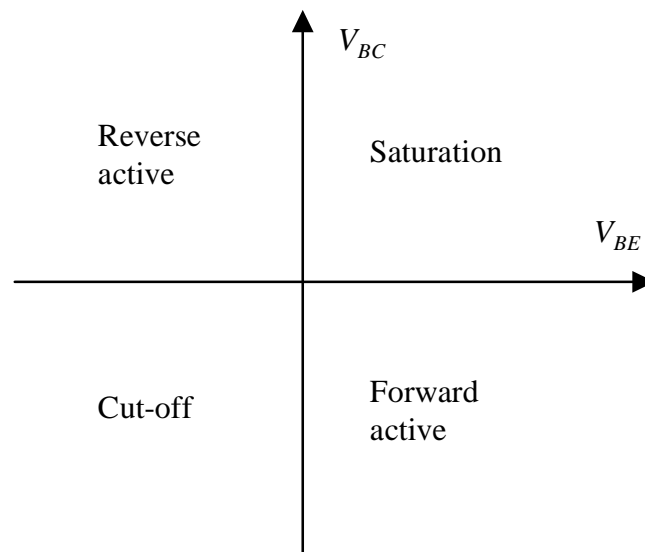


Figure 5.3. Possible bias modes of operation of a bipolar junction transistor.

The forward active mode is the one where we forward bias the base-emitter junction,  $V_{BE} > 0$  and reverse bias the base-collector junction,  $V_{BC} < 0$ . This mode, as discussed in section 5.2.1, is the one used in bipolar transistor amplifiers. In bipolar transistor logic circuits, one frequently switches the transistor from the “off” state to the low resistance “on” state. This “off” state is the cut-off mode and the “on” state is the saturation mode. In the cut-off mode, both junctions are reverse biased,  $V_{BE} < 0$  and  $V_{BC} < 0$ , so that very little current goes through the device. This corresponds to the “off” state of the device. In the saturation mode, both junctions are forward biased,  $V_{BE} > 0$  and  $V_{CB} > 0$ . This corresponds to the low resistance “on” state of the transistor.

Finally, there is the reverse active mode of operation. This mode is one where we reverse the function of the emitter and the collector. We reverse bias the base-emitter junction and forward bias the base-collector junction, or  $V_{BE} < 0$  and  $V_{BC} > 0$ . In this mode, the transistor has an emitter efficiency and base transport factor as described by equations (5.19) and (5.21), where we replace the emitter parameters by the collector parameters. Most transistors, however, have poor emitter efficiency in reverse active bias since the collector doping density is typically much less than the base doping density to ensure high base-collector breakdown voltages. In addition, the collector-base area is typically larger than the emitter-base area, so that even fewer electrons can make it from the collector into the emitter.

Having described the forward active mode of operation, there remains the saturation mode, which needs further discussion. Cut-off requires little further analysis, while the reverse active mode of operation is analogous to the forward active mode with the added complication that the areas of the base-emitter and base-collector junction,  $A_E$  and  $A_C$ , differ. The Ebers-Moll model describes all of these bias modes.

### 5.3.3. The Ebers-Moll model

The Ebers-Moll model is a model for a bipolar transistor, which can be used, in the forward

active mode of operation, in the reverse active mode, in saturation and in cut-off. This model is the predecessor of today's computer simulation models and contains only the “ideal” diode currents.

The model contains two diodes and two current sources as shown in Figure 5.4. The two diodes represent the base-emitter and base-collector diodes. The current sources quantify the transport of minority carriers through the base region. These are current sources depend on the current through each diodes. The parameters  $I_{E,s}$ ,  $I_{C,s}$ ,  $\alpha_F$  and  $\alpha_R$  are the saturation currents of the base-emitter and base collector diode and the forward and reverse transport factors.

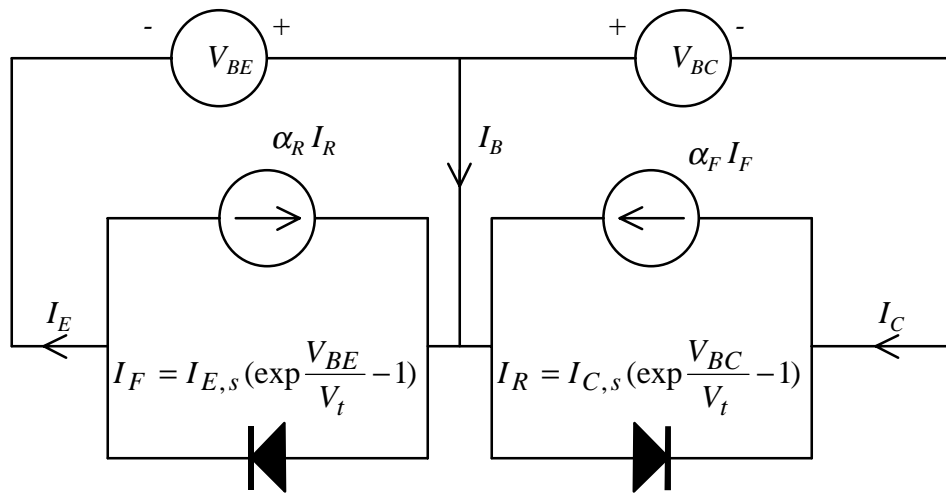


Figure 5.4 Equivalent circuit for the Ebers-Moll model of an NPN bipolar junction transistor

Using the parameters identified in Figure 5.4, we can relate the emitter, base and collector current to the forward and reverse currents and transport factors, yielding:

$$(5.22) \quad I_E = I_F - \alpha_R I_R$$

$$(5.23) \quad I_B = (1 - \alpha_F) I_F + (1 - \alpha_R) I_R$$

$$(5.24) \quad I_C = -I_R + \alpha_F I_F$$

The Ebers-Moll parameters are related by the following equation:

$$(5.25) \quad I_{E,s} \alpha_F = I_{C,s} \alpha_R$$

This relation ship is also referred as the reciprocity relation and can be derived by examining the minority carrier current through the base. For the specific case where the base-emitter and base-collector voltage are the same, there can be no minority carrier diffusion in the base so that:

$$(5.26) \quad I_F(V_{BE}) \alpha_F = I_R(V_{BC} = V_{BE}) \alpha_R$$

from which the reciprocity relation is obtained.

The forward- and reverse-bias transport factors are obtained by measuring the current gain in the forward active and reverse active mode of operation. The saturation currents  $I_{E,s}$  and  $I_{C,s}$  are obtained by measuring the base-emitter (base-collector) diode saturation current while shorting the base-collector (base-emitter) diode.

### 5.3.4. Saturation.

In the low resistance “on” state of a bipolar transistor, one finds that the voltage between the collector and emitter is less than the forward bias voltage of the base-emitter junction. Typically the “on” state voltage of a silicon BJT is 100 mV and the forward bias voltage is 700 mV. Therefore, the base-collector is forward biased. Using the Ebers-Moll model, we can calculate the “on” voltage from:

$$(5.27) \quad V_{CE,sat} = V_{BE} - V_{BC} = V_t \ln \left\{ \frac{I_F I_{C,s}}{I_R I_{E,s}} \right\}$$

and using equations (5.23) and (5.24) and the reciprocity relation (5.25), one obtains:

$$(5.28) \quad V_{CE,sat} = V_t \ln \left\{ \frac{1 + \frac{I_C}{I_B} (1 - \alpha_R)}{\alpha_R \left[ 1 - \frac{I_C}{I_B} \frac{(1 - \alpha_F)}{\alpha_F} \right]} \right\} \quad \text{Derivation 5.1}$$

Saturation also implies that a large amount of minority carrier charge is accumulated in the base region. As a transistor is switched from saturation to cut-off, this charge initially remains in the base and a collector current will flow until this charge is removed by recombination. This causes an additional delay before the transistor is turned off. Since the carrier lifetime can be significantly longer than the base transit time, the turn-off delay causes a large and undesirable asymmetry between turn-on and turn-off time. Saturation is therefore avoided in high-speed bipolar logic circuits. Two techniques are used: 1) adding a Schottky diode in parallel to the base-emitter junction and 2) using an emitter-coupled circuit configuration.