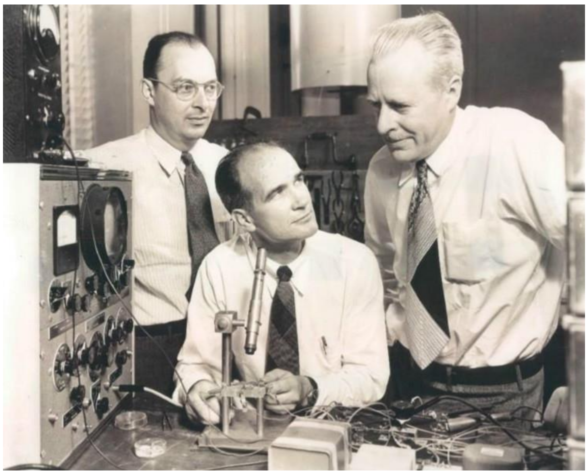
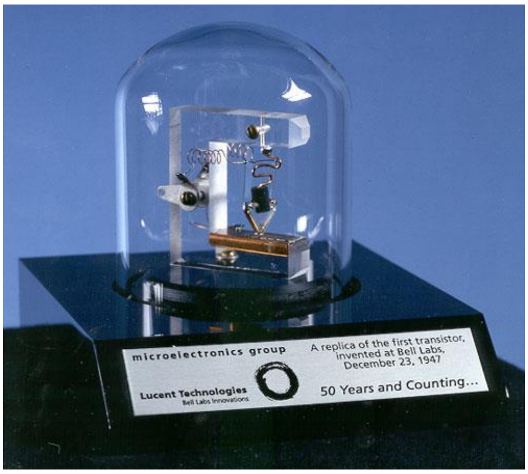
3. The Bipolar Junction Transistor

The bipolar point-contact transistor was invented in 1947 at the Bell Telephone Laboratories (USA) by John Bardeen and Walter Brattain under the direction of William Shockley. The junction version known as the bipolar junction transistor (BJT), invented by Shockley in 1948, is the version we are going to study hereafter. In acknowledgement of this accomplishment, Shockley, Bardeen, and Brattain were jointly awarded the 1956 Nobel Prize in Physics "for their researches on semiconductors and their discovery of the transistor effect."



John Bardeen, William Shockley and Walter Brattain at Bell Labs, 1948



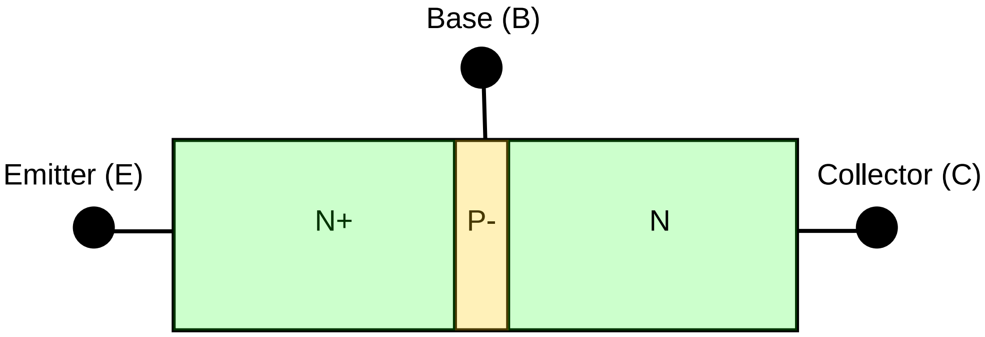
A replica of the first working transistor

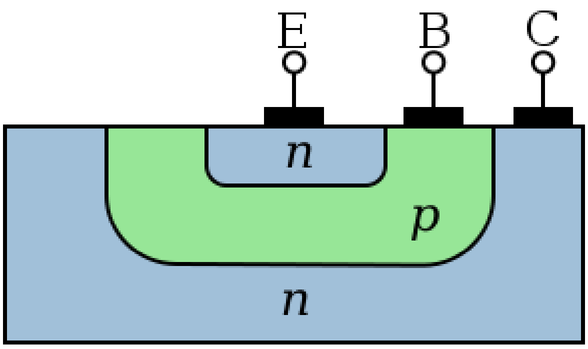
The transistor (in its various forms, not only BJT) is the key active component in practically all modern electronics. Many consider it to be one of the greatest inventions of the 20th century. Its importance in today's society rests on its ability to be mass produced using a highly automated process (semiconductor device fabrication) that achieves astonishingly low per-transistor costs.

Although several companies each produce over a billion individually packaged (known as discrete) transistors every year, the vast majority of transistors now are produced in integrated circuits, along with diodes, resistors, capacitors and other electronic components, to produce complete electronic circuits.

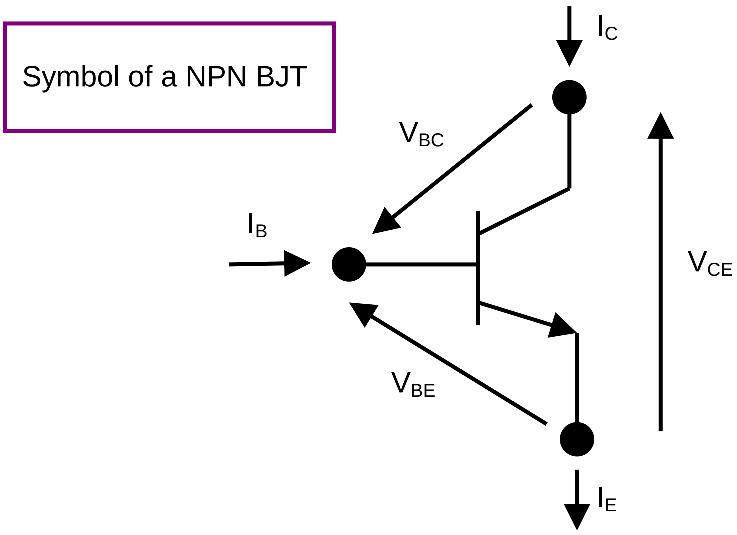
- **Basic Operation**

A NPN BJT is a semiconductor device consisting of a narrow P-type region between two N-type regions. The three regions are called the *emitter* (E), *base* (B), and *collector* (C), respectively. The emitter region is heavily doped with the appropriate impurity, while the base region is very lightly doped. The collector region has a moderate doping level. Note that the structure is not symmetrical.





Simplified cross section of a planar NPN bipolar junction transistor



Symbol of a NPN BJT

We consider throughout this chapter a device consisting of N, P, and N regions in order, but we can also build equivalent devices in P, N, and P order instead. In fact, it is sometimes useful to have both types of devices available.

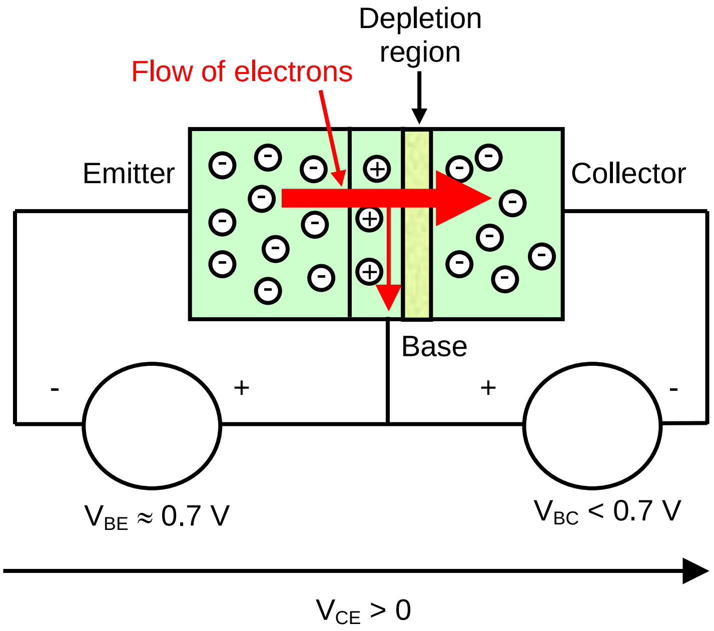
Let us see what happens when bias voltages are applied to such device. Let us assume the use of a silicon BJT.

Consider first that a forward bias is applied to the base-emitter junction and a reverse bias is applied to the base-collector junction. These are the normal operating conditions of a bipolar junction transistor. These conditions imply that $V_{BE} \approx 0.7$ V and $V_{BC} < 0.7$ V. If we take the emitter as a reference, these conditions can be re-written as $V_{BE} \approx 0.7$ volt and $V_{CE} = V_{CB} + V_{BE} = V_{BE} - V_{BC} > 0$ V.

Since we already know how a PN junction operates, we would expect to have electrons move from emitter to base and leave the device through the base at that point. With the collector junction reverse biased, we would expect no current to flow through that junction.

But something happens inside the base region. The forward bias on the base-emitter junction does indeed attract electrons from the emitter into the base. As the base is very thin, electrons entering the base find themselves close to the depletion region formed by the reverse bias of the base-collector junction.

While the reverse-bias voltage acts as a barrier to holes in the base, it actively propels electrons across it. Thus, any electrons entering the junction area are swept across the depletion area into the collector and give rise to a collector current.



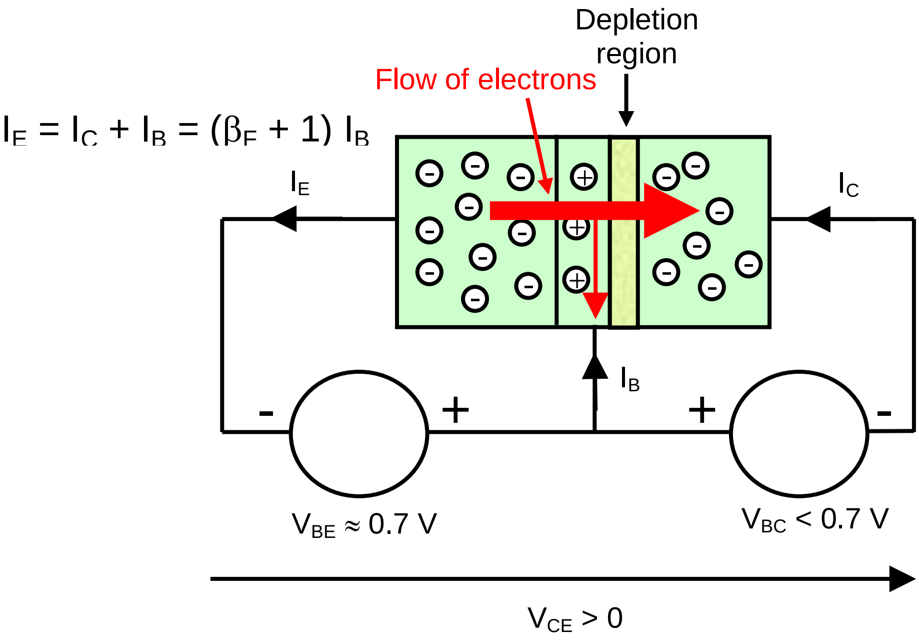
Careful design ensures that the majority of the electrons entering the base are swept across the base-collector junction into the collector.

Thus the flow of electrons from emitter to collector is many times greater than the flow from emitter to the base. In fact, the collector current I_C is proportional to the base current I_B:

$$I_C = \beta_F I_B ,$$

where β_F is a constant that can take its value in the range from approximately 50 to 300 for typical bipolar technologies.

This current amplification phenomenon is known as the transistor effect.



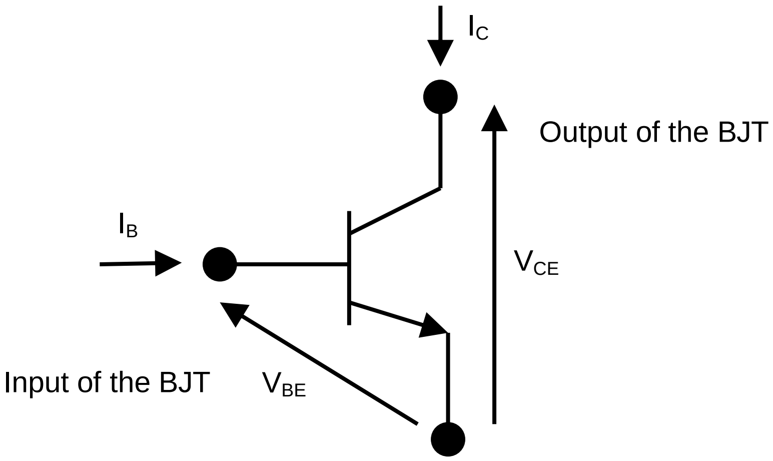
As previously mentioned, it is also possible to build a transistor with the region types reversed (PNP structure). In this case, holes will be drawn from the emitter into the base region by the forward bias, and will then be pulled into the collector region by the higher negative bias. Otherwise, this device works the same way and has the same general properties as the one described above. To distinguish between the two types of transistors, we refer to them by the order in which the different regions appear. Thus, this is a PNP transistor while the device described above is an NPN transistor.

However, PNP transistors often have lower β_F values and are slower (i.e., operate at lower frequencies) than their NPN counterparts.

• Common-Emitter Configuration of a BJT

The BJT is viewed as a semiconductor device with an input and an output. Usually, the input parameters are the current I_B and the voltage V_{BE}, whereas the output parameters are the current I_C and the voltage V_{CE}. This particular arrangement is referred to as common-emitter configuration because the emitter terminal is common to both input and output.

Note that common-collector and common-base configurations are also sometimes considered.

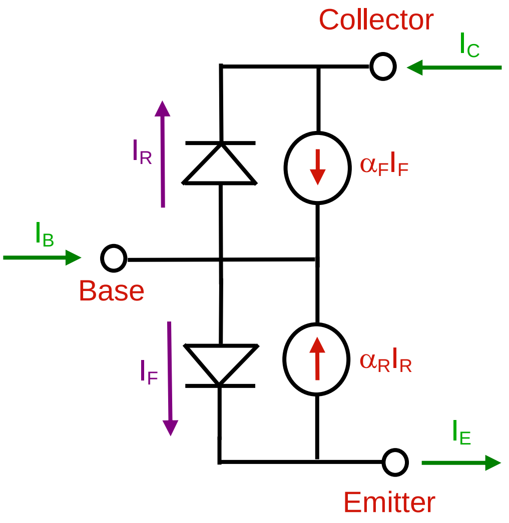


- **The Four Different Modes of Operation of a Transistor**

In 1954, Jewell Ebers and John Moll introduced their (static) model of a BJT. The Ebers-Moll model is depicted below.

$$I_F = I_S\left(\exp\left\{\frac{V_{BE}}{V_T}\right\} - 1\right)$$

$$I_R = I_S\left(\exp\left\{\frac{V_{BC}}{V_T}\right\} - 1\right)$$



In this model, α_F is the forward common-base current gain (typically ranging from 0.98 to 0.998 for most BJT technologies, i.e. α_F slightly smaller than the unit), and α_R is the reverse common-base current gain (typically, $\alpha_R \approx 0.5$).

We can write the general equations for the Ebers-Moll model:

Base current: $I_B = I_F + I_R - \alpha_F I_F - \alpha_R I_R = (1-\alpha_F)I_F + (1-\alpha_R)I_R$

Collector current: $I_C = \alpha_F I_F - I_R$

Emitter current: $I_E = I_F - \alpha_R I_R = I_B + I_C$

The physical phenomena behind the Ebers-Moll model are rather simple to understand:

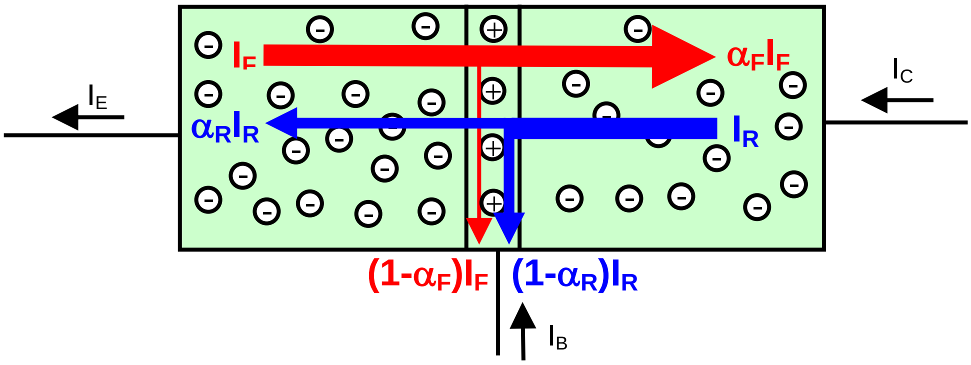
1. Both diodes represent the base-emitter and base-collector PN junctions.

2. The parameter α_F represents the proportion of electrons coming from the emitter that are able to reach the collector. The fact that α_F is very close to the unit implies that the majority of electrons coming from the emitter do reach the collector, while the remaining electrons leave the device through the base.

3. The parameter α_R represents the proportion of electrons coming from the collector that are able to reach the emitter. The fact that the value of α_R is (typically) approximately equal to 0.5 means that roughly half of the electrons coming from the collector end up leaving the transistor through the emitter.

The difference in values between α_F and α_R is due to the inherent non-symmetrical physical structure of a BJT.

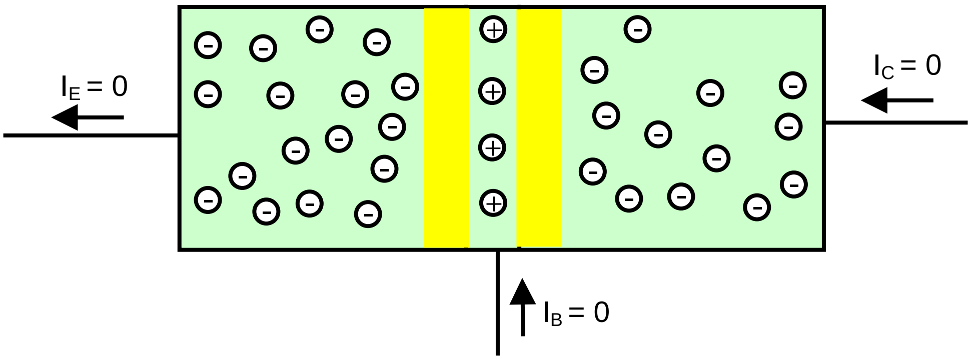
Arrows showing the flows of electrons



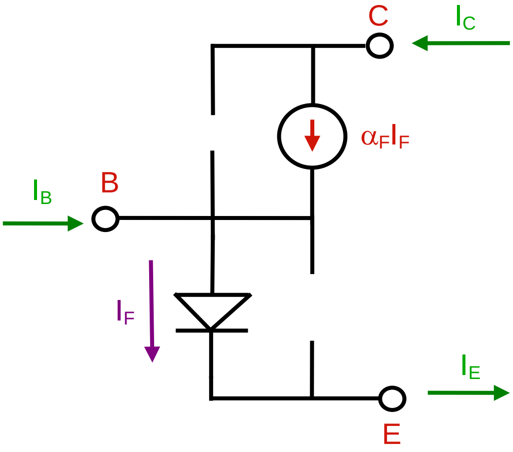
A BJT has four modes of operation.

- *First mode of operation*: The transistor is in the cut-off mode when $V_{BE} < 0.7$ V and $V_{BC} < 0.7$ V. In such case, we have $I_F = I_R = 0$, which leads to $I_B = I_C = I_E = 0$.

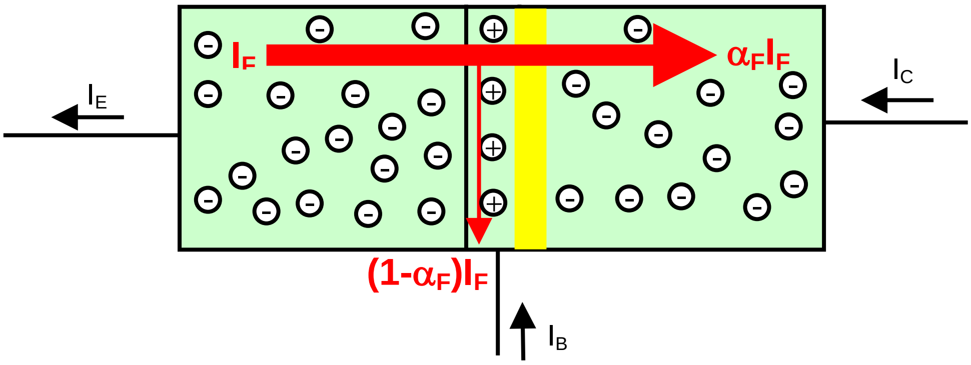
Two depletion regions



<span>-</span> *Second mode of operation*: The transistor is in the forward active mode when $V_{BE} \approx 0.7$ V and $V_{BC} < 0.7$ V (thus implying $V_{CE} > 0$).



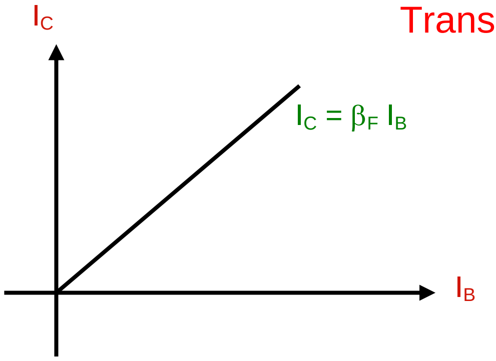
Arrows showing the flows of electrons



In the forward active mode, we have $I_R = 0$, which yields $I_B = (1-\alpha_F) I_F$, $I_C = \alpha_F I_F$, and $I_E = I_F$. By combining those three expressions, we obtain $I_C = \alpha_F I_F = \frac{\alpha_F}{1-\alpha_F} I_B = \beta_F I_B$.

The parameter β_F is known as the forward current gain. If we take $0.98 < \alpha_F < 0.998$, we have $49 < \beta_F < 499$. Hereafter, we will adopt the value $\beta_F = 100$.

In the forward active mode, the collector current I_C is proportional to the base current I_B



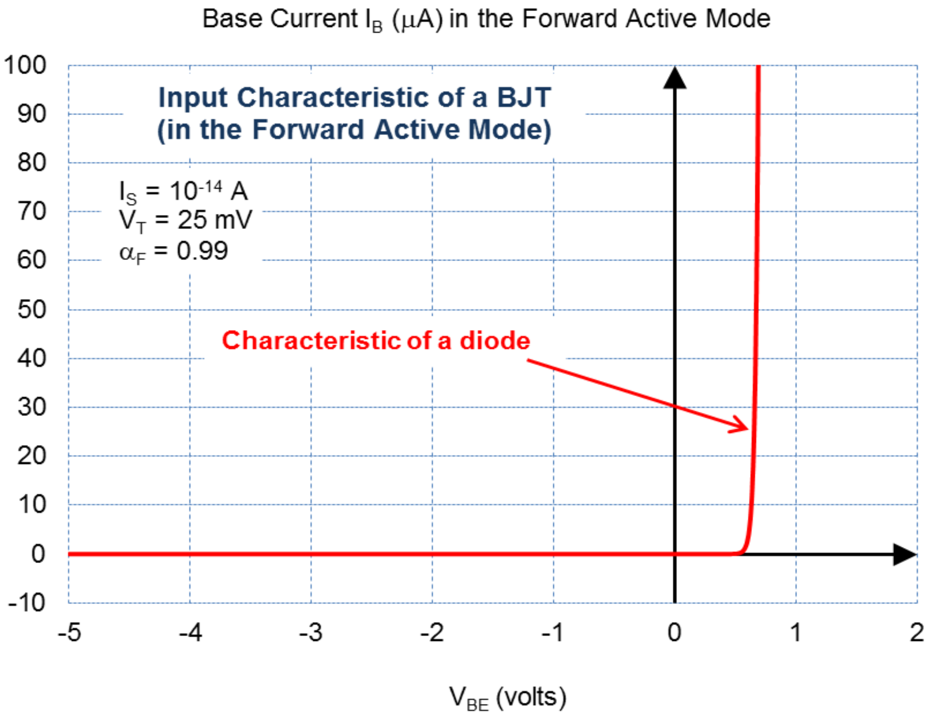
Transfer characteristic of a BJT

We can also notice that the emitter current is given by $I_E = I_F \approx I_S \exp\left(\frac{V_{BE}}{V_T}\right) = I_C + I_B = (\beta_F + 1) I_B$.

This result indicates that the base current varies exponentially with the voltage V_{BE}:

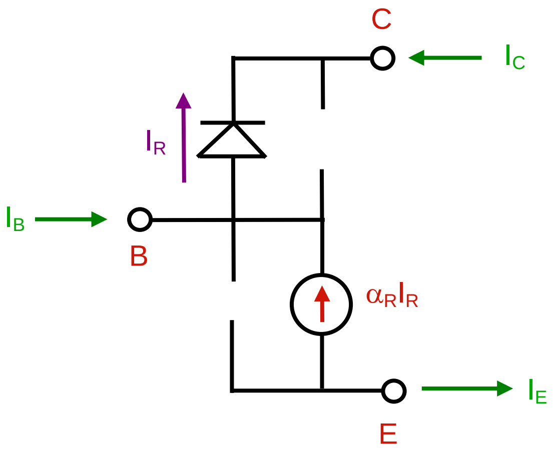
$$I_B = \frac{I_E}{\beta_F + 1} \approx \frac{I_S}{\beta_F + 1} \exp\left(\frac{V_{BE}}{V_T}\right) = I_S{}' \exp\left(\frac{V_{BE}}{V_T}\right).$$

This equation linking I_B and V_{BE} provides us with the input characteristic of a BJT in the forward active mode. In fact, the equation corresponds to that of a diode as if the current I_B was the current flowing through the base-emitter junction.

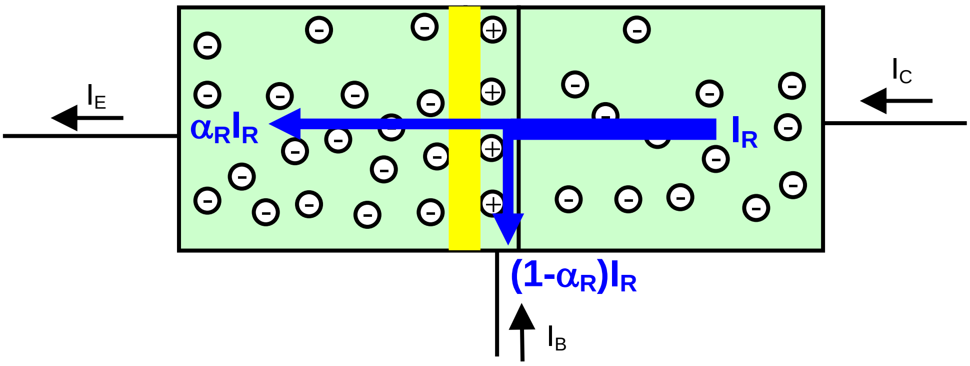


The forward active mode is the mode used for designing amplifiers in analogue electronics.

- *Third mode of operation*: The transistor is in the reverse active mode when $V_{BE} < 0.7$ V and $V_{BC} \approx 0.7$ V (thus implying $V_{CE} = V_{CB} + V_{BE} = V_{BE} - V_{BC} < 0$).



Arrows showing the flows of electrons

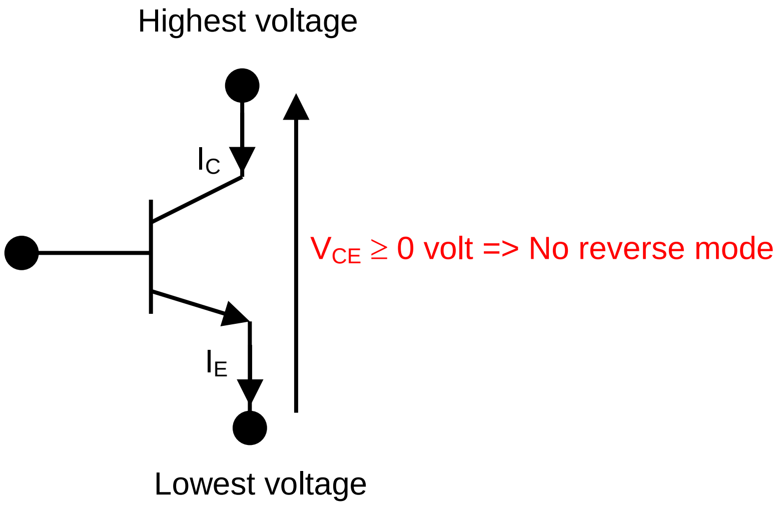


In the reverse active mode, we have $I_F = 0$, which yields $I_B = (1-\alpha_R)I_R$, $I_C = -I_R$, and $I_E = -\alpha_R I_R = I_B + I_C$. By combining these three expressions, we obtain $I_C = -I_R = -\frac{I_B}{1-\alpha_R}$ and $I_E = -\alpha_R I_R = -\frac{\alpha_R}{1-\alpha_R} I_B = -\beta_R I_B$.

The parameter β_R is known as the reverse current gain. If we take $\alpha_R \approx 0.5$, we have $\beta_R \approx 1$.

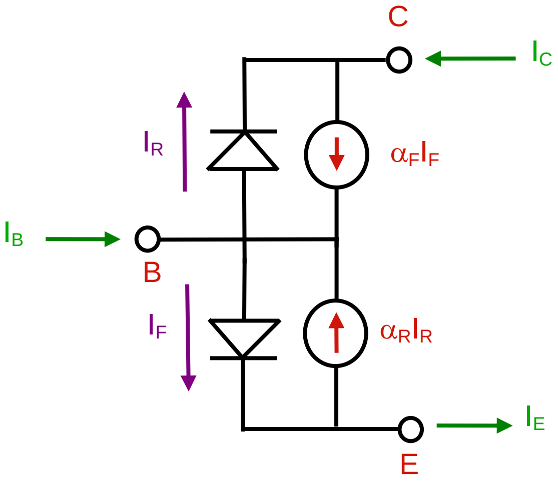
Since $\beta_R << \beta_F$, it is clear that the transistor effect obtained in the reverse active mode is much weaker that that achieved in the forward active mode. This is why the reverse active mode is of no particular interest in practice.

In EEE1002/EEE1010, we will never have to consider the reverse mode of operation in any of our circuits because the collector will always be on the side of the highest voltage whereas the emitter will be on the side of the lowest voltage, thus implying that we will always have $V_{CE} \geq 0$ volt (which contradicts the condition required for the reverse mode of operation).

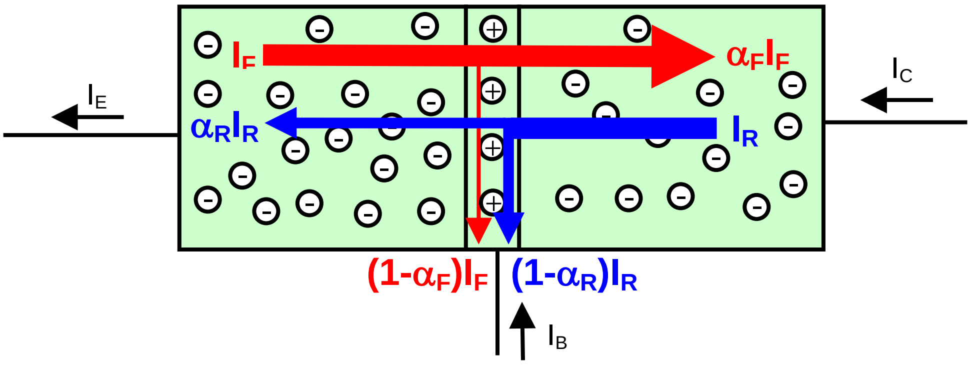


- *Fourth mode of operation*: The transistor is in the saturation mode of operation when $V_{BE} \approx 0.7$ V and $V_{BC} \approx 0.7$ V (thus implying $V_{CE} \approx 0$).

In the saturation mode, the expressions for the base, collector, and emitter currents are complicated and, in fact, not very interesting. The most important thing to remember is that $V_{CE} \approx 0$ volt in this mode of operation.



Arrows showing the flows of electrons



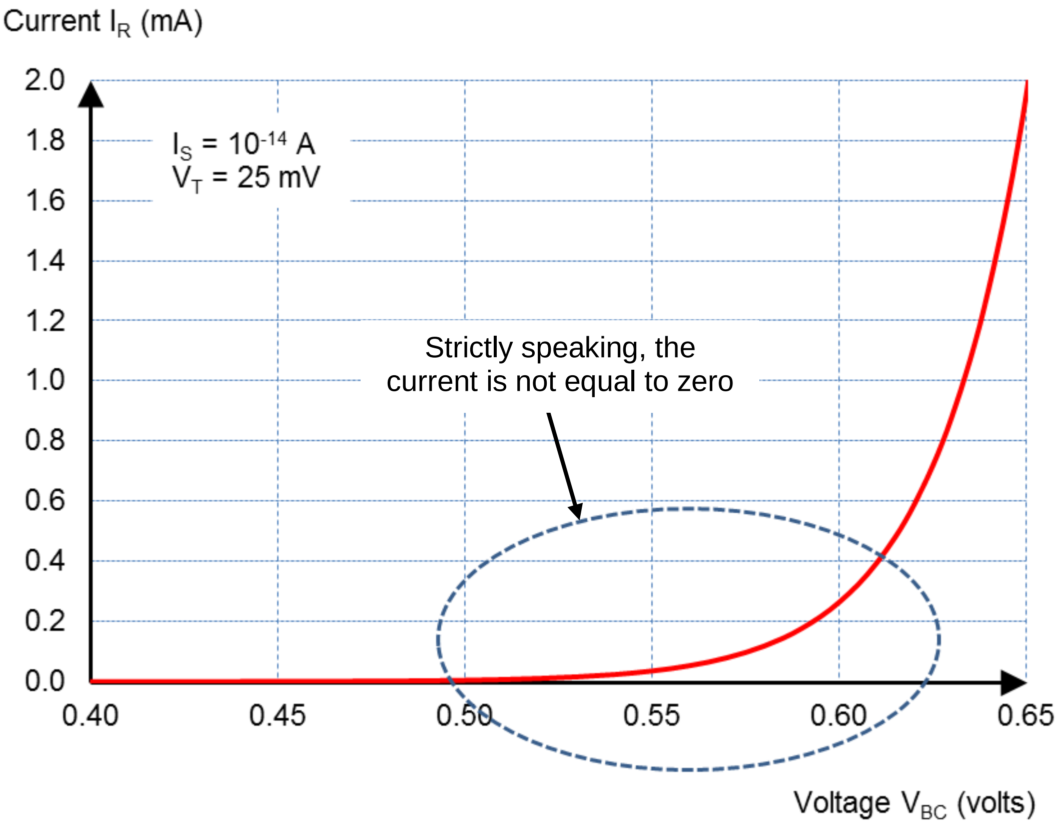
We now know enough to be able to understand the way the collector current I_C varies with the voltage V_{CE} when the BJT is either in the saturation mode or in the forward active mode, i.e. when $V_{BE} \approx 0.7$ volt.

When V_{CE} is close to zero, the BJT is in the saturation mode. In this case, we could show that a small increase in V_{CE} above 0 volt results in a very large increase in the collector current. This was not demonstrated earlier as it is of little interest to us.

In practice, the saturation mode corresponds to any value of the voltage V_{CE} ranging from 0 to roughly 0.2 volts.

This value of 0.2 volt can be explained as follows: Strictly speaking, the condition for the saturation mode is that a current does flow through the base-collector junction (i.e., $I_R \neq 0$). In a practical PN junction, a forward bias ranging from approximately 0.5 to 0.7 volt is often sufficient for the existence of a non-negligible current (see figure below).

In other words, the condition $V_{BC} > 0.5$ V can be considered sufficient for the BJT to be in the saturation mode. It is thus reasonable to say that, for $V_{CE} = V_{BE} - V_{BC} < 0.2$ V, the BJT is actually in the saturation mode.

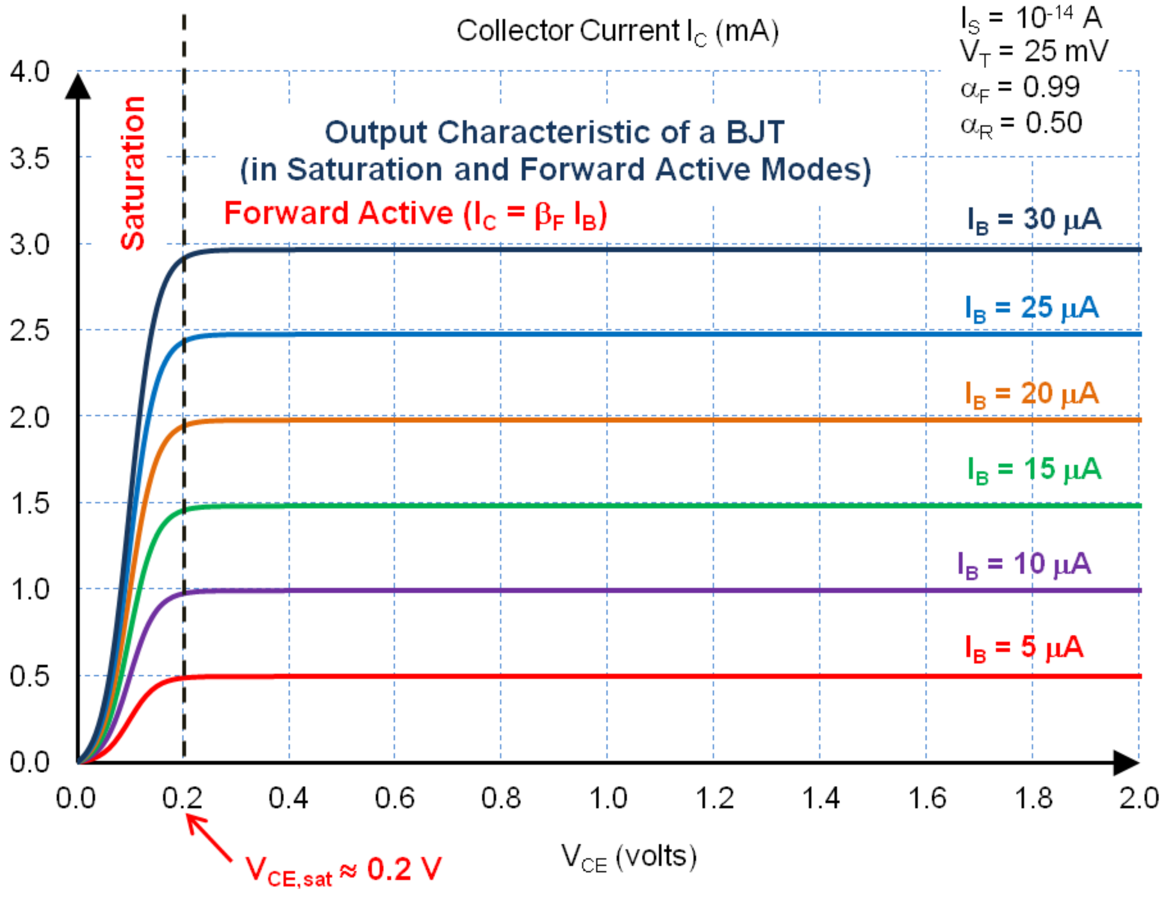


Once V_{CE} is increased beyond $V_{CE} \approx 0.2$ volt, i.e. $V_{BC} < 0.5$ volt, the current I_R is completely negligible, and the BJT is clearly in the forward active mode.

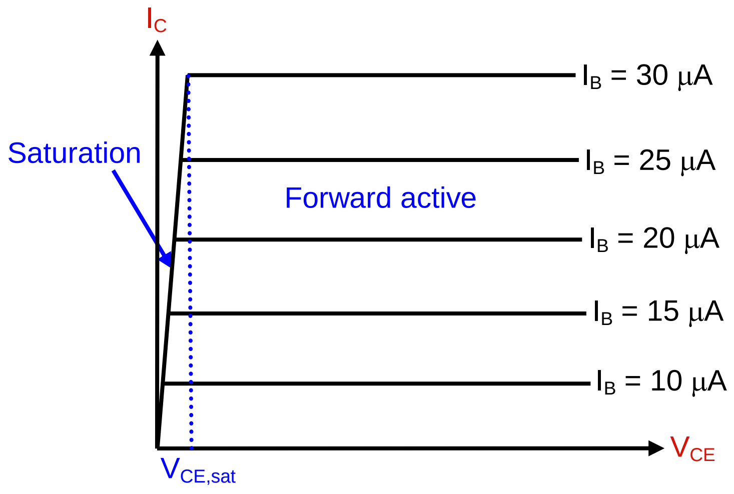
It may not always be easy to determine the exact value of the voltage V_{CE} in the saturation mode when performing the manual analysis of a circuit. However, we can make our life easier by simply assuming that, in the saturation mode, the voltage V_{CE} is a constant slightly greater than zero and called $V_{CE,sat}$.

Throughout these lecture notes, we will use the value $V_{CE,sat} \approx 0.2$ volt whenever the BJT is in the saturation mode of operation. This simplification does not result in any significant error as the actual value of V_{CE} always lies somewhere between 0 volt and $V_{CE,sat} \approx 0.2$ volt.

We finally obtain the output characteristic of a BJT which shows the variation of the collector current I_C (output current) as a function of the voltage V_{CE} (output voltage).



Output characteristic of a BJT


I_C
Saturation
Forward active
I_B = 30 µA
I_B = 25 µA
I_B = 20 µA
I_B = 15 µA
I_B = 10 µA
V_{CE}
$V_{CE,sat}$