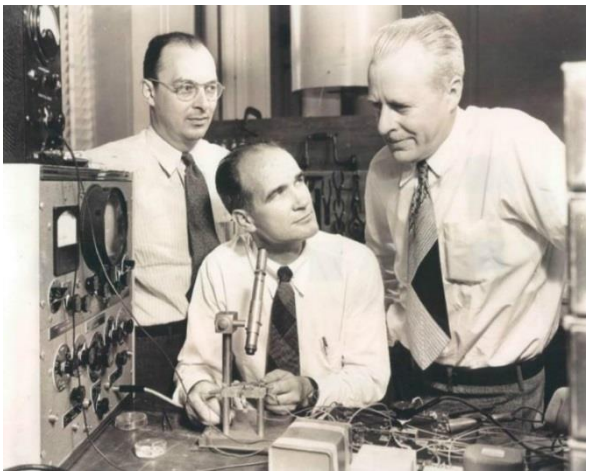
Part 3 – Bipolar Junction Transistors

The bipolar junction transistor (BJT) was invented in 1947 at the Bell Telephone Laboratories (USA) by John Bardeen, Walter Brattain, and William Shockley.

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."



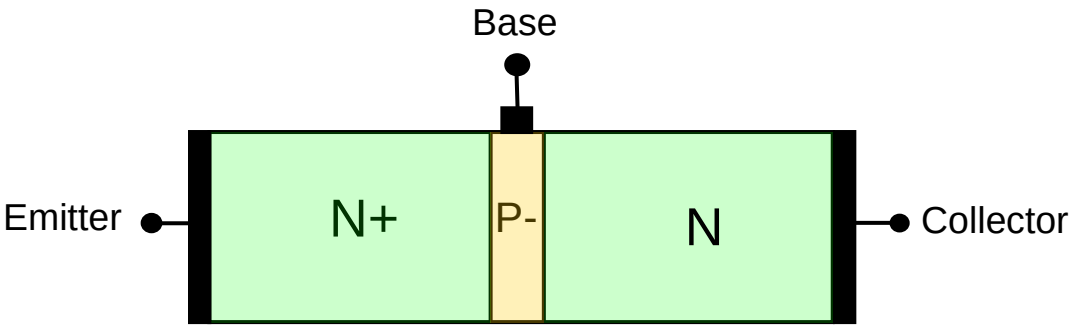
John Bardeen (1908 – 1991), William Bradford Shockley (1910 – 1989) and Walter Houser Brattain (1902 – 1987) at Bell Labs, 1948

The transistor, in its various forms, is the key active component in practically all modern electronics. Many consider it to be one of the greatest inventions of the 20th century.

Semiconductor companies manufacture billions of individual transistors every year. However, the vast majority of transistors are now produced in integrated circuits that also contain diodes, resistors, capacitors and other electronic components.

Basic Operation of a BJT

An NPN BJT is a semiconductor device consisting of a narrow P-type region, called the base, between two N-type regions, called the emitter and the collector.



The N-type emitter region is heavily doped (N+), whereas the P-type base region is lightly doped (P-). The N-type collector region has a moderate doping level (N). The whole structure is therefore not symmetrical.

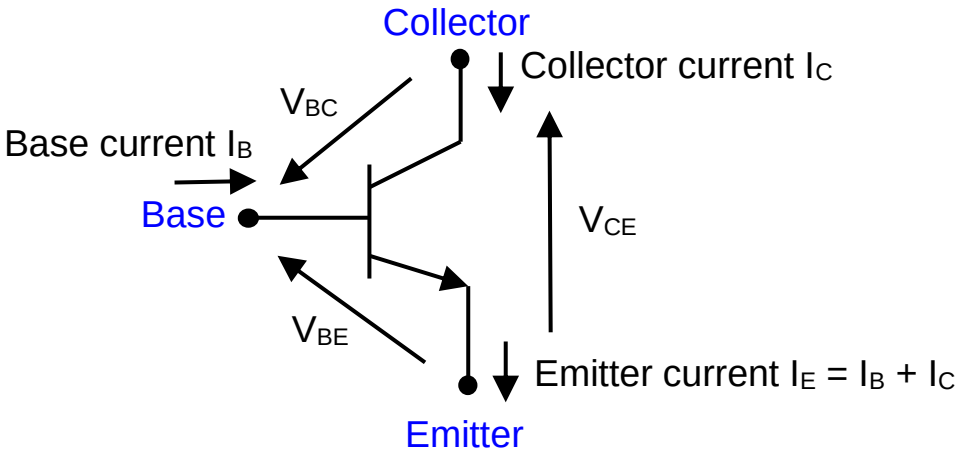
Electrons are the majority carriers in both emitter and collector. Holes are the majority carriers in the base.

We consider here a device consisting of N, P, and N regions in order, referred to as *NPN bipolar junction transistor*.

It is worth mentioning that we can also build equivalent devices in P, N, and P order instead, called *PNP bipolar junction transistors*.

In fact, it is sometimes useful to have both types of devices available in the same circuit.

Symbol of an NPN BJT



The figure above shows the symbol of an NPN BJT. We see that this device has indeed three terminals, namely the emitter, the base, and the collector.

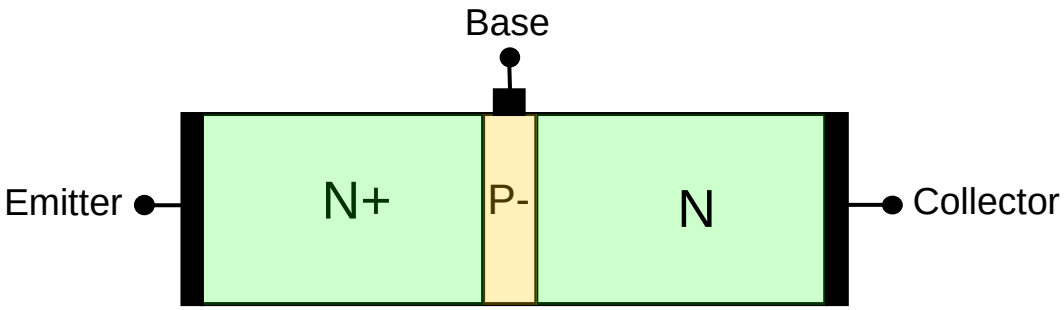
We can define three currents for this device:

- The collector current I_C that enters the transistor through the collector;
- The base current I_B that enters the transistor through the base;
- The emitter current I_E that leaves the transistor via the emitter.

We can always write $I_E = I_B + I_C$. This equation means that the current leaving a BJT, I_E, is equal to the sum of the two currents, I_B and I_C, entering it.

It is also possible to define three voltages for this device:

- The voltage, V_{BE}, between base and emitter;
- The voltage, V_{BC}, between base and collector;
- The voltage, V_{CE}, between collector and emitter.



The structure of an NPN BJT shows the existence of two diodes: the base-emitter PN junction and the base-collector PN junction.

Since each diode can be either on or off, i.e. forward-biased or reverse-biased, there are clearly four possibilities.

In other words, we can expect the BJT to have four modes of operation:
(1) Cut-off mode when both BE and BC junctions are off.
(2) Forward active mode when the BE junction is on and the BC junction is off;
(3) Reverse active mode when the BE junction is off and the BC junction is on;
(4) Saturation mode when both BE and BC junctions are on.

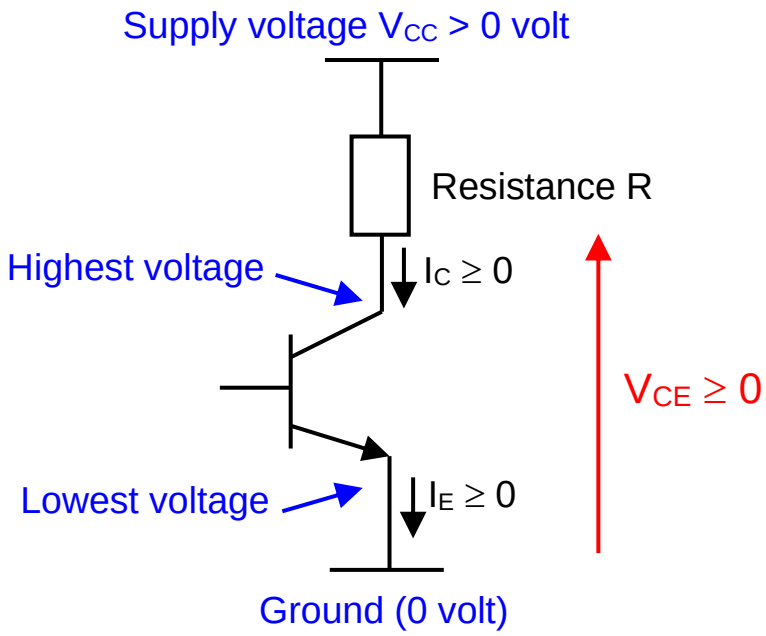
Before we start analysing these modes of operation, we are going to make an initial assumption: the voltage V_{CE} can never be negative. In other words, V_{CE} is greater than or equal to zero.

This assumption is important because it is going to simplify things tremendously.

In practical BJT circuits, this assumption is (almost) always valid.

In fact, it is very simple to make sure that $V_{CE} \geq 0$ by connecting the collector to the highest voltage in the circuit, either directly or through a resistance. The highest voltage is the positive DC supply voltage.

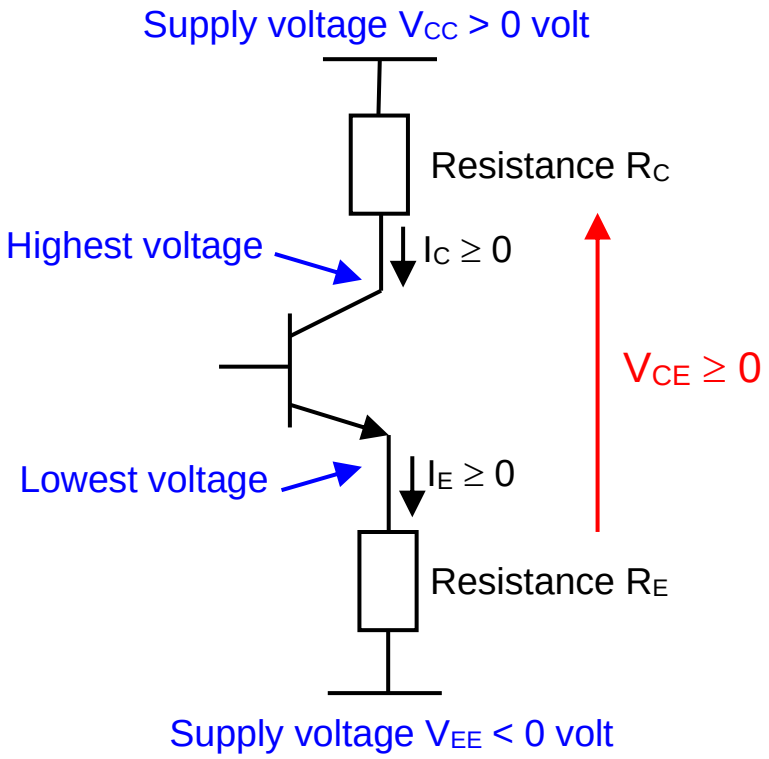
A typical configuration is shown in the figure below.



We can see that the collector is connected to the positive supply voltage, called V_{CC}, through a resistance R, whereas the emitter is directly connected to ground, i.e. the reference voltage defined as being equal to zero volt.

It is easy to show that, with such configuration, the voltage, V_{CE}, between collector and emitter can never be negative, i.e. we must have $V_{CE} \geq 0$.

Another typical configuration using two supply voltages is shown below.



Since the voltage V_{CE} can be written as $V_{CE} = V_{CB} + V_{BE} = V_{BE} - V_{BC}$, the assumption $V_{CE} \geq 0$ volt implies that $V_{BE} - V_{BC} \geq 0$ volt, i.e. $V_{BE} \geq V_{BC}$. This inequality must be remembered as it will be often used later on.

1st mode of operation: Cut-off mode, when both BE and BC diodes are off.

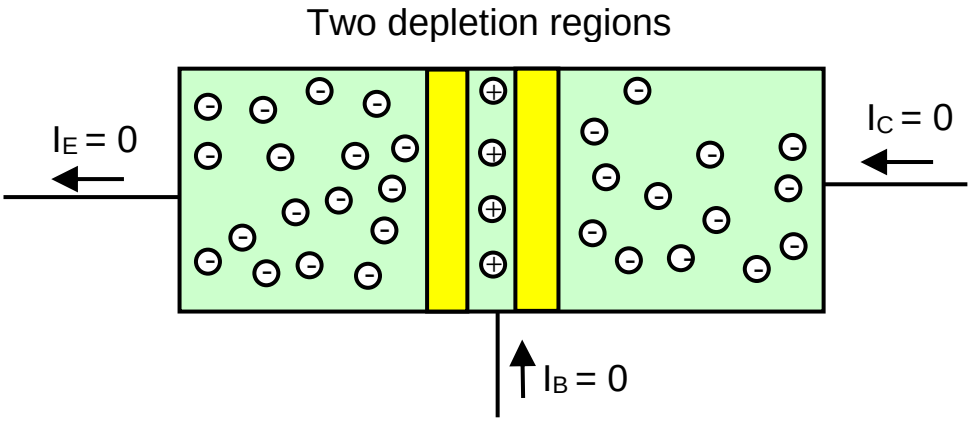
A BJT is in the cut-off mode when the voltage V_{BE} is less than the threshold voltage of the base-emitter junction and the voltage V_{BC} is less than the threshold voltage of the base-collector junction. It is written as $V_{BE} < V_{BE,on}$ and $V_{BC} < V_{BE,on}$, where $V_{BE,on}$ is the threshold voltage of a PN junction.

Obviously, since the base-emitter and base-collector junctions have the same threshold voltage, it would not be a good idea to use two different notations

$V_{BE,on}$ and $V_{BC,on}$ for this threshold voltage. We can adopt the same notation, $V_{BE,on}$, in order to avoid unnecessary complications.

Note that, in the context of BJTs, we do not use the notation V_d for the threshold voltage of a PN junction. Instead we replace it with the notation $V_{BE,on}$. The quantity $V_{BE,on}$ clearly represents the voltage required to turn on the base-emitter and base-collector junctions.

For a silicon BJT, we have $V_{BE,on} \sim 0.7$ volt.



The cut-off mode corresponds to the case where both base-emitter and base-collector junctions are off at the same time. In other words, the base is isolated from both emitter and collector due to the existence of two depletion regions, shown in yellow in the figure above.

No current can therefore flow through the device and, consequently, the three currents I_B, I_C, and I_E are all equal to zero.

At this stage, it is useful to remember our initial assumption, $V_{CE} \geq 0$, which was shown to be strictly equivalent to $V_{BE} \geq V_{BC}$.

By assuming that $V_{BE} \geq V_{BC}$, the two conditions $V_{BE} < V_{BE,on}$ and $V_{BC} < V_{BE,on}$ can actually be combined into only one condition: $V_{BE} < V_{BE,on}$.

Summary: If $V_{BE} < V_{BE,on}$, where $V_{BE,on}$ denotes the threshold voltage of a PN junction, the BJT is in the cut-off mode of operation, and we then have $I_B = 0$, $I_C = 0$, and $I_E = 0$.

2nd mode of operation: Forward-active mode, when the BE diode is on and the BC diode is off.

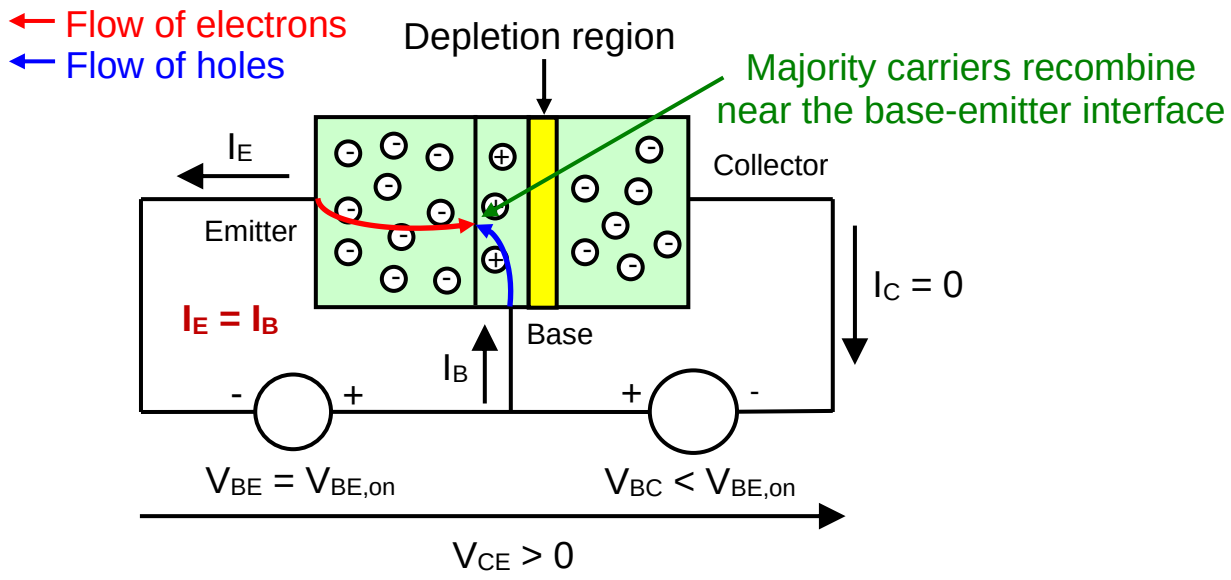
A BJT is said to be in the forward active mode if $V_{BE} = V_{BE,on}$ and $V_{BC} < V_{BE,on}$, where $V_{BE,on}$ denotes the threshold voltage of a diode.

By using the emitter as a reference, these two conditions can be re-written as $V_{BE} = V_{BE,on}$ and $V_{CE} > 0$ volt as $V_{CE} = V_{CB} + V_{BE} = V_{BE} - V_{BC}$.

Note that we like to use the emitter as a reference for voltages. This is because the emitter is often connected to ground, either directly or through a resistance, and is thus an ideal reference point in a circuit.

That explains why, when analysing BJT circuits, electronic engineers generally favour the use of the two voltages V_{BE} and V_{CE} and try to avoid as much as possible the use of the voltage V_{CB}.

What to expect?



In the forward-active mode, we would, at first glance, expect to have electrons move from the emitter to the base-emitter interface, holes move from the base to the base-emitter interface, and the two types of carriers constantly recombining in the vicinity of this interface. There would therefore be a current > 0 flowing though the base-emitter junction.

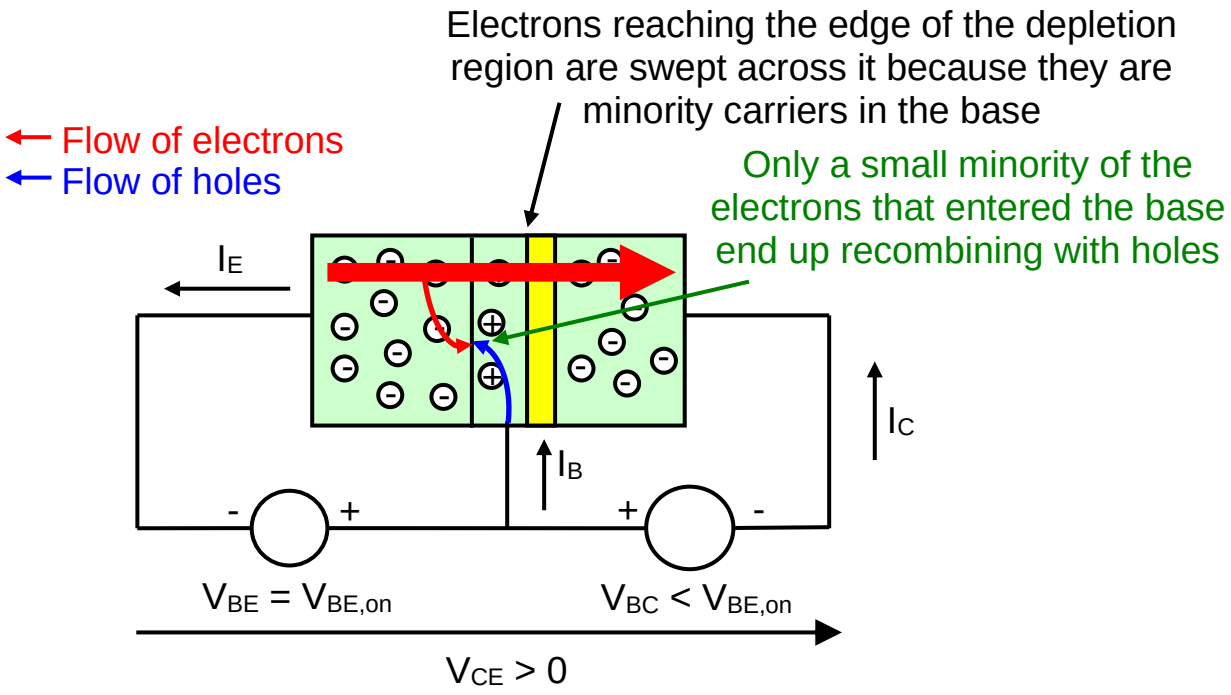
With the BC junction reverse biased, we would expect no current to flow through that junction.

As a result, we would have $I_B = I_E$ and $I_C = 0$.

But this is NOT what happens.

The forward bias on the base-emitter junction does indeed attract electrons from the emitter into the base. Once they are in the base, these electrons become

minority carriers and are expected to quickly recombine with holes which are the majority carriers in the base.



However, this does not happen because the base region is so thin that most minority carriers (electrons) can travel across the base region without ever recombining with majority carriers (holes).

In addition, the light P-type doping of the base region also ensures that the probability of recombination of electron-hole pairs is kept to a minimum in that region.

Therefore, most electrons injected into the base from the emitter are able to travel through the base region and reach 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 coming close to the depletion region are swept across it into the collector and give rise to a collector current.

To understand this, we must remember what happens with a diode under reverse bias: the depletion region cannot be crossed by the majority carriers on both sides of a PN junction, but it can definitely be crossed by the minority carriers. This is something that we already mentioned in the chapter on PN junctions.

In a simple diode, the current due to minority carriers under reverse bias is negligible because there are so few of them on each side of the junction.

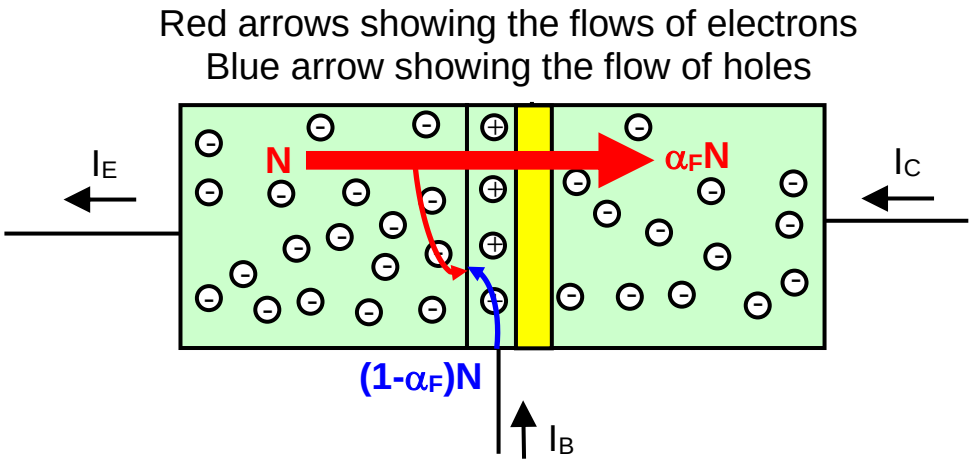
But, in a BJT, there are plenty of minority carriers (electrons) injected in the base and most of them manage to reach the edge of the depletion region without recombining with majority carriers (holes) in the base. These minority carriers crossing the reverse-biased base-collector junction generate the collector current.

Remarkably, there is actually an expression linking the emitter and collector currents:

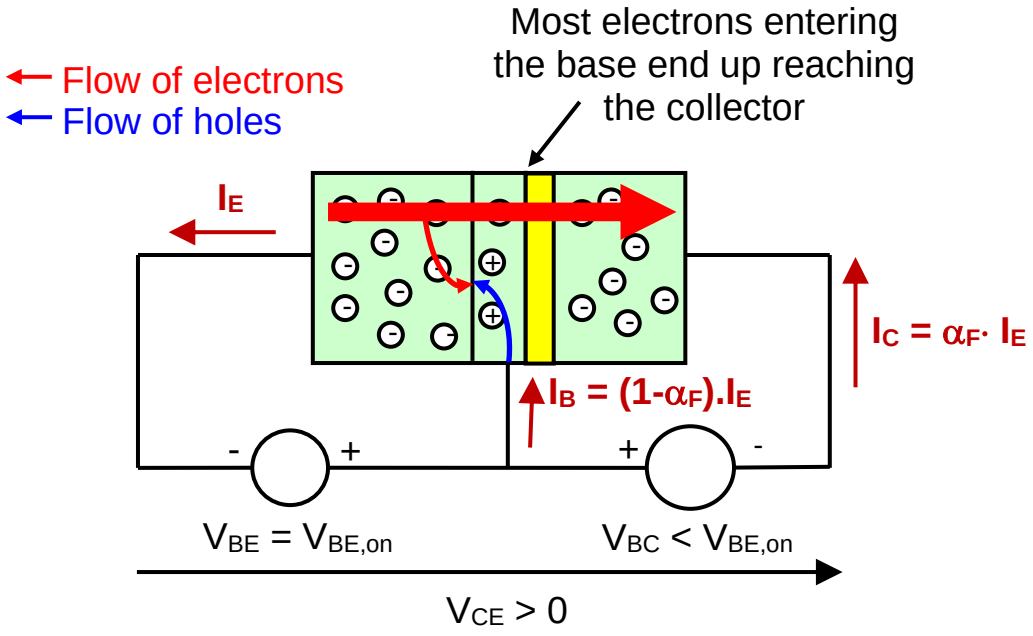
$$I_C = \alpha_F \cdot I_E,$$

where α_F is a constant parameter. In fact, the quantity α_F simply indicates the proportion of electrons coming from the emitter that are able to reach the collector. The value of α_F is slightly smaller than the unit for a well-designed BJT. We typically have $\alpha_F \sim 0.99$.

To better understand the operation of an NPN BJT, we can consider the following illustration.



Let us assume that N electrons enter the base from the emitter. The equation $I_C = \alpha_F \cdot I_E$ implies that $\alpha_F \cdot N$ electrons will then exit the device through the collector, while $(1-\alpha_F) \cdot N$ electrons will recombine with holes in the base. This means that $(1-\alpha_F) \cdot N$ holes must be brought into the base.



In terms of electric currents, this simple reasoning indicates that, if the expression of the collector current is $I_C = \alpha_F \cdot I_E$, then the expression of the base current must be $I_B = (1-\alpha_F) \cdot I_E$.

Let us assume, for instance, that α_F is equal to 0.99. In this case, we can write $I_C = 0.99 \times I_E$ and $I_B = 0.01 \times I_E$. These two equations imply that 99% of the electrons entering the emitter end up reaching the collector, whereas 1% of them end up recombining with holes during their journey through the base.

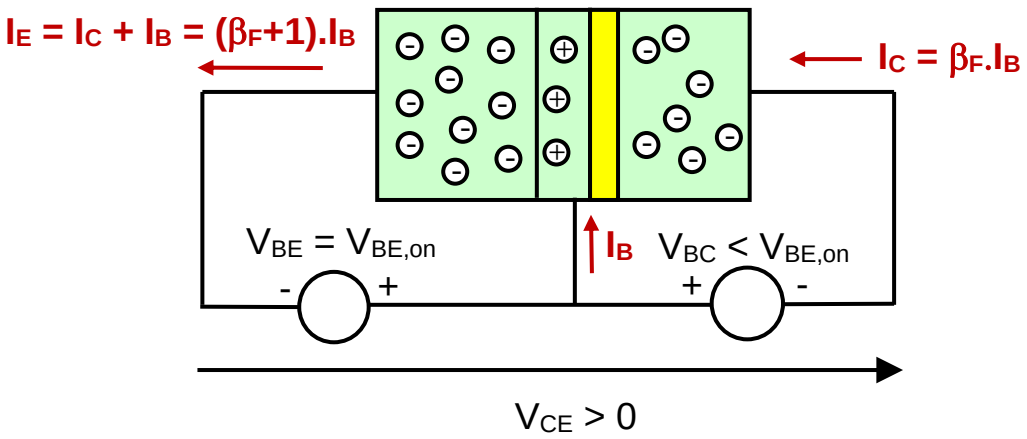
The fact that most of the electrons entering the emitter can exit the BJT through the collector is referred to as the *transistor effect*.

By combining the expressions $I_C = \alpha_F \cdot I_E$ and $I_B = (1 - \alpha_F) \cdot I_E$, we can easily express the collector current as a function of the base current as follows:

$$I_C = \frac{\alpha_F}{1-\alpha_F} \cdot I_B \Rightarrow I_C = \beta_F \cdot I_B,$$

where $\beta_F = \frac{\alpha_F}{1-\alpha_F}$ is a constant, called the *forward current gain*, that can take its value in the range from approximately 50 to 300 for typical bipolar technologies. If we assume $\alpha_F = 0.99$, we obtain

$$\beta_F = \frac{\alpha_F}{1-\alpha_F} = \frac{0.99}{1-0.99} = 99 \sim 100.$$



The expression $I_C = \beta_F \cdot I_B$ shows that the collector current I_C is proportional to the base current I_B, and also much larger than the latter. This is an important finding because it means that the BJT is able to perform significant current amplification.

Summary: If $V_{BE} = V_{BE,on}$ and $V_{CE} > 0$ volt, the BJT is in the forward active mode, and we can then write $I_C = \beta_F \cdot I_B$ and $I_E = I_C + I_B = (\beta_F+1) \cdot I_B$. Since $\beta_F >> 1$, the last equation implies that $I_E \sim \beta_F . I_B = I_C$.

In the forward active mode, the BJT behaves as a current amplifier in the sense that the collector current I_C is an amplified version of the base current I_B. This fascinating property is what has made the BJT so useful for many applications in the field of analogue electronics.

The forward active mode is the most important mode of operation for a BJT. In fact, the design of amplifiers in analogue electronics requires the use of BJTs operating in the forward active mode (at least some of the time) because the only mode of operation in which a BJT acts as a current amplifier is the forward active mode.

3rd mode of operation: Reverse-active mode, when the BE diode is off and the BC diode is on.

A BJT is said to be in the reverse active mode if $V_{BE} < V_{BE,on}$ and $V_{BC} = V_{BE,on}$, where $V_{BE,on}$ denotes the threshold voltage of a diode.

If we, once again, take the emitter as a reference, these two conditions can be re-written as $V_{BE} < V_{BE,on}$ and $V_{CE} < 0$ volt as $V_{CE} = V_{CB} + V_{BE} = V_{BE} - V_{BC}$.

The latter condition is in contradiction with the original assumption that $V_{CE} \geq 0$. Under this assumption, the BJT can never operate in the reverse-active mode. We can thus ignore this mode of operation.

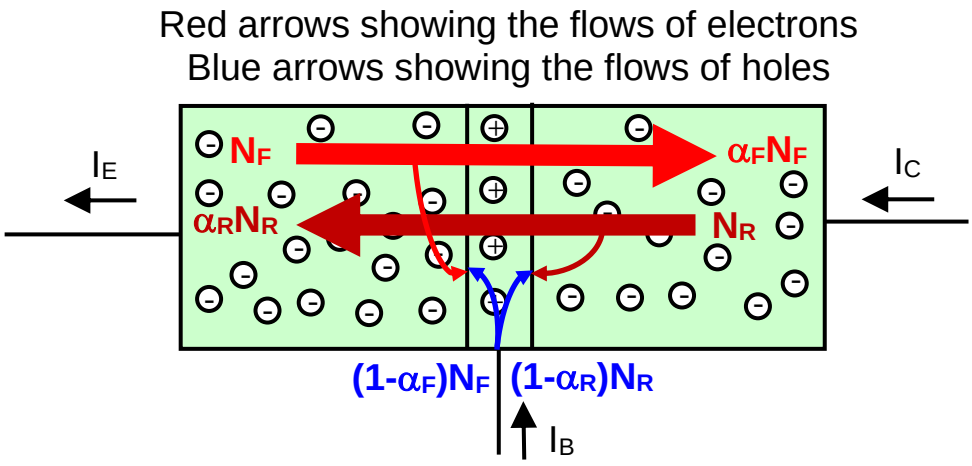
4th mode of operation: Saturation mode, when both BE and BC diodes are on.

A BJT is said to be in the saturation mode if $V_{BE} = V_{BE,on}$ and $V_{BC} = V_{BE,on}$, where $V_{BE,on}$ denotes the threshold voltage of a diode.

As both PN junctions are on, we can expect electrons flowing from emitter to collector, and also flowing from collector to emitter. There are therefore two transistor effects happening at the same time. This makes everything more complicated.

If we take the emitter as a reference, the two conditions $V_{BE} = V_{BE,on}$ and $V_{BC} = V_{BE,on}$ can be re-written as $V_{BE} = V_{BE,on}$ and $V_{CE} = 0$ volt as $V_{CE} = V_{CB} + V_{BE} = V_{BE} - V_{BC}$.

To try to understand what happens in the saturation mode of operation, let us consider the following example.



Assume that N_F electrons enter the base from the emitter, and N_R electrons enter the base from the collector. The letter F stands for “forward direction”, whereas the letter R stands for “reverse direction”.

As a result, $\alpha_F \cdot N_F$ electrons exit the device through the collector, while $(1-\alpha_F) \cdot N_F$ electrons recombine with holes in the base. This means that $(1-\alpha_F) \cdot N_F$ holes must be brought into the base. Remember that the constant parameter α_F represents the proportion of electrons coming from the emitter that are able to reach the collector.

At the same time, $\alpha_R \cdot N_R$ electrons also exit the device through the emitter, while $(1-\alpha_R) \cdot N_R$ electrons recombine with holes in the base. This means that $(1-\alpha_R) \cdot N_R$ holes must be brought into the base. Here, the constant parameter α_R represents the proportion of electrons coming from the collector that are able to reach the emitter.

As a summary, the net number of electrons entering the emitter is equal to $N_F - \alpha_R \cdot N_R$. The net number of electrons leaving the collector is $\alpha_F \cdot N_F - N_R$. Finally, the number of holes entering the base is $(1-\alpha_R) \cdot N_R + (1-\alpha_F) \cdot N_F$.

These numbers can be translated into current expressions easily, but we are not going to do it because this would lead us nowhere.

What we need to understand is that there are two transistor effects to consider in the saturation mode: a first one in the forward direction and a second one in the reverse direction.

Note that the transistor effect in the forward direction is in practice more “powerful” than that in the reverse direction because we always have $\alpha_F > \alpha_R$. Typically, we would have $\alpha_F \sim 0.99$ and $\alpha_R \sim 0.5$.

This discrepancy in the values of α_F and α_R should not come as a surprise because the BJT was never intended to be a symmetrical device. Remember that the N-type doping level is intentionally made higher in the emitter than in the collector. Since the concentration of free electrons is higher in the emitter than in the collector, it is much easier for electrons to diffuse in the forward direction than in the reverse direction.

So, the transistor effect is only supposed to work well in the forward direction, providing a large current gain in that direction only. As a matter of fact, the transistor effect in the reverse direction is more a parasitic effect than anything else.

We have several current contributions to take into account in order to find expressions of the various currents I_E, I_B, and I_C in the saturation mode. These expressions will be rather complicated and, in fact, not really helpful in practice. That is why there is no real need to derive the equations for the three currents.

Summary: A BJT operates in the saturation mode when $V_{BE} = V_{BE,on}$ and $V_{CE} = 0$ volt. That is only what we need to remember.

We have now completed the study of all modes of operation for a BJT, but before we bring everything together as a conclusion, we need to make a small correction.

So far, we have seen that the forward-active and saturation modes can be distinguished based on the value of the voltage V_{CE}:

- If $V_{BE} = V_{BE,on}$ and $V_{CE} > 0$, the BJT is forward active;
- If $V_{BE} = V_{BE,on}$ and $V_{CE} = 0$, the BJT is saturated.

Strictly speaking, the forward-active mode is defined as the mode for which the currents flowing in the reverse direction are negligible compared to the currents flowing in the forward direction. In practice, this is the case only if V_{BE} is slightly greater than V_{BC}, i.e. only if $V_{CE} = V_{BE} - V_{BC}$ is slightly greater than zero.

This is the reason why the statement “the BJT is forward active when $V_{BE} = V_{BE,on}$ and $V_{CE} > 0$” should in fact be replaced with the more accurate statement “the BJT is forward active when $V_{BE} = V_{BE,on}$ and $V_{CE} > V_{CE,sat}$”.

The constant quantity $V_{CE,sat}$, whose value must obviously be very close to zero volt, represents the voltage between collector and emitter at which the BJT enters saturation. Hereafter, we will use $V_{CE,sat}$ = 0.2 volt.

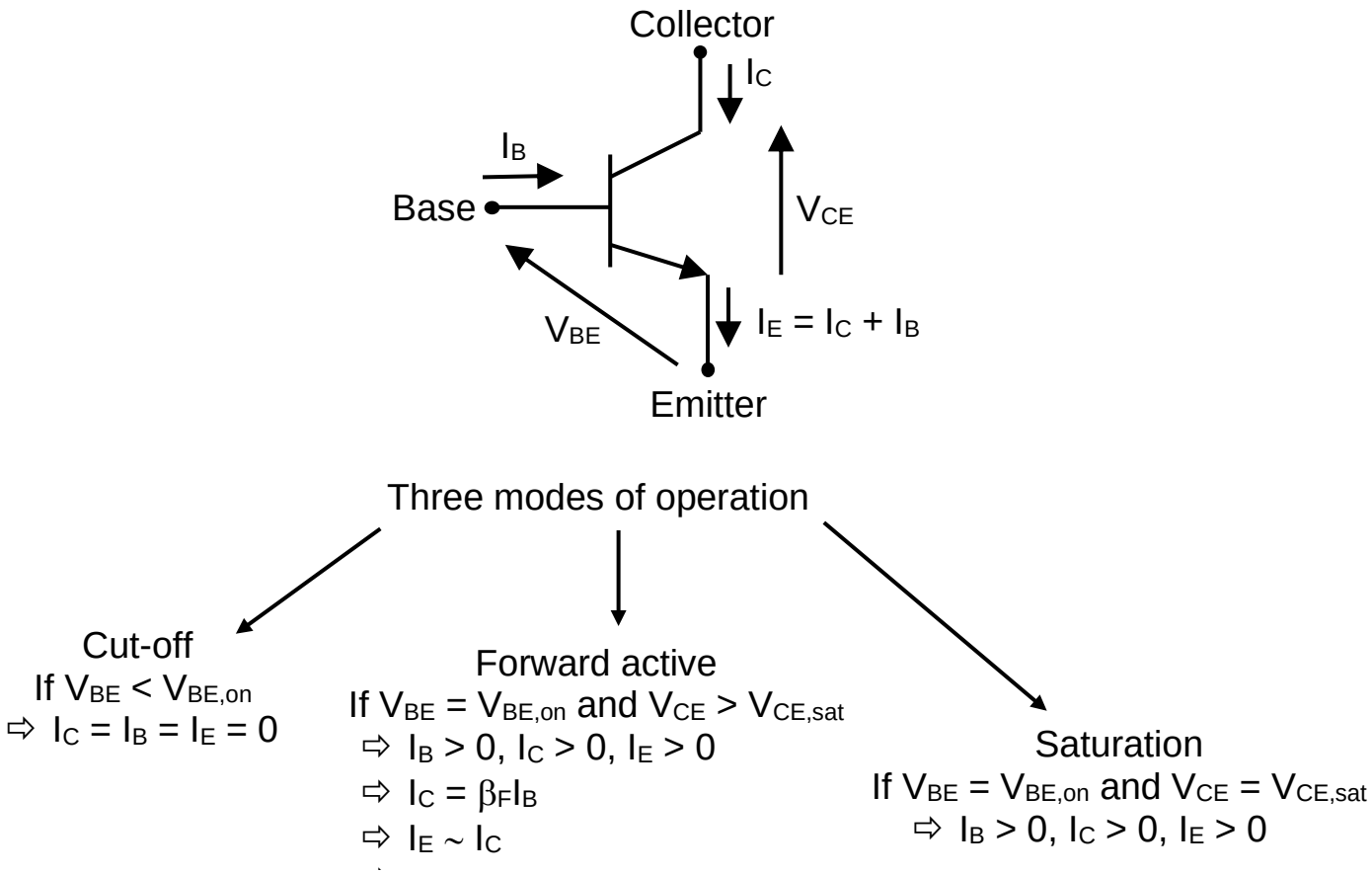
Also, the statement “the BJT is saturated when $V_{BE} = V_{BE,on}$ and V_{CE} = 0 volt” should be replaced with the more accurate statement “the BJT is saturated when $V_{BE} = V_{BE,on}$ and $V_{CE} = V_{CE,sat}$”.

For simplicity sake, we have assumed that $V_{CE} = V_{CE,sat}$ rather than $V_{CE} < V_{CE,sat}$ in the saturation mode. This simplification does not result in any significant error as the actual value of V_{CE} lies somewhere between 0 volt and $V_{CE,sat}$ = 0.2 volt.

This small correction gives us a more accurate model for a BJT.

What to remember?

An NPN BJT is a semiconductor device that has, in practical settings, three possible modes of operation. They are detailed in the figure below.



What can be done with a BJT?

The answer is simple: so many things!

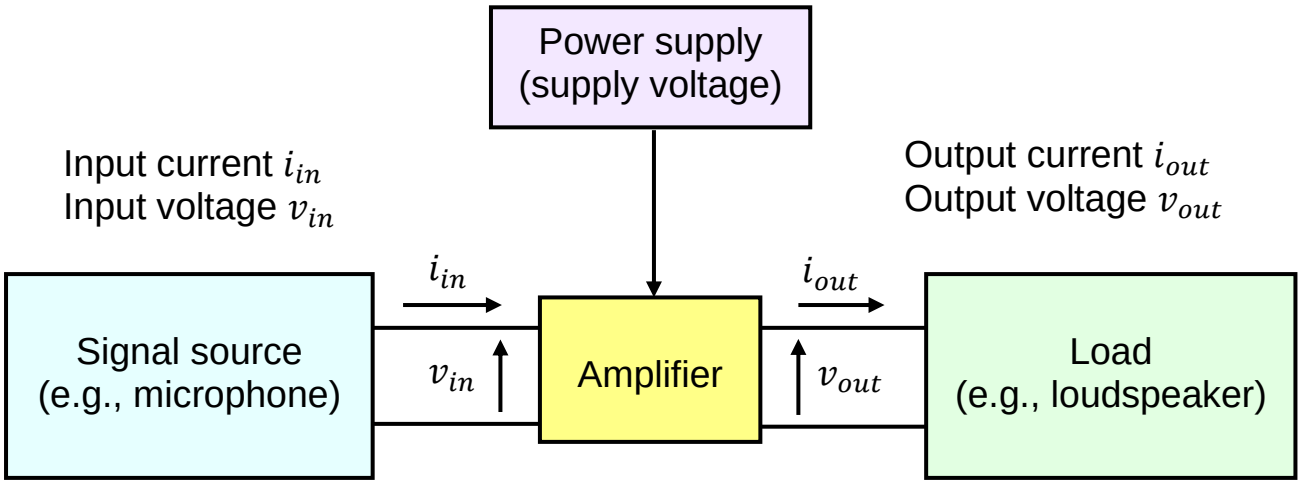
We can design logic gates using BJTs. This used to be commercially viable from the 1950s to the 1980s. There are many ways to implement logic functions

based on BJTs. This led to the development of various logic families, such as the resistor-transistor logic (RTL), diode-transistor logic (DTL), the transistor-transistor logic (TTL), and the emitter-coupled logic (ECL).

Until the end of the 1990s, the ECL family used to be the fastest logic family on silicon, but it could never be employed for designing very complex electronic circuits due to its high power consumption.

For the past several decades, the field-effect transistors have completely replaced the BJTs for all digital applications.

The world of analogue electronics is where BJTs can truly be useful due to their ability to provide high gains when amplifying signals. As an illustration, a traditional amplifier setting is shown in the figure below.



A signal source, e.g. a microphone, produces the signal to be amplified. This signal can be either a voltage v_{in} or a current i_{in} that carries information. This signal is to be used by a loading circuit, e.g. a loudspeaker or another amplifier.

In most cases, the loading circuit can only make efficient use of the information-carrying signal if the latter has a magnitude that is sufficiently large. As the signal produced by the source generally has a small amplitude, it is then necessary to insert, between source and loading circuits, an amplifier in order to increase the magnitude of the information-carrying signal.

For a linear amplifier, the voltage, v_{out}, at the amplifier output is given by $v_{out} = A_v \cdot v_{in}$, where A_v denotes the voltage gain of the amplifier. For a "good" voltage amplifier, we should have $A_v \gg 1$.

The expression $v_{out} = A_v \cdot v_{in}$ clearly implies that v_{out} is, at all times, proportional to v_{in}. In other words, there is no distortion introduced by the amplifier and the information carried by v_{in} is therefore fully preserved.

In the same way, we can say that, for a linear amplifier, the current, i_{out}, provided to the loading circuit by the amplifier, is expressed as $i_{out} = A_i \cdot i_{in}$, where A_i denotes the current gain of the amplifier. For a "good" current amplifier, we should have $A_i \gg 1$.

The expression $i_{out} = A_i \cdot i_{in}$ clearly implies that i_{out} is, at all times, proportional to i_{in}. In other words, there is no distortion introduced by the amplifier and the information carried by i_{in} is therefore fully preserved.

The amplifier provides more electric power to the loading circuit than it absorbs from the source. This electric power must come from somewhere. This is why

an amplifier is always connected to a power supply that, in most cases, is a battery that applies a DC voltage across the circuit.

In linear amplifiers, the BJT should, at all times, operate in the forward-active mode. In this case, the price to pay is a poor power efficiency, meaning that much of the power delivered by the battery is actually wasted.

If we can accept some distortion of the input signals during the amplification process, which is often the case in practice, it is possible to design non-linear amplifiers that possess the advantage of having higher power efficiency.

Higher power efficiency means that more power from the battery is actually delivered to the loading circuit. This is particularly important in applications where power consumption is a critical issue (smartphones, laptops, etc.). Amplifiers with high power efficiency are referred to as power amplifiers.

The key to increase the power efficiency is to allow the transistor to operate at times in the cut-off mode rather than always staying in the forward-active mode. The good thing about the cut-off mode is that no power is consumed by a BJT in that mode.

Examples of applications where a voltage amplifier is required

(1) An electrocardiogram (ECG) is used to detect the heart's electrical activity to identify eventual heart disorders. It utilises a pair of electrodes to reveal the ionic potential difference between their respective points of application on the skin. The voltage detected between the electrodes is around 1 – 5 mV.

(2) A thermocouple produces a voltage proportional to a temperature.

(3) A light-to-voltage (LTV) sensor provides a voltage output proportional to light.

(4) Most microphones produce a voltage signal from mechanical vibrations.

(5) Bio-signals are recorded as very low-level voltages (from 1 μV to 100 mV) generated by nerves, muscles, etc.

(6) The brain waves use electroencephalography (EEG) to furnish potential differences in the microvolt range measured across locations on the user's scalp.

For these examples, voltage amplification is typically required for further signal processing or to simply drive a display device or a loudspeaker system.

Example of application where a current amplifier is required

Photovoltaic cells and photodiodes act like current sources since they generate a current proportional to the intensity of light. In this case, current amplification is typically required when the amplified signal is used to drive a meter.

Analysis of a BJT Circuit

To analyse an electronic circuit including semiconductor devices such as diodes and transistors, we must always follow the same steps:

(1) Write a set of general equations using KVL, KCL, and Ohm's law. In other words, gather as much information as possible using the laws of circuit theory;

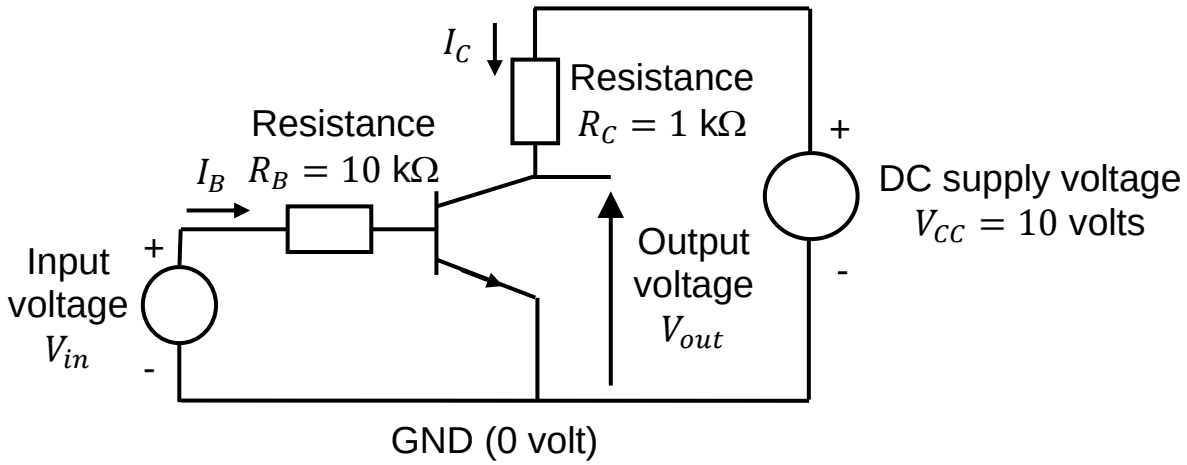
(2) Consider each possible mode of operation for the semiconductor device(s) and rewrite the general equations for each mode;

(3) Draw conclusions by combining all obtained equations for all modes of operation. Do not forget anything as all information is precious.

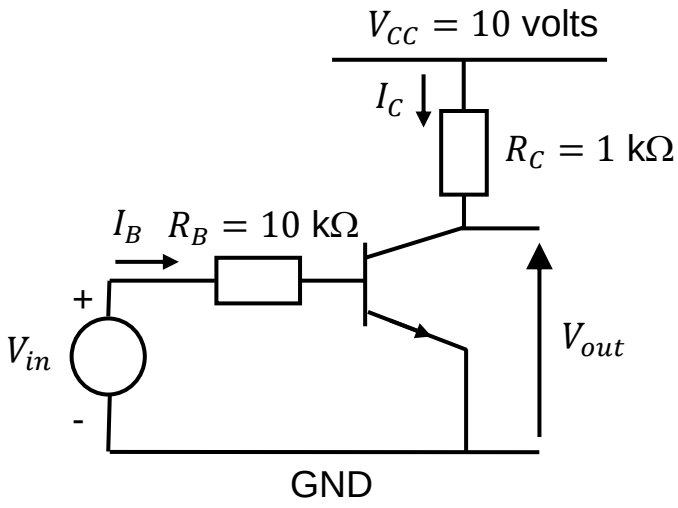
This methodology of manual analysis never fails, but it requires a bit of practice to be comfortable with it. At some stage, it will become a second nature.

Let us consider an example of BJT circuit in order to illustrate this methodology.

Consider the circuit depicted below. The NPN BJT has a forward current gain $\beta_F = 100$.

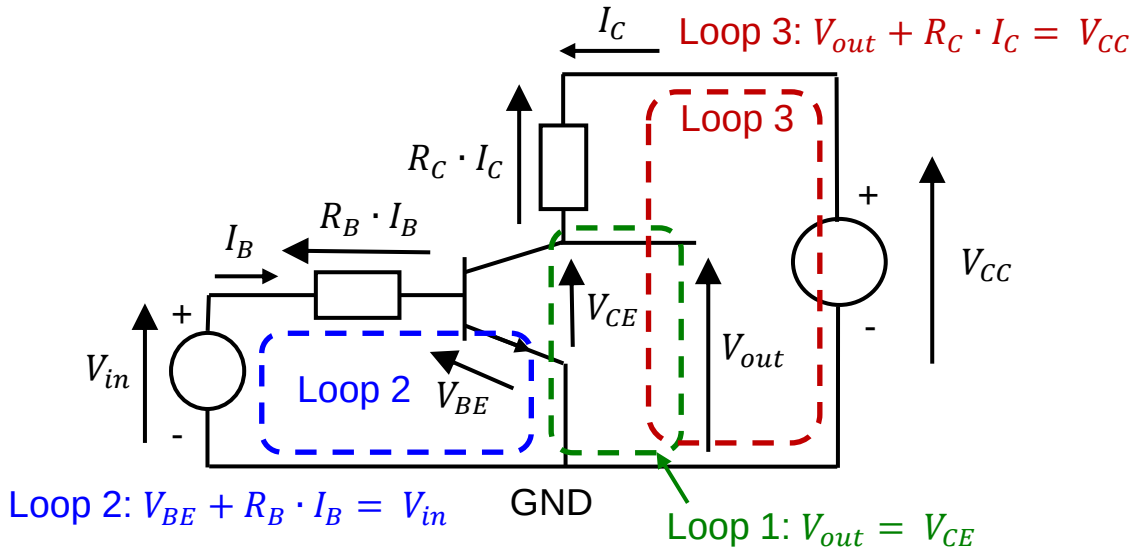


Electronic engineers would prefer the lighter figure shown below.



Our goal throughout this exercise is to find the DC voltage transfer characteristic for this circuit. In other words, for each value of the input voltage V_{in} ranging from 0 to $V_{CC} = 10$ volts, we want to determine the corresponding output voltage V_{out}.

Let us write the general equations using Kirchhoff voltage law (KVL) and Ohm's law. Kirchhoff current law (KCL) cannot be applied here as there is no node in the circuit.



Using KVL around loop 1, we obtain $V_{out} = V_{CE}$.

By jointly applying KVL and Ohm's law around loop 2, we have $V_{BE} + R_B \cdot I_B = V_{in}$.

In the same way, by jointly applying KVL and Ohm's law around loop 3, we can write $V_{out} + R_C \cdot I_C = V_{CC}$.

We thus obtain a set of three equations that are always valid, regardless the mode in which the BJT operates. These equations are referred to as general equations.

Equation 1: $V_{out} = V_{CE}$;

Equation 2: $V_{BE} + R_B \cdot I_B = V_{in}$;

Equation 3: $V_{out} + R_C \cdot I_C = V_{CC}$.

Now, we can move to the next step. Consider the three possible modes of operation and examine what happens to the general equations.

(1) First mode of operation: Cut-off mode if $V_{BE} < V_{BE,on}$ and $I_B = I_C = I_E = 0$.

In this mode of operation, Equation 1, $V_{out} = V_{CE}$, remains unchanged.

Equation 2, $V_{BE} + R_B \cdot I_B = V_{in}$, becomes $V_{BE} = V_{in}$ because $I_B = 0$. This result yields $V_{in} < V_{BE,on}$ as $V_{BE} < V_{BE,on}$.

The inequality $V_{in} < V_{BE,on}$ provides us with a condition on the input voltage V_{in} for the BJT to be cut-off.

Equation 3, $V_{out} + R_C \cdot I_C = V_{CC}$, leads to $V_{out} = V_{CC}$ because $I_C = 0$.

Conclusion: when $V_{in} < V_{BE,on}$, the BJT is cut-off and we have $V_{out} = V_{CC}$.

(2) Second mode of operation: Forward-active mode if $V_{BE} = V_{BE,on}$ and $V_{CE} > V_{CE,sat}$. In this case, we have $I_C = \beta_F \cdot I_B$, $I_B > 0$, $I_C > 0$, $I_E > 0$, and $I_E \sim I_C$.

Some of these equations are not necessarily going to be useful in this particular exercise, and some are even redundant. But it is always good practice to bring as much information as possible in case we need it.

In the forward-active mode, Equation 1, $V_{out} = V_{CE}$, yields $V_{out} > V_{CE,sat}$ because $V_{CE} > V_{CE,sat}$.

Equation 2, $V_{BE} + R_B \cdot I_B = V_{in}$, becomes $V_{BE,on} + R_B \cdot I_B = V_{in}$ as $V_{BE} = V_{BE,on}$, which leads to $V_{BE,on} < V_{in}$ as $I_B > 0$.

Equation 3, $V_{out} + R_C \cdot I_C = V_{CC}$, leads to $V_{out} + R_C \cdot \beta_F \cdot I_B = V_{CC}$ since $I_C = \beta_F \cdot I_B$. It also yields $V_{out} < V_{CC}$ as $I_B > 0$.

So, at this stage, we have discovered that the BJT is forward active when $V_{in} > V_{BE,on}$ and $V_{CE,sat} < V_{out} < V_{CC}$.

But we need to know more than that. Remember that our goal is to express the output voltage V_{out} as a function of the input voltage V_{in}. This can be done by further exploiting Equations (2) and (3) as follows.

Equation (3) can be written as $V_{out} = V_{CC} - R_C \cdot \beta_F \cdot I_B$. The only unknown in this equation is the base current I_B because the values of the circuit parameters V_{CC}, R_C, and β_F are known.

Using Equation (2), we can obtain an expression of I_B as a function of V_{in}:

$$I_B = \frac{V_{in} - V_{BE,on}}{R_B}.$$

Equation (3) finally leads to

$$V_{out} = V_{CC} - R_C \cdot \beta_F \cdot \frac{V_{in} - V_{BE,on}}{R_B} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right).$$

This shows that the output voltage V_{out} varies linearly with V_{in}. The slope of this linear function is negative and given by $-\beta_F \cdot \frac{R_C}{R_B} = -10$.

Conclusion: when $V_{in} > V_{BE,on}$ and $V_{CE,sat} < V_{out} < V_{CC}$, the BJT is forward active and we have $V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right)$.

(3) Third mode of operation: Saturation mode if $V_{BE} = V_{BE,on}$ and $V_{CE} = V_{CE,sat}$. In this case, we have $I_B > 0$, $I_C > 0$, and $I_E > 0$.

In the saturation mode, Equation 1, $V_{out} = V_{CE}$, becomes $V_{out} = V_{CE,sat}$ as $V_{CE} = V_{CE,sat}$.

Equation 2, $V_{BE} + R_B \cdot I_B = V_{in}$, leads to $V_{BE,on} + R_B \cdot I_B = V_{in}$ since $V_{BE} = V_{BE,on}$. This result also yields $V_{BE,on} < V_{in}$ as $I_B > 0$;

Equation 3, $V_{out} + R_C \cdot I_C = V_{CC}$, leads to $V_{out} < V_{CC}$ as $I_C > 0$.

Conclusion: when $V_{in} > V_{BE,on}$ and $V_{out} = V_{CE,sat}$, the BJT is saturated and we have $V_{out} = V_{CE,sat}$.

This statement may seem a bit weird, but there is nothing to be afraid of. This is the result obtained using the equations rigorously and it is therefore absolutely correct. We only need to make sense of it.

Now, we are ready to move to the third and final step: Putting everything together.

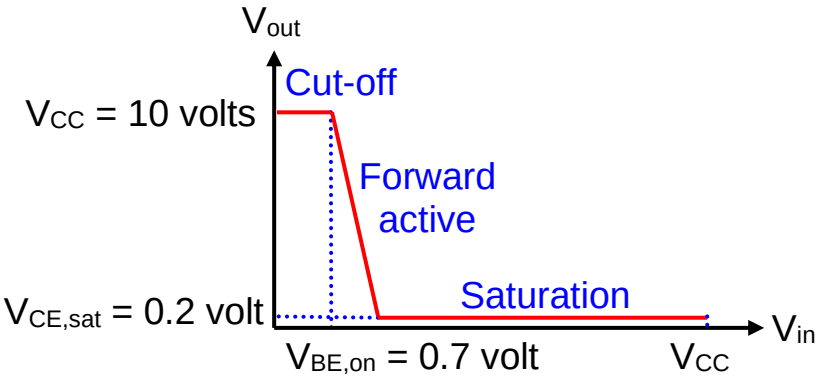
When $V_{in} < V_{BE,on}$, the BJT is cut-off and we have $V_{out} = V_{CC}$.

As V_{in} is increased beyond $V_{BE,on}$, the transistor enters the forward-active mode and the output voltage V_{out} decreases linearly with V_{in} according to the equation

$$V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right).$$

Once V_{out} becomes equal to $V_{CE,sat}$, the transistor enters the saturation mode, and we thus have $V_{out} = V_{CE,sat}$.

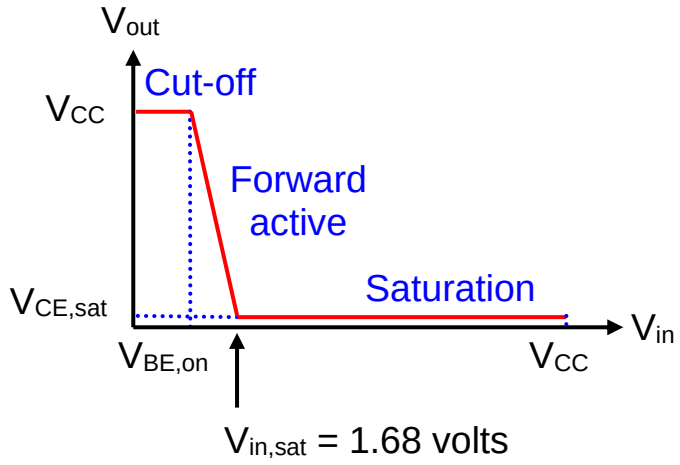
The resulting DC voltage transfer characteristic is shown below.



We can also determine the value, $V_{in,sat}$, of the input voltage at which the BJT leaves the forward-active mode to enter the saturation mode. At $V_{in} = V_{in,sat}$, the output voltage V_{out} is equal to $V_{CE,sat}$. We can thus write

$$V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in,sat} - V_{BE,on}\right) = V_{CE,sat},$$

which yields $V_{in,sat} = V_{BE,on} + \frac{R_B}{\beta_F \cdot R_C} \cdot \left(V_{CC} - V_{CE,sat}\right) = 1.68$ volts.



Finally, we can summarize the operation of the circuit as follows.

- When $V_{in} < V_{BE,on}$, the BJT is cut-off and we have $V_{out} = V_{CC}$.
- When $V_{BE,on} < V_{in} < V_{in,sat}$, the BJT is forward active and we have

$$V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right).$$

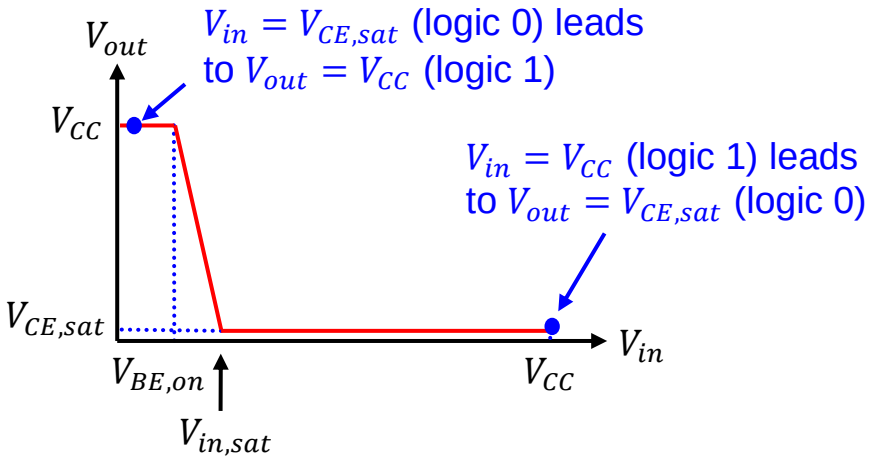
• When $V_{in,sat} < V_{in}$, the BJT is saturated and we have $V_{out} = V_{CE,sat}$.

What could be the possible applications of this circuit?

This circuit could be used as a logic inverter (NOT gate).

Let us define, for this circuit, a logic 0 as a (low) voltage $V_{CE,sat} = 0.2$ volt and a logic 1 as a (high) voltage $V_{CC} = 10$ volts.

In this case, we see that $V_{in} = V_{CE,sat}$ (logic 0) leads to $V_{out} = V_{CC}$ (logic 1), while $V_{in} = V_{CC}$ (logic 1) leads to $V_{out} = V_{CE,sat}$ (logic 0). This circuit clearly implements a logic inversion function.



We can also observe the potential of this circuit as a voltage amplifier when the BJT is forward active because, in that mode, the output voltage V_{out} is a linear function of the input voltage V_{in}: $V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot (V_{in} - V_{BE,on})$.

To clarify this important point, let us assume that the BJT is forward active and an initial input voltage V_{in} is increased by an amount ΔV_{in}. The new input voltage is thus expressed as $V_{in}' = V_{in} + \Delta V_{in}$.

The new output voltage V_{out}' corresponding to V_{in}' is given by

$$V_{out}' = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in}' - V_{BE,on}\right),$$

which is equivalent to

$$V_{out}' = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right) - \beta_F \cdot \frac{R_C}{R_B} \cdot \Delta V_{in}.$$

This equation indicates that the new output voltage V_{out}' can be written as the sum of the initial output voltage V_{out}, given by

$$V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right),$$

and an increase

$$\Delta V_{out} = -\beta_F \cdot \frac{R_C}{R_B} \cdot \Delta V_{in}.$$

In other words, increasing the input voltage by an amount ΔV_{in} leads to a corresponding increase $\Delta V_{out} = -\beta_F \cdot \frac{R_C}{R_B} \cdot \Delta V_{in}$ in the output voltage.

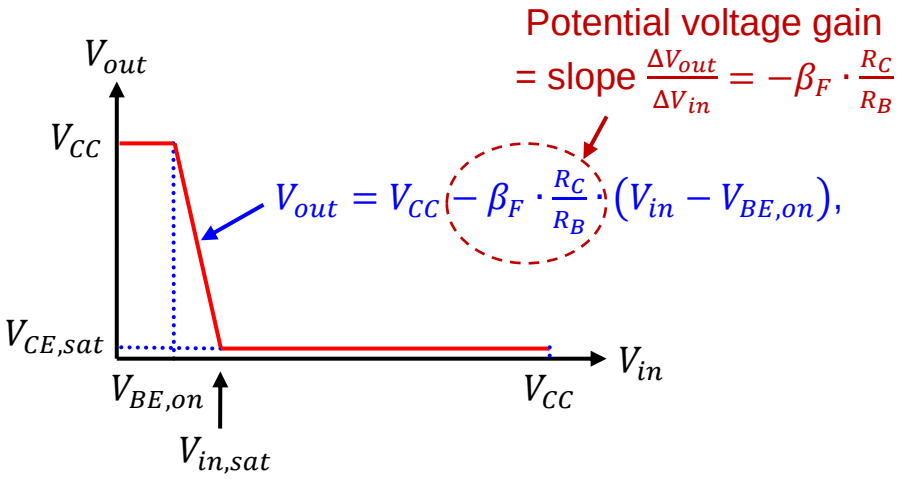
Since the term $\beta_F \cdot \frac{R_C}{R_B}$ is typically much larger than the unit, we conclude that a small variation ΔV_{in} in the input voltage can potentially generate a much larger variation ΔV_{out} in the output voltage.

This is the behaviour of a linear voltage amplifier with a voltage gain

$$\frac{\Delta V_{out}}{\Delta V_{in}} = -\beta_F \cdot \frac{R_C}{R_B}.$$

Note that the presence of a minus sign in the voltage gain means that the variations in V_{out} and V_{in} are in opposite directions, which is not necessarily a major issue in practice.

Finally, note that the potential voltage gain $\frac{\Delta V_{out}}{\Delta V_{in}}$ of this circuit is simply the slope of the DC voltage transfer characteristic in the forward-active region.

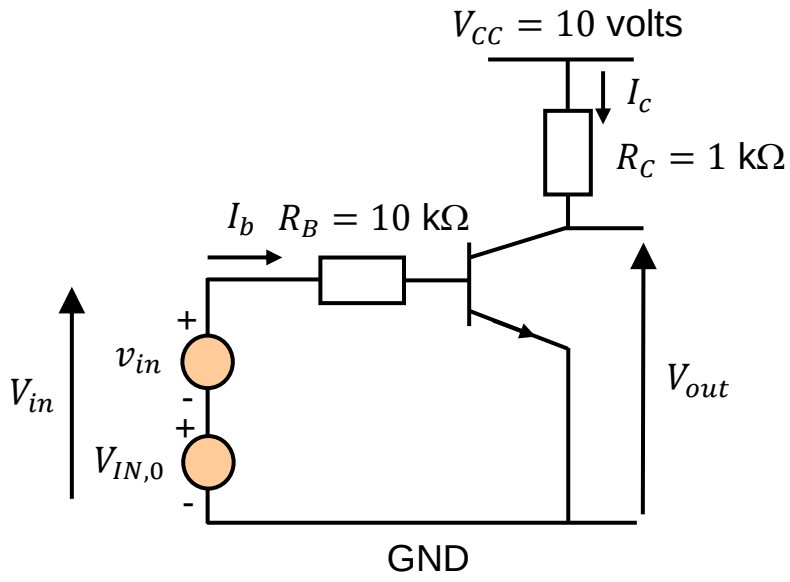


To make our circuit work as a linear amplifier in practice, we will have to make sure that the BJT remains at all times in the forward-active mode and enters neither saturation nor cut-off mode.

This issue is going to be addressed in the next section of this document.

Design of a Linear BJT amplifier

We now slightly modify the previous circuit in order to design a *linear amplifier*. The modified circuit is shown in the figure below. We are now going to study this circuit in order to understand the principles of amplifier design in analogue electronics.

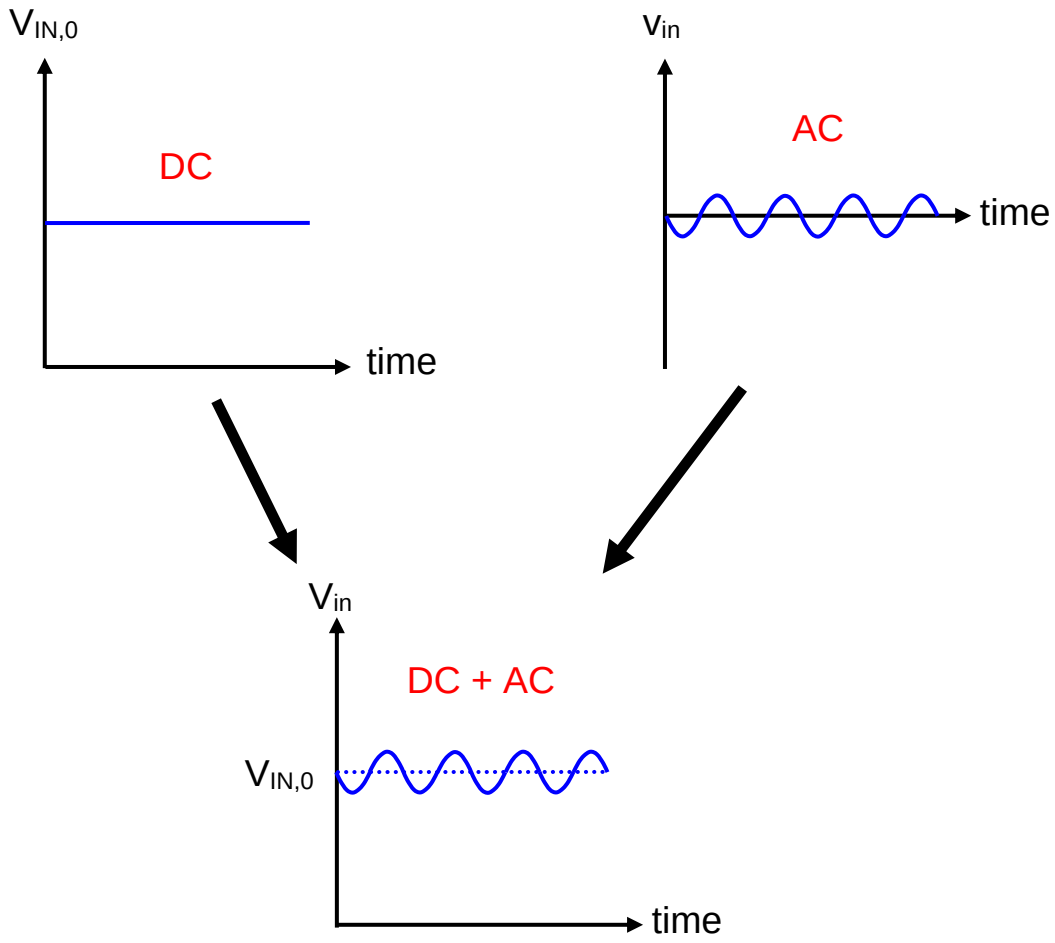


The input voltage V_{in} is now composed of two voltage sources $V_{IN,0}$ and v_{in} connected in series. Hence, we can write $V_{in} = V_{IN,0} + v_{in}$.

The DC voltage source $V_{IN,0}$ is used to bias the circuit so that the BJT operates at all times in the forward-active mode. The quantity $V_{IN,0}$ is often referred to as bias voltage and its value has to be chosen very carefully by the designer. Remember that a BJT has the ability to amplify a current or voltage only when it operates in the forward-active mode.

The AC voltage source v_{in} generates the AC signal to be amplified. Here, we assume, without loss of generality, that v_{in} is a pure AC voltage signal with zero mean and small-amplitude symmetrical swings around zero volt.

The AC voltage signal v_{in} carries the precious information, e.g. a voice signal, that needs to be amplified.



In the previous circuit, we have seen that, as long as the BJT operates in the forward active mode, the output voltage V_{out} is given by

$$V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{in} - V_{BE,on}\right).$$

So, assuming operation in the forward-active mode, we can write here that

$$V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(v_{in} + V_{IN,0} - V_{BE,on}\right),$$

which is equivalent to $V_{out} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{IN,0} - V_{BE,on}\right) - \beta_F \cdot \frac{R_C}{R_B} \cdot v_{in}$.

This result means that we can express the output voltage as follows:

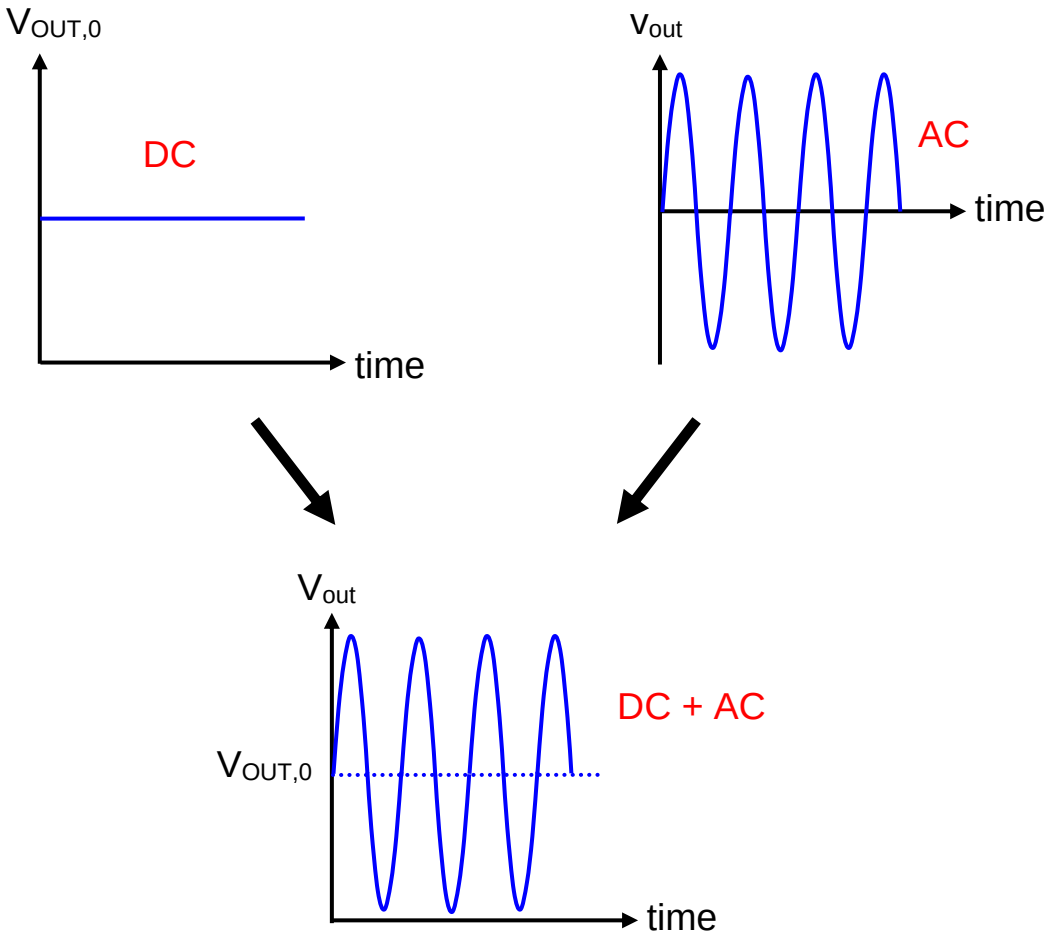
$$V_{out} = V_{OUT,0} + v_{out},$$

where

- $V_{OUT,0} = V_{CC} - \beta_F \cdot \frac{R_C}{R_B} \cdot \left(V_{IN,0} - V_{BE,on}\right)$ represents the DC component of the output voltage;
- $v_{out} = -\beta_F \cdot \frac{R_C}{R_B} \cdot v_{in}$ is an amplified replica of the input AC signal.

This result shows that the output voltage V_{out} is composed of a DC component, $V_{OUT,0}$, and a pure AC component, v_{out}, with zero mean and symmetrical swings around zero volt.

This result also indicates that the AC component v_{out} of the output signal is an amplified version of the AC component v_{in} of the input signal.



The voltage gain of the circuit is defined as

$$A_v = \frac{v_{out}}{v_{in}},$$

and is thus given here by

$$A_v = -\beta_F \cdot \frac{R_C}{R_B} = -10,$$

since $\beta_F = 100$, $R_B = 10\ k\Omega$, and $R_C = 1\ k\Omega$.

The DC voltages $V_{IN,0}$ and $V_{OUT,0}$ are the *bias voltages*. Their values must be chosen carefully by the circuit designer. To understand the reason behind this statement, we need to introduce the concept of maximum voltage swing for an amplifier.

The maximum output voltage swing, $\Delta V_{out,max}$, is the maximum peak-to-peak amplitude of the AC output voltage v_{out} that guarantees no distortion in v_{out}.

In the same way, the maximum input voltage swing, $\Delta V_{in,max}$, can be defined as the maximum peak-to-peak amplitude of the AC input voltage v_{in} that guarantees no distortion in v_{out}.

For a linear amplifier, we must ensure that the AC output voltage v_{out} remains, at all times, proportional to the AC input voltage v_{in}, i.e. the expression $v_{out} = A_v \cdot v_{in}$, where A_v denotes the voltage gain of the circuit, remains valid at all times. If it is not the case, the amplifier does not perform linear amplification.

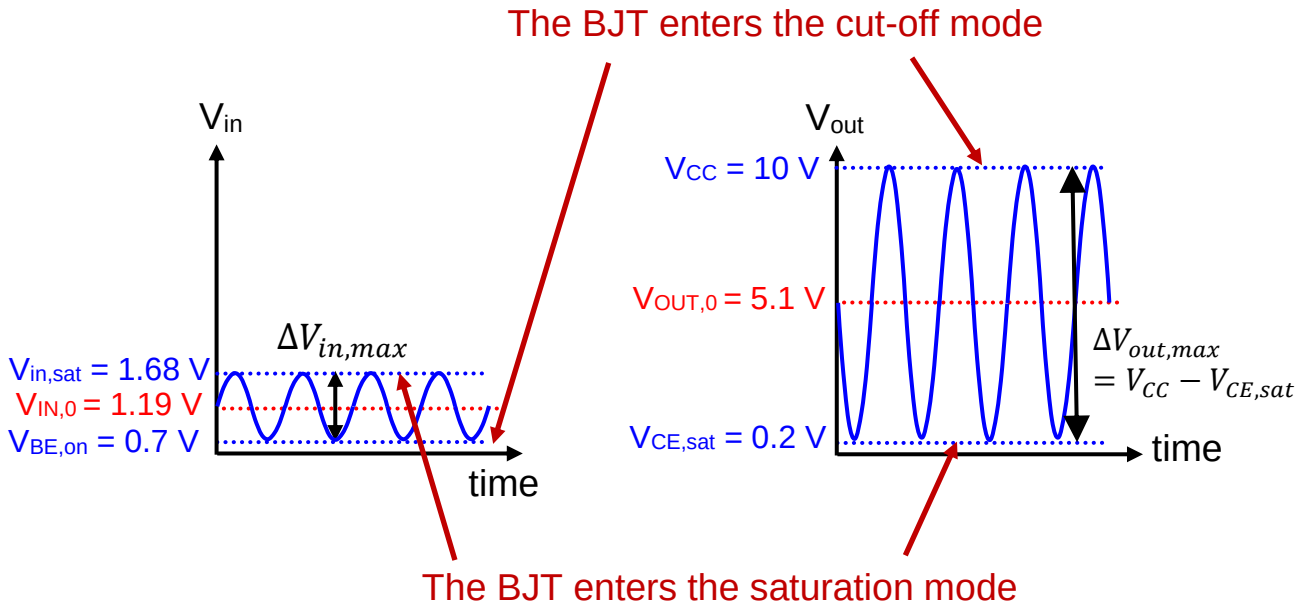
This is why it is indeed crucial to avoid introducing any distortion in the output signal. If distortion occurs, v_{out} will no longer be an amplified version of v_{in}, and the circuit will no longer implement a linear amplification function.

A well-designed amplifier should have the largest possible maximum voltage swings at its input and output. An important part of the analogue designer's work is thus to maximise both maximum voltage swings $\Delta V_{in,max}$ and $\Delta V_{out,max}$.

By the way, there is a very simple equation linking both maximum voltage swings: $\Delta V_{out,max} = |A_v| \cdot \Delta V_{in,max}$, where A_v is the voltage gain of the amplifier. This expression is a direct consequence of the equation $v_{out} = A_v \cdot v_{in}$.

The "space" in which input and output voltages can swing is that located between the start of the cut-off region and the start of the saturation region.

For the circuit studied here, it is easy to see that the maximum value for the maximum output voltage swing, $\Delta V_{out,max}$, is given by $V_{CC} - V_{CE,sat} = 9.8$ volts. No matter how well our circuit is designed, the value of $\Delta V_{out,max}$ cannot exceed $V_{CC} - V_{CE,sat} = 9.8$ volts.



In the same way, the maximum value for the maximum input voltage swing, $\Delta V_{in,max}$, is given by $V_{in,sat} - V_{BE,on} = 0.98$ volts. In other words, the maximum peak-to-peak amplitude of the AC input voltage, v_{in}, that guarantees no distortion in v_{out} cannot exceed $V_{in,sat} - V_{BE,on} = 0.98$ volts.

The goal of the analogue designer is to choose the values of the bias voltages $V_{IN,0}$ and $V_{OUT,0}$ so that $\Delta V_{in,max} = V_{in,sat} - V_{BE,on}$ and $\Delta V_{out,max} = V_{CC} - V_{CE,sat}$.

In our circuit, it is easy to see that, in order to maximise the maximum output voltage swing, we have to ensure that $V_{OUT,0}$ is located midway between V_{CC} and $V_{CE,sat}$:

$$V_{OUT,0} = \frac{V_{CC} + V_{CE,sat}}{2} = 5.1 \text{ volts.}$$

In the same way, the optimal value of $V_{IN,0}$ is also located midway between the start of the cut-off region, corresponding to $V_{in} = V_{BE,on}$, and the start of the saturation region, corresponding to $V_{in} = V_{in,sat}$:

$$V_{IN,0} = \frac{V_{BE,on} + V_{in,sat}}{2} = \frac{0.7V + 1.68V}{2} = 1.19 \text{ volts.}$$

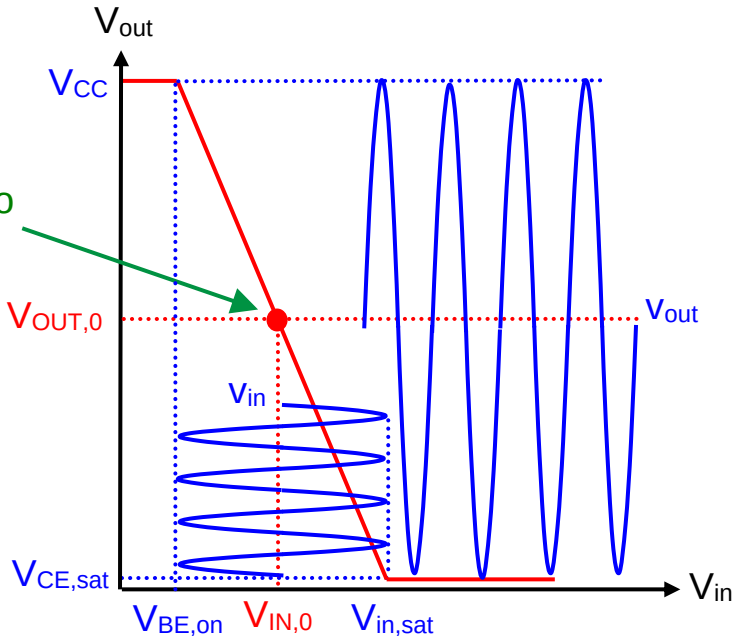
Optimally-located DC bias voltages $V_{IN,0}$ and $V_{OUT,0}$ allow maximal swings for both input and output voltages, as illustrated in the figure below.

The BJT amplifier studied here is able to amplify and track changes in the AC input voltage v_{in} as long as the BJT remains within the limits set by cut-off and saturation. If the magnitude of v_{in} increases beyond those limits, the output voltage v_{out} is clipped and distortion then occurs.

$V_{IN,0}$ and $V_{OUT,0}$ located midway between cut-off and saturation regions, leading to maximum input and output voltage swings:

$$\Delta V_{out,max} = V_{CC} - V_{CE,sat}$$

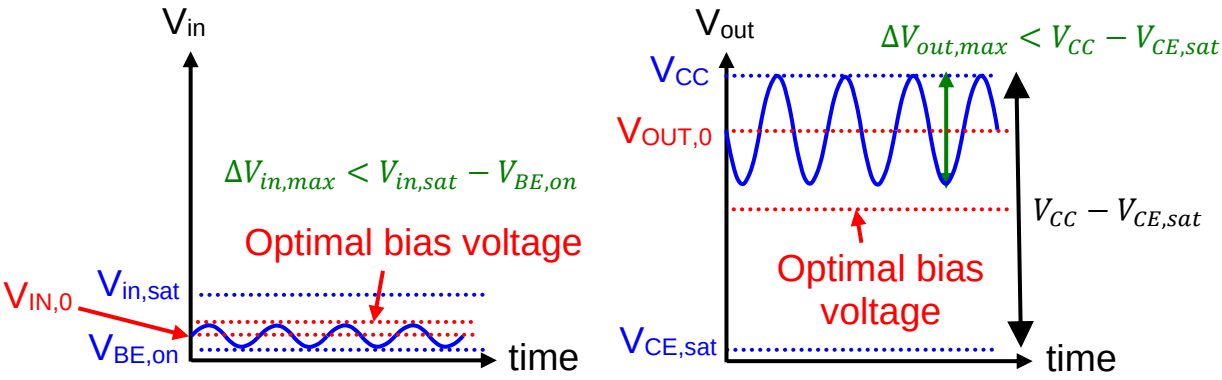
$$\Delta V_{in,max} = V_{in,sat} - V_{BE,on}$$



Badly-located bias voltages $V_{IN,0}$ and $V_{OUT,0}$, i.e. non-optimal biasing, can significantly reduce the maximal swings for both input and output voltages.

This point is illustrated in the figures shown below for the case when the BJT operates too close to cut-off due to a bias voltage $V_{IN,0}$ chosen too close to $V_{BE,on}$ and too far from $V_{in,sat}$.

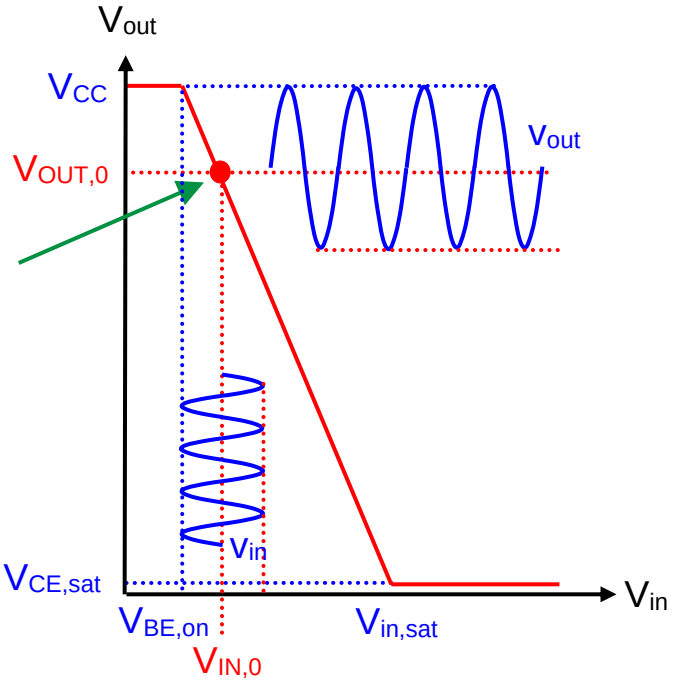
Maximum voltage swings are not optimal



Badly-located $V_{IN,0}$ and $V_{OUT,0}$, leading to a reduction in max input and output voltage swings:

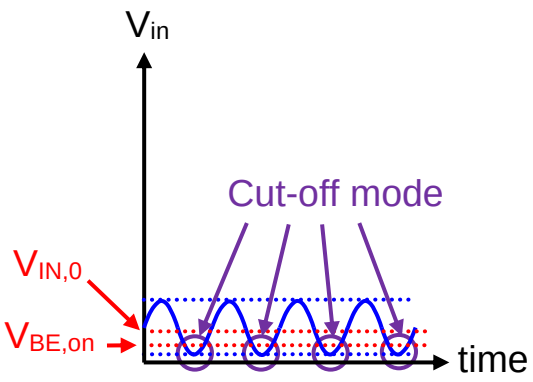
$\Delta V_{out,max} < V_{CC} - V_{CE,sat}$

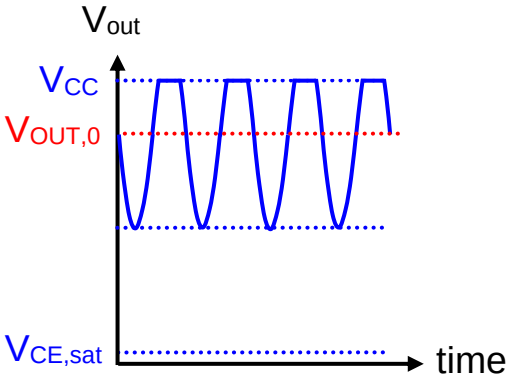
$\Delta V_{in,max} < V_{in,sat} - V_{BE,on}$

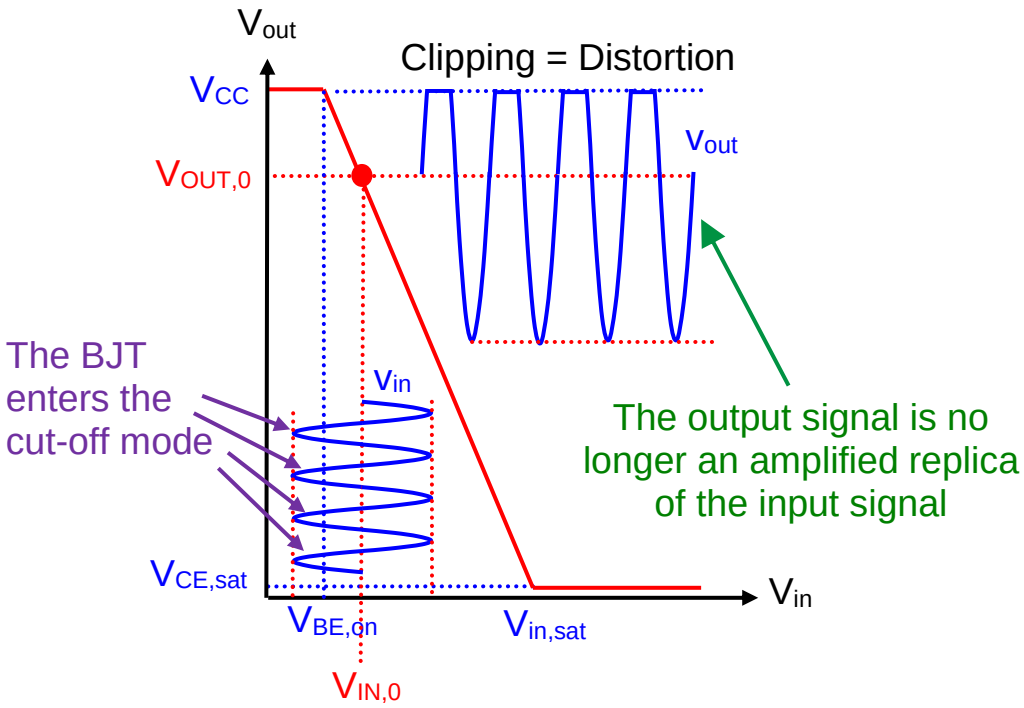


If the transistor is allowed to enter either the cut-off mode or the saturation mode, the output signal will then be clipped, and distortion will be introduced in v_{out}.

This will definitely happen if the peak-to-peak amplitude of the input signal v_{in} is larger than the maximum voltage swing $\Delta V_{in,max}$. In this case, our circuit is no longer a linear amplifier.







An example is shown in the figure above where the BJT is allowed to be cut-off due to badly-located bias voltages $V_{IN,0}$ and $V_{OUT,0}$ combined with an AC input voltage v_{in} with excessive magnitude.

In the figure above, it clearly appears that the BJT enters the cut-off mode when $V_{in} < V_{BE,on}$, i.e. when $V_{IN,0} + v_{in} < V_{BE,on}$, i.e. when $v_{in} < V_{BE,on} - V_{IN,0}$. This inequality indicates that, if we want to prevent the BJT from entering the cut-off mode, we must reduce the magnitude of v_{in} and/or raise the value of $V_{IN,0}$.

Note that there are applications where some distortion is acceptable, for instance, when the clipped part of the output voltage signal v_{out} can be extrapolated based on the part that has been preserved. This happens, for example, when it is known that v_{out} must be a sinusoidal waveform.

When an amplifier is intentionally designed to introduce some degree of distortion in the output voltage, such amplifier is known as a non-linear amplifier. Generally, the BJT is allowed to enter the cut-off mode rather than having to stay in the forward-active mode at all times. This reduces the power consumption, and thus increases the power efficiency of the circuit.

As a conclusion, we have shown how to design a linear BJT amplifier circuit that has a voltage gain $A_v = -\beta_F \cdot \frac{R_C}{R_B} = -10$ and can amplify, without any distortion, an AC input signal v_{in}, as long as the peak-to-peak amplitude of the latter does not exceed $\Delta V_{in,max} = V_{in,sat} - V_{BE,on} = 0.98$ volt.