Bipolar Junction Transistors (BJT)

The diode opened up a world of possibilities by being able to respond very differently to different circuit conditions -- conducting when forward biased and not conducting when reverse biased. The transistor takes that about an order of magnitude farther, because transistors can be externally controlled, and can provide a range of responses rather than just "on" (with a small voltage drop) and "off".

There are a number of different transistors -- to begin with, there are BJTs and FETs (next topic, and there are three main families of FETs). Then, in each group there are P-type and N-type devices. So, this isn't going to be quite as straightforward as diodes!

Introduction to Amplification

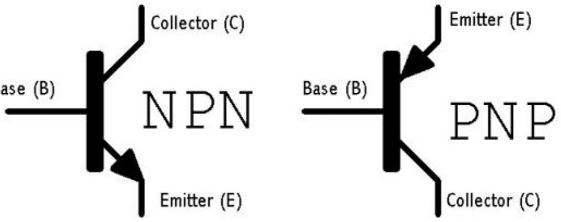
Transistors are designed for current amplification, which can be turned into voltage amplification and power amplification. Amplification means to turn a small signal into a large one. Clearly, that can't be done without the input of energy from outside the device. A transistor can be seen as a device that can "read" a small incoming signal and control a large external power source to produce a large output signal. Without the external power source, nothing happens.

It's like making a car move. You push down on the accelerator with just your foot, and the massive vehicle surges ahead -- but only if the engine is running. No power source, no response -- your foot didn't do the work; the engine did, but at the command of your foot.

BJTs respond to a very small input current by producing a much larger output current, if there's a DC power supply connected.

Transistors are three-pin devices: The big current from the power supply goes through two of the pins, and the third pin provides the control.

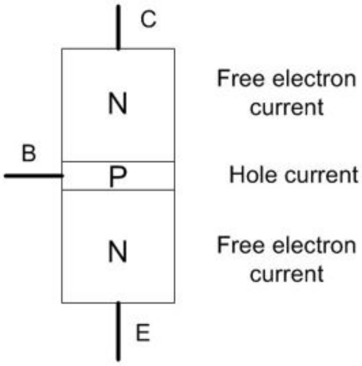
The schematic symbols for the two flavours of BJT are shown below.



In each of these, a large current can flow through the Collector-Emitter path, controlled by a small current injected at the Base. Notice the difference in the way the pins are drawn: The Collector and Emitter are drawn in-line with each other to show the primary current path, whereas the Base is off to the side to provide control. Also, the Emitter always has an arrow on it to show the direction of conventional current, whereas the Collector has no arrow.

Without getting into too much Quantum Physics, here's a brief explanation as to how the NPN version of the BJT works from the perspective of conventional current, and a similar explanation holds for the PNP, but from the point of view of electron current.

Picture the NPN transistor as three blocks of semiconductor material.



This transistor is called "bipolar" because it uses both hole and free electron current in order to conduct.

In its natural state, there are two P-N junctions, and therefore two barrier potentials to deal with.

Transistors are always biased so that the Base to Collector junction (B-C junction) is reverse biased. This naturally prevents the flow of current, so a BJT at rest doesn't conduct. In the case of the NPN, this biasing arrangement means that the Collector voltage, V_C, will always be the at the highest potential of the three, with the Emitter at a lower potential so current can flow from the collector to it. In the case of the PNP, V_C is always the lowest voltage and V_E is always the highest voltage.

The Base to Emitter junction is the one we control. If the B-E junction is not forward biased, the transistor does not conduct (unless you blast it with a higher external voltage than it can handle -- more on that later). When the B-E junction is forward biased, the amount of current flowing through it directly determines how much current flows from the Collector to the Emitter and out. This happens, roughly, because the layer of "P" material connected to the Base is really thin, and when current flows from Base to Emitter, it "erodes" the B-C junction, allowing current to cross over. The more Base current, the deeper the "erosion", and therefore the more current flows from the Collector.

This means that both the current from the Collector, I_C and the current from the Base, I_B end up flowing into and through the Emitter, so that $I_E = I_C + I_B$. Since the Base current is relatively tiny, I_C and I_E are pretty similar, with I_E being just a tiny bit bigger.

By the way, we don't make transistors by "attaching" three blocks of semiconductor material together. We start with a layer called the Substrate, say N material in the NPN transistor, sputter a thin layer of P material onto that, then sputter another layer of N material onto that. Next comes a metallization layer that allows us to make physical connections to the transistor. As a result, we can make these things so tiny that an Apple A13 microprocessor has 8.5 billion of them in a single layer in the area of about the size of your fingernail! These transistors are on the order of 7 nm in size!

Another thing to know is that the material used in the Collector is a bit different from the material used in the Emitter. The Emitter material is very heavily doped, the Collector material is heavily doped but not like the Emitter, and the Base material is lightly doped. So it matters which end is the Emitter and which is the Collector.

BJT Relationships

The BJT is, as mentioned, a Current-Controlled Current Source. A tiny Base current controls a large Collector current, and the two combine to produce the Emitter current.

$$I_E = I_C + I_B$$

The current gain, given the symbol β, tells us how much bigger the useful current -- usually I_C -- is than the controlling current, I_B.

$$\beta = \frac{I_C}{I_B}$$

or, more often this is written as $I_C = \beta I_B$

Sometimes β is referred to in a different system of "h-parameters" as h_{FE}. This is the only h-parameter you'll see in this course, but watch for it and remember that it's the same as β.

The efficiency of the device, given the symbol α, compares the total current to the useful current.

$$\alpha = \frac{I_C}{I_E}$$

This will always be slightly less than one, or, if we were to consider it in percents, less than 100%. (We actually don't ever convert it to percents, but this might help you remember it can't be greater than one.)

Since α and β are based upon the same current values, they are related to each other. If you want, you can do the math to prove these:

$$\alpha = \frac{\beta}{\beta+1}$$

$$\beta = \frac{\alpha}{1-\alpha}$$

Time to play with the mathematical relationships.

A particular NPN transistor has a measured Collector current of 10.896 mA and a measured Emitter current of 11.205 mA.

1. What is the Base current, in microamps? 309 μA
2. What is α, to four decimal places? 0.9724+5E-4
3. What is β? 35.3

A PNP transistor has a current efficiency of 0.9965 and a measured Base current of 560 μA.

4. What is β? 285

5. What would the Collector current be, in milliamps? 159 mA
6. What would the Emitter current be, in milliamps? 160 mA

An NPN transistor has a current gain of 375 and a measured Base current of 220 μA.

7. What is the current efficiency, α, for four decimal places? 0.9973+3E-4
8. What would you expect the Collector current to be. in milliamps? 82.5 mA
9. What would you expect the Emitter current to be? 82.7 mA

Standard Transistor Nomenclature

Voltages

- Any singly-subscripted voltage (e.g. V_C, V_E, V_B) is the voltage at the labelled pin referenced to ground.
- Any doubly-subscripted voltage (e.g. V_{CE}, V_{BE}) is the voltage at the first labelled pin referenced to the second labelled pin. Mathematically, $V_{BE} = V_B - V_E$ and $V_{CE} = V_C - V_E$
- Any doubled subscript (e.g. V_{CC}, V_{BB}, V_{EE}) indicates an external power supply that powers a particular pin, either directly or indirectly. These power supplies are also referenced to ground.

Currents

The currents are labelled according to the pin they are flowing through: I_B, I_C, and I_E. When we draw an arrow on a diagram to indicate our intention in terms of the direction of current flow, we can simply use positive numbers.

However, on a lot of specification sheets, you will see positive and negative currents shown. The convention is this: If the doubly-subscripted voltage driving the current is negative, the current will be negative. So, for example, if V_{CE} is actually negative (V_E is more positive than V_C), then I_C will also be negative.

Since PNP transistors have their Emitters more positive than the Bases and Collectors, their standard V_{BE} and V_{CE} voltages will be negative, and so will be their I_C and I_B currents.

Biasing Resistors

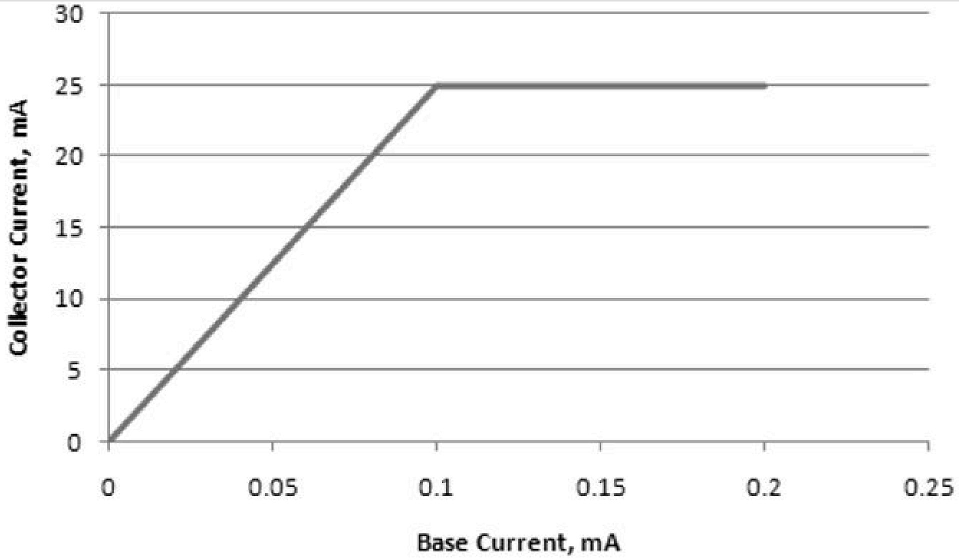
Often, we use letter representations for the resistors that provide current limit or voltage referencing for the pins: R_C, R_E, and R_B. If more than one resistor is used in biasing a pin, they will be designated with a number in the subscript: R_{B1}, R_{B2}, R_{E1}, or R_{E2}.

BJT Modes of Operation

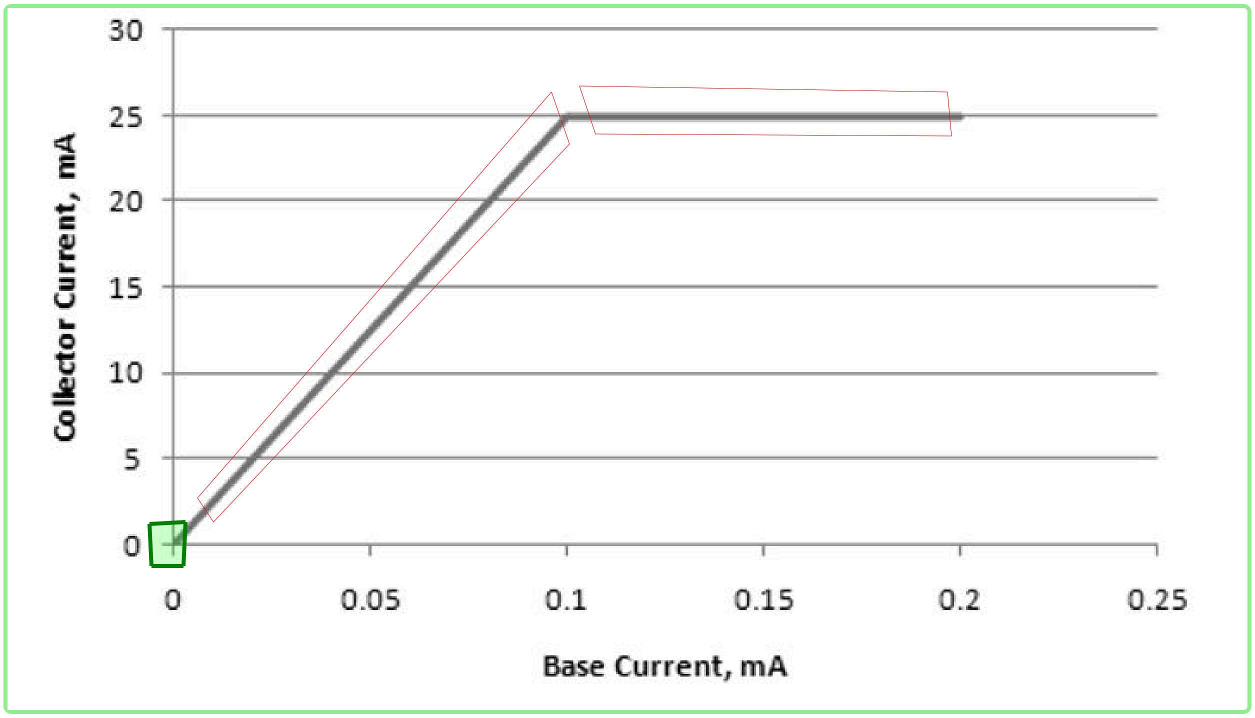
Like diodes, BJTs operate differently depending on the electrical conditions they are exposed to. The following terms describe the four most important conditions for BJTs.

1. ***Cut-off*** -- in this condition, the transistor does not conduct. The B-E junction is not forward biased, so there is no current between the Collector and the Emitter.
2. ***Active Mode*** -- in this condition, an increase in Base current results in an increase in Collector current, in the linear relationship defined by $I_C = \beta I_B$. Consequently, this is also called the ***Linear Active Region***.
3. ***Saturation*** -- In this condition, the transistor has reached the maximum current that can be supplied for the given circuit, and any further increase in Base current has no effect on the Collector current. In this condition, the voltage drop across the transistor from Collector to Emitter (V_{CE}) is reduced very nearly to zero, and is called V_{CEsat}.
4. ***Reverse Breakdown*** -- like any insulator, the reverse-biased C-B junction can be forced to conduct with a high enough potential. Above the Reverse Breakdown voltage, the transistor will conduct with a similar tiny internal resistance as it would have in Saturation, and with significant current flowing as driven by a large voltage, the power will be significant and will likely result in Thermal Destruction.

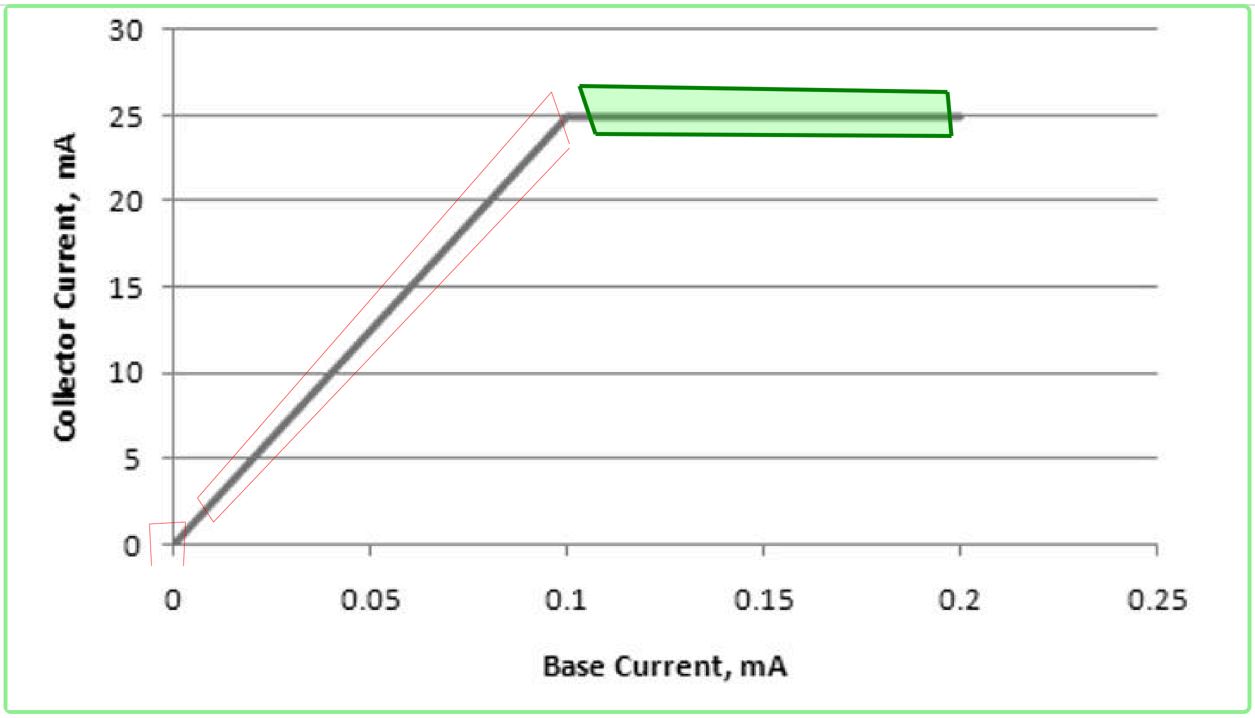
The following graph shows the Transfer Characteristic (relationship between input current and output current) for a particular NPN transistor, and shows the first three of the conditions described above.



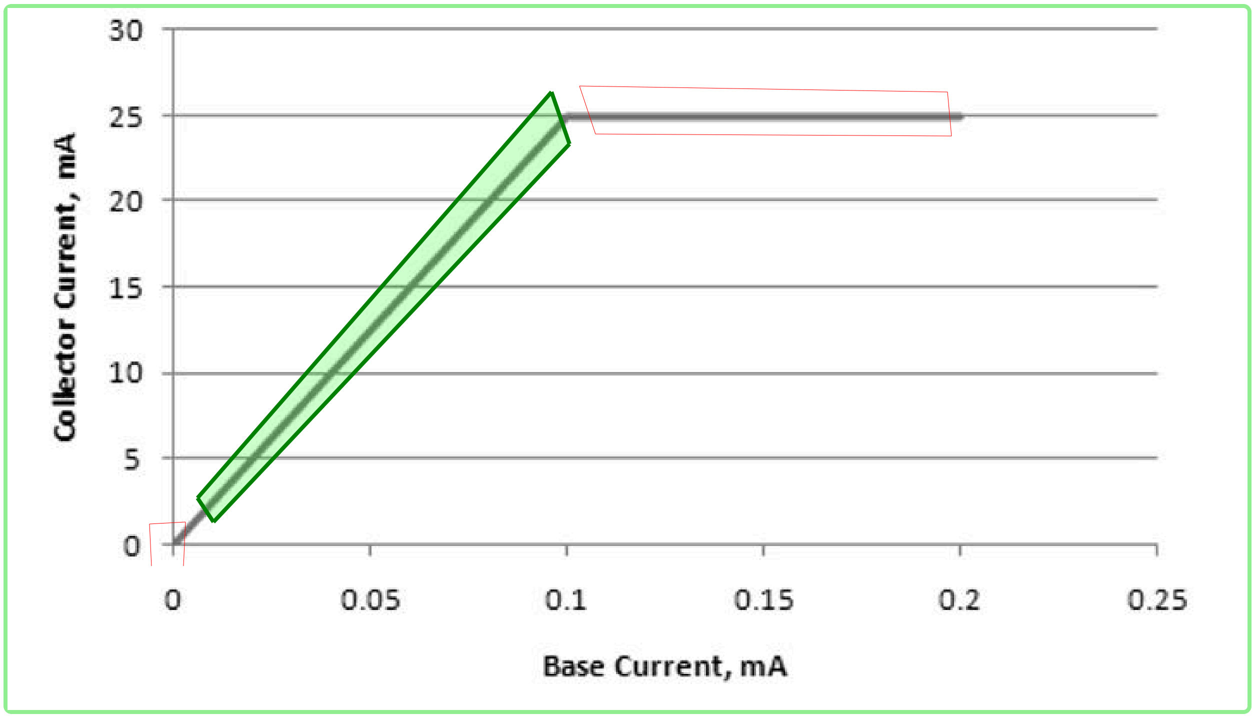
10. Use your mouse to identify the Cut-off Region.



11. Use your mouse to identify the Saturation Region.



12. Use your mouse to identify the Active Region.

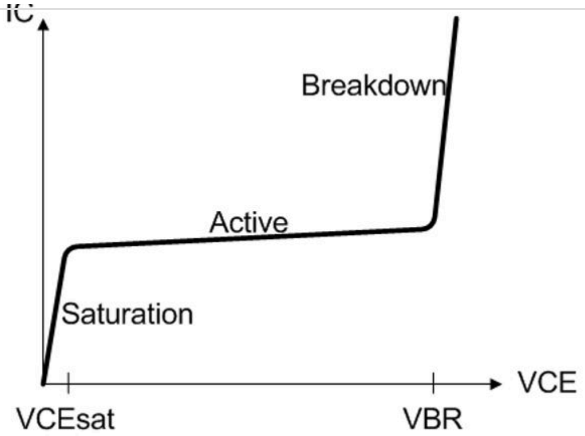


13. What is β in the Active Region? 250

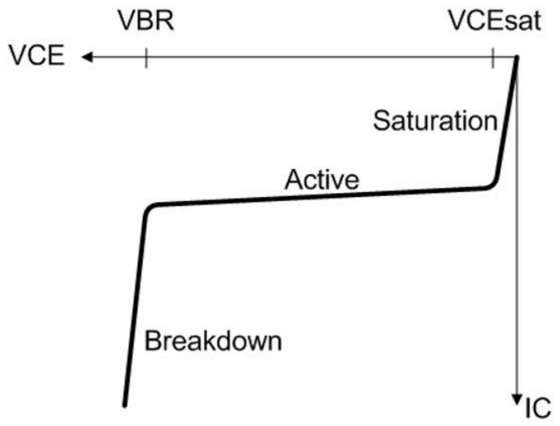
14. Does this value of β apply in the Saturation Region? No -- the output current has reached a maxi…

BJT Characteristic Curves

When testing transistors, a piece of equipment called a ***Curve Tracer*** generates a display indicating how the transistor responds to changes in electrical conditions. The curve tracer applies a constant Base Current, then increases the voltage across the transistor, V_{CE}, to show how the transistor responds. For a single run, the following traces would be typical.



NPN Characteristic Curve



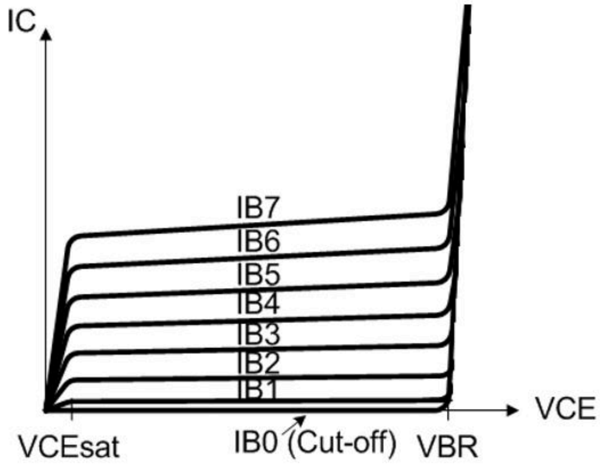
PNP Characteristic Curve

Notice that, for the NPN, V_{CE} and I_C are both positive, whereas for the PNP they are both negative, in keeping with the previous discussion.

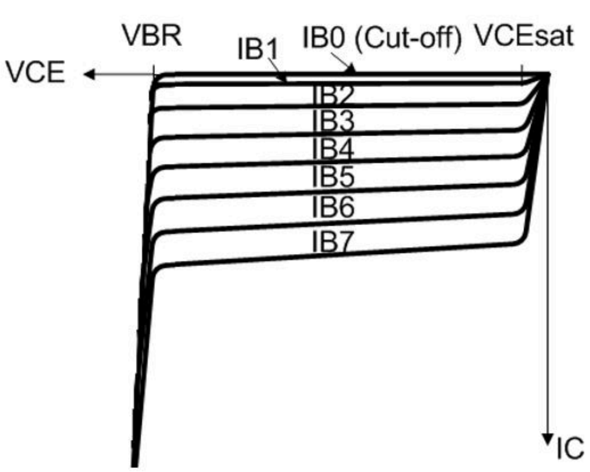
Here are some salient points to be made from these traces:

- In the Active region, I_C is very nearly constant. That's because the Base current is constant in this test, so we would expect $I_C = \beta I_B$. The fact that the line rises slightly as the voltage is raised across the transistor means that there is a large but not infinite internal impedance for this device. In other words, it is very nearly an Ideal Current Source, but not quite, because its current is affected by voltage. The flatter the line, the larger the internal impedance.
- In the very small region prior to entering the Active region, the transistor is allowing whatever current the supply can allow to flow through it. In other words, here it behaves as though it had a very small internal impedance, something you would expect from an Ideal Voltage Source. However, again the line is not straight up as it would be for a truly ideal voltage source. A transistor that allows any available current to pass through it is in Saturation.
- Related to the above discussion, notice that V_{CEsat} is a very small number, usually even smaller than it appears on this diagram.
- Beyond the Reverse Breakdown Voltage, V_{BR}, the transistor again conducts as with a small internal resistance, and the current rises rapidly, very likely leading to Thermal Destruction.

In reality, the Curve Tracer can be set up do draw a number of traces, each for a different constant Base Current. The following shows the "Family of Curves" produced.



NPN Family of Curves



PNP Family of Curves

Here, you can see that as I_B increases, I_C also increases. And, since the spacing between the curves is essentially constant, the relationship between I_B and I_C must be fairly constant.

In addition, we now can see all four of the transistor conditions:

- The I_{B0} curve shows that with no Base current in, there is no Collector current out -- the transistor is not conducting, and is in ***Cut-off***.
- When the transistor is conducting freely with only a tiny voltage drop across it, it is in ***Saturation***.
- Once the transistor begins to act as a Current source, the output is linearly related to the input, so it is in the ***Active Region***.

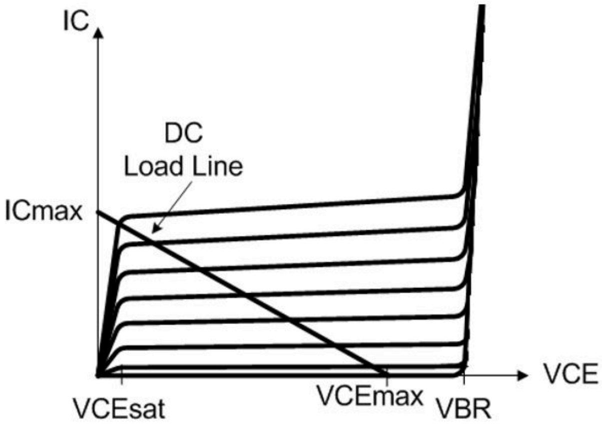
- When the voltage exceeds the transistor's ability to control the current, the transistor enters ***Reverse Breakdown***.

DC Load Line

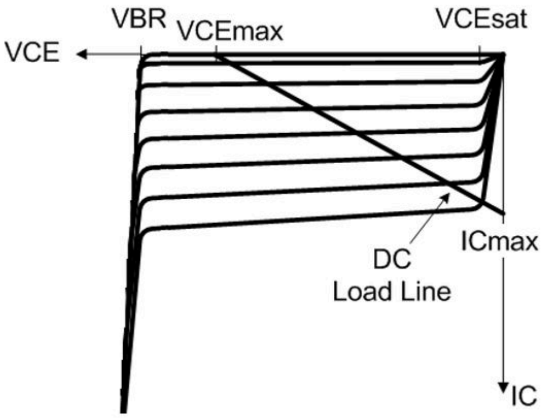
We're going to do one more thing to this set of graphs. As you will soon be learning, the power supplies and resistors used to bias the transistor in a circuit also define limits on what the maximum current through the circuit can be and what the maximum voltage across the transistor can be.

When the transistor is not conducting at all, the maximum Supply Voltage will appear across it, and no current will flow through it. That defines the Cut-off conditions for the circuit.

If we consider the tiny voltage across the transistor as negligible, we can determine the maximum current through the circuit based upon the resistors only, This defines the ideal Saturation conditions for the circuit. And, since the relationship between I_B and I_C is considered to be constant, we can draw a straight line between these two endpoints and call it the DC Load Line. For the circuit as wired, this DC Load Line defines all the possible values for I_C as controlled by I_B and their corresponding values of V_{CE}, or the voltage across the transistor in that circuit.



NPN Curves with DC Load Line



PNP Curves with DC Load Line

Question: In a particular circuit, an NPN transistor is powered by $V_{CC} = 20\ V_{DC}$, and the transistor is biased by a 1.0 kΩ R_C resistor.

15. What is V_{CEmax}? 20 V

16. What is I_{Cmax}? 20 mA

17. When $I_C = 10$ mA, what do you expect V_{CE} to be? 10 V

Now that we've done all the hard work on figuring out how transistors behave, it's time to create some simple models to help us predict how they will respond in different circuits