BJT: Bipolar Junction Transistors

BJT.1 Basic Operation

A bipolar junction transistor is a three-terminal device that, in most logic circuits, acts like a current-controlled switch. If we put a small current into one of the terminals, called the base, then the switch is “on”—current may flow between the other two terminals, called the emitter and the collector. If no current is put into the base, then the switch is “off”—no current flows between the emitter and the collector.

To study the operation of a transistor, we first consider the operation of a pair of diodes connected as shown in Figure BJT-1(a). In this circuit, current can flow from node B to node C or node E, when the appropriate diode is forward biased. However, no current can flow from C to E, or vice versa, since for any choice of voltages on nodes B, C, and E, one or both diodes will be reverse biased. The pn junctions of the two diodes in this circuit are shown in (b).

Now suppose that we fabricate the back-to-back diodes so that they share a common p-type region, as shown in Figure BJT-1(c). The resulting structure is called an npn transistor and has an amazing property. (At least, the physicists working on transistors back in the 1950s thought it was amazing!) If we put current across the base-to-emitter pn junction, then current is also enabled to flow across the collector-to-base np junction (which is normally impossible) and from there to the emitter.

The circuit symbol for the npn transistor is shown in Figure BJT-1(d). Notice that the symbol contains a subtle arrow in the direction of positive current flow. This also reminds us that the base-to-emitter junction is a pn junction, the same as a diode whose symbol has an arrow pointing in the same direction.

Figure BJT-1 Development of an npn transistor: (a) back-to-back diodes; (b) equivalent pn junctions; (c) structure of an npn transistor; (d) npn transistor symbol.
It is also possible to fabricate a *pnp transistor*, as shown in Figure BJT-2. However, *pnp* transistors are seldom used in digital circuits, so we won’t discuss them any further.

The current $I_e$ flowing out of the emitter of an *npn* transistor is the sum of the currents $I_b$ and $I_c$ flowing into the base and the collector. A transistor is often used as a signal *amplifier*, because over a certain operating range (the *active region*) the collector current is equal to a fixed constant times the base current ($I_c = \beta \cdot I_b$). However, in digital circuits, we normally use a transistor as a simple switch that’s always fully “on” or fully “off,” as explained next.

Figure BJT-3 shows the *common-emitter configuration* of an *npn* transistor, which is most often used in digital switching applications. This configuration uses two discrete resistors, $R1$ and $R2$, in addition to a single *npn* transistor. In this circuit, if $V_{IN}$ is 0 or negative, then the base-to-emitter diode junction is reverse biased, and no base current ($I_b$) can flow. If no base current flows, then no collector current ($I_c$) can flow, and the transistor is said to be cut off (OFF).
Since the base-to-emitter junction is a real diode, as opposed to an ideal one, \( V_{IN} \) must reach at least +0.6 V (one diode-drop) before any base current can flow. Once this happens, Ohm’s law tells us that
\[
I_b = \frac{(V_{IN} - 0.6)}{R1}
\]
(We ignore the forward resistance \( R_f \) of the forward-biased base-to-emitter junction, which is usually small compared to the base resistor \( R1 \).) When base current flows, then collector current can flow in an amount proportional to \( I_b \), that is,
\[
I_c = \beta \cdot I_b
\]
The constant of proportionality, \( \beta \), is called the gain of the transistor, and is in the range of 10 to 100 for typical transistors.

Although the base current \( I_b \) controls the collector current flow \( I_c \), it also indirectly controls the voltage \( V_{CE} \) across the collector-to-emitter junction, since \( V_{CE} \) is just the supply voltage \( V_{CC} \) minus the voltage drop across resistor \( R2 \):
\[
V_{CE} = V_{CC} - I_c \cdot R2
\]
\[
= V_{CC} - \beta \cdot I_b \cdot R2
\]
\[
= V_{CC} - \beta \cdot (V_{IN} - 0.6) \cdot R2 / R1
\]

However, in an ideal transistor \( V_{CE} \) can never be less than zero (the transistor cannot just create a negative potential), and in a real transistor \( V_{CE} \) can never be less than \( V_{CE\text{sat}} \), a transistor parameter that is typically about 0.2 V.

If the values of \( V_{IN} \), \( \beta \), \( R1 \), and \( R2 \) are such that the above equation predicts a value of \( V_{CE} \) that is less than \( V_{CE\text{sat}} \), then the transistor cannot be operating in the active region and the equation does not apply. Instead, the transistor is operating in the saturation region, and is said to be saturated (ON). No matter how much current \( I_b \) we put into the base, \( V_{CE} \) cannot drop below \( V_{CE\text{sat}} \), and the collector current \( I_c \) is determined mainly by the load resistor \( R2 \):
\[
I_c = \frac{(V_{CC} - V_{CE\text{sat}})}{(R2 + R_{CE\text{sat}})}
\]
Here, \( R_{CE\text{sat}} \) is the saturation resistance of the transistor. Typically, \( R_{CE\text{sat}} \) is 50 \( \Omega \) or less and is insignificant compared with \( R2 \).

Computer scientists might like to imagine an npn transistor as a device that continuously looks at its environment and executes the program in Table BJT-1 on the next page..
Table BJT-1  A C program that simulates the function of an npn transistor in the common-emitter configuration.

```c
/* Transistor parameters */
#define DIODEDROP 0.6 /* volts */
#define BETA 10
#define VCE_SAT 0.2 /* volts */
#define RCE_SAT 50 /* ohms */

main()
{
    float Vcc, Vin, R1, R2; /* circuit parameters */
    float Ib, Ic, Vce;   /* circuit conditions */

    if (Vin < DIODEDROP) { /* cut off */
        Ib = 0.0;
        Ic = 0.0;
        Vce = Vcc;
    }
    else { /* active or saturated */
        Ib = (Vin - DIODEDROP) / R1;
        if (((Vcc - ((BETA * Ib) * R2)) >= VCE_SAT) { /* active */
            Ic = BETA * Ib;
            Vce = Vcc - (Ic * R2);
        }
        else { /* saturated */
            Vce = VCE_SAT;
            Ic = (Vcc - Vce) / (R2 + RCE_SAT);
        }
    }
}
```
BJT.2 Transistor Logic Inverter

Figure BJT-4 shows that we can make a logic inverter from an npn transistor in the common-emitter configuration. When the input voltage is LOW, the output voltage is HIGH, and vice versa.

In digital switching applications, bipolar transistors are often operated so they are always either cut off or saturated. That is, digital circuits such as the inverter in Figure BJT-4 are designed so that their transistors are always (well, almost always) in one of the states depicted in Figure BJT-5. When the input voltage $V_{IN}$ is LOW, it is low enough that $I_b$ is zero and the transistor is cut off; the collector-emitter junction looks like an open circuit. When $V_{IN}$ is HIGH,

**Figure BJT-4** Transistor inverter: (a) logic symbol; (b) circuit diagram; (c) transfer characteristic.

**Figure BJT-5** Normal states of an npn transistor in a digital switching circuit: (a) transistor symbol and currents; (b) equivalent circuit for a cut-off (OFF) transistor; (c) equivalent circuit for a saturated (ON) transistor.
it is high enough (and $R_1$ is low enough and $\beta$ is high enough) that the transistor will be saturated for any reasonable value of $R_2$; the collector-emitter junction looks almost like a short circuit. Input voltages in the undefined region between LOW and HIGH are not normally encountered, except during transitions. This undefined region corresponds to the noise margin that we discussed with Figure 1-2 on page 8.

Another way to visualize the operation of a transistor inverter is shown in Figure BJT-6. When $V_{IN}$ is HIGH, the transistor switch is closed, and the output terminal is connected to ground, definitely a LOW voltage. When $V_{IN}$ is LOW, the transistor switch is open and the output terminal is pulled to +5 V through a resistor; the output voltage is HIGH unless the output terminal is too heavily loaded (i.e., improperly connected through a low impedance to ground).
BJT.3 Schottky Transistors

When the input of a saturated transistor is changed, the output does not change immediately; it takes extra time, called *storage time*, to come out of saturation. In fact, storage time accounts for a significant portion of the propagation delay in the original TTL logic family.

Storage time can be eliminated and propagation delay can be reduced by ensuring that transistors do not saturate in normal operation. Contemporary TTL logic families do this by placing a *Schottky diode* between the base and collector of each transistor that might saturate, as shown in Figure BJT-7. The resulting transistors, which do not saturate, are called *Schottky-clamped transistors* or *Schottky transistors* for short.

When forward biased, a Schottky diode’s voltage drop is much less than a standard diode’s, 0.25 V vs. 0.6 V. In a standard saturated transistor, the base-to-collector voltage is 0.4 V, as shown in Figure BJT-8(a). In a Schottky transistor, the Schottky diode shunts current from the base into the collector before the transistor goes into saturation, as shown in (b). Figure BJT-9 is the circuit diagram of a simple inverter using a Schottky transistor.

---

**Figure BJT-7**
Schottky-clamped transistor: (a) circuit; (b) symbol.

**Figure BJT-8**
Operation of a transistor with large base current: (a) standard saturated transistor; (b) transistor with Schottky diode to prevent saturation.
Figure BJT-9
Inverter using Schottky transistor.