

3. Transistors

S. Orsi, Université de Genève, October 2009

1 The Transistors

The name transistor comes from the phrase “transferring an electrical signal across a resistor”.

Transistors are four-terminal semiconductor devices commonly used to amplify or switch electronic signals. Often one terminal is not used, therefore we will refer to transistors as **3-terminal** devices. They are *active* devices that can amplify an input signal, i.e. produce an output signal with more power in it than the input signal. The transistor is the fundamental building block of modern electronic devices, and is used in radio, telephone, computer and other electronic systems. The transistor is often cited as being one of the greatest achievements in the 20th century, and some consider it one of the most important technological breakthroughs ever in human history. Most transistors are found in integrated circuits, but in our course we will work with individually packaged transistors. In particular, we will use two types of transistors:

- The **Bipolar Junction Transistor (BJT)** is a current controlled valve. The 3 terminals are named base, collector and emitter. The current flowing through the *base* (I_B) controls the current through the *collector* (I_C).
- The **Field Effect Transistor (FET)** is a voltage controlled valve. The 3 terminals are named gate, drain and source. The gate-source voltage (V_{GS}) controls the drain current (I_D).

Figure 1 shows the schematic drawing of a BJT transistor (left) and of a JFET (a particular type of FET transistor, section 5; right). Both transistors shown are of *npn* type. As rule of thumb to recognize transistors, the arrow is always *pointing in*, i.e. is towards the *n* type region. This is valid for all junctions.

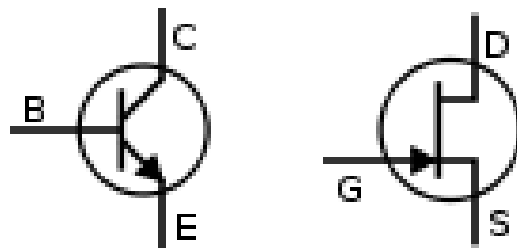


Figure 1: Schematic drawing of a BJT (left) and of a JFET transistors.

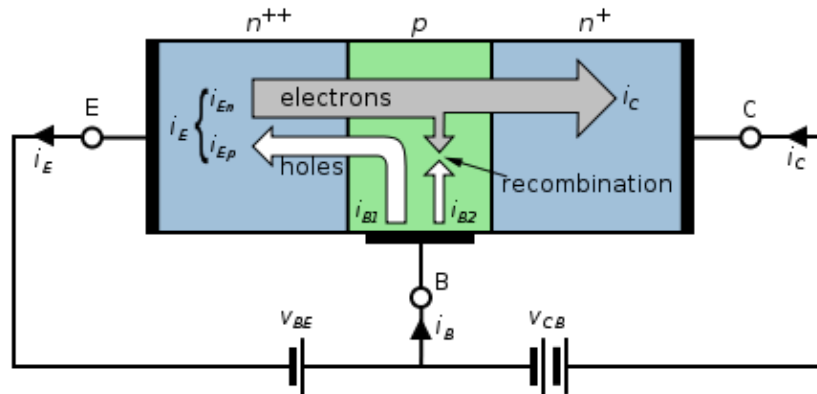


Figure 2: Structure of a BJT.

The transistor is a three terminal device, thus the input and the output must share one terminal in common. This is the origin of the nomenclature of the three types of transistor amplifiers: common collector, common emitter, and common base.

The **gain** is defined as the ratio of the output signal to the input signal. Because transistor amplifiers often have a quiescent output (a non zero output when the input is zero) we define gain as the derivative of the output with respect to the input. For systems where the quiescent output is zero, this reduces to the ratio of the output to the input, therefore the gain is defined as the ratio of the change in output to the change in input, for any given output and input quantities: voltages, currents or power $V_{in}, V_{out}, I_{in}, I_{out}, P_{in}, P_{out}$. Negative gain means that the sign of the signal is inverted, while Power Gain cannot be negative. $|A|$ less than unity indicates that the output is smaller than the input. If the input and output quantities are different, the gain is no longer unitless; the most common examples are transimpedance gain and transadmittance gain.

Clearly a transistor cannot be constructed on the bench by combining two diodes (Why is that?), but an ohmmeter in diode mode can identify the base and the type of transistor. Look at the laboratory instructions to see a drawing of the equivalent circuits.

2 Bipolar Junction Transistors (BJT)

A BJT is a device with two diode junctions, divided into 3 differently doped regions: emitter, base and collector. Depending on the direction of the junctions (i.e. on the doping of each area), BJT may be *npn* or *pnp* type transistors. BJT are called *bipolar* because both holes and electrons are involved in the current flow.

In an *npn* BJT, electrons are the principal charge carriers in the emitter, flowing into the collector, as shown in figure 2. For this reason the emitter is heavily doped (indicated as n^+). The base is slightly doped (p) and made very thin. This allows the recombination current (holes moving from the base and recombining with electrons from the emitter) to be small. Almost all electrons from the emitter will therefore reach the collector (n), and then exit to the external circuit you have built on the 'protoboard'. For *pnp* the description is analogous, inverting electrons/holes and p/n doping. Figure 3 shows the

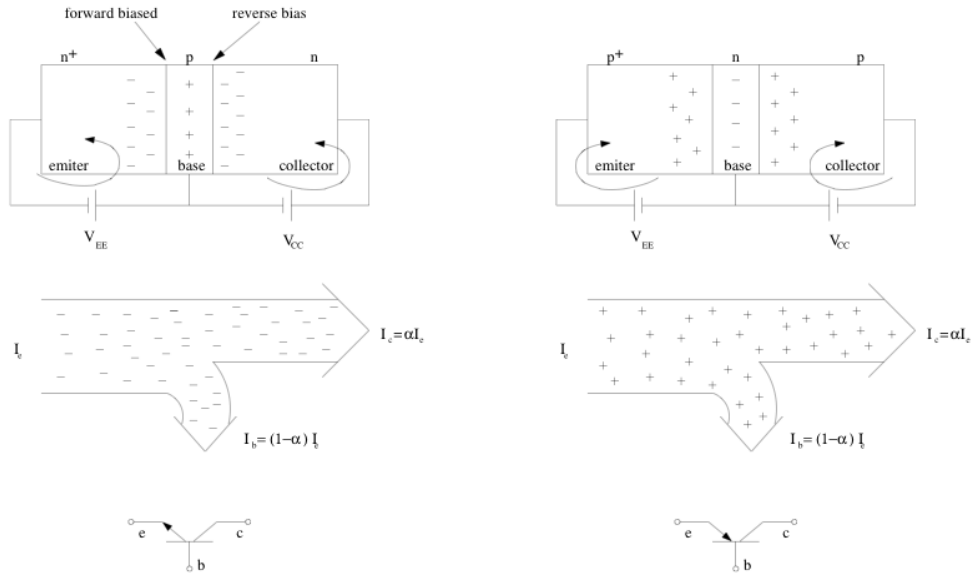


Figure 3: Charge carriers flows inside a BJT: *npn* on the left, *pnp* on the right.

flow of charge carriers inside a BJT.

The **transport factor** of a BJT is defined as $\alpha = I_C/I_E$ (the “transistor effect”), and the **current gain** is given by $\beta = I_C/I_B = \alpha/(1 - \alpha)$.

There are 3 regions of operation, as shown in figure 4:

- **Cut-off region:** The transistor is off. There is no conduction between the collector and the emitter: $I_B=0$ therefore $I_C=0$;
- **Active region:** The transistor is on. The collector current is proportional to and controlled by the base current ($I_C=\beta I_B$) and relatively insensitive to V_{CE} . In this region the transistor can be an amplifier.
- **Saturation region:** The transistor is on. The collector current varies very little with a change in the base current in the saturation region. The voltage V_{CE} is small, a few tenths of volt: it is below the threshold. The collector current is strongly dependent on V_{CE} unlike in the active region. It is desirable to operate transistor switches in or near the saturation region when in their ‘on-state’.

In circuits with nonlinear elements such as a transistor, the **input impedance** of a BJT is defined as the reciprocal of the slope of the graph $V-I$. This is simply the derivative of V_{in} with respect to I_{in} :

$$Z_{in} = \frac{dV_{in}}{dI_{in}} \quad (1)$$

It can be calculated that (see figure 5, left):

$$Z_{in} = (\beta + 1)R_E \simeq \beta R_E. \quad (2)$$

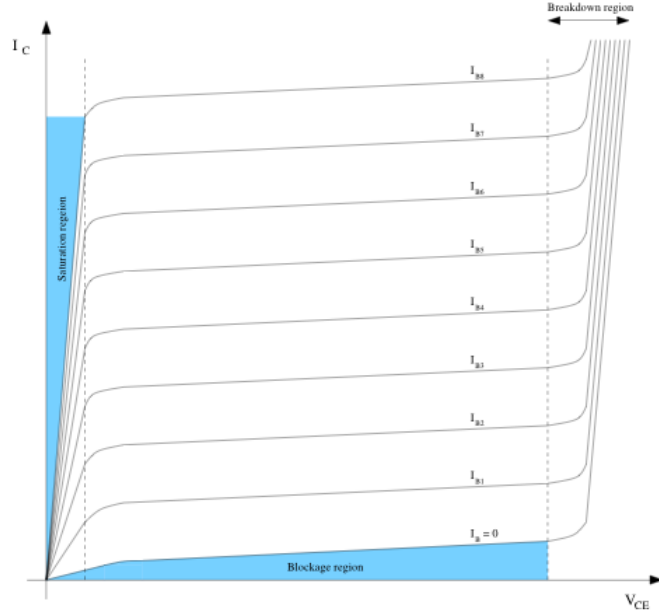


Figure 4: Characteristic curves of a BJT, where $I_{B1} < I_{B2} < \dots$; the large region in the center of the graph is the active region. We want the transistor to work far away from the breakdown region.

The **output impedance** of a transistor for the emitter follower (common collector) can be calculated from figure 5, right. It is found that the impedance of the source, as viewed by the load, is reduced by a factor β^{-1} :

$$Z_{out} = \frac{R_S}{\beta + 1} \simeq \frac{R_S}{\beta}. \quad (3)$$

3 Field Effect Transistors (FET)

The field-effect transistor (FET) relies on an electric field to control the shape and hence the conductivity of a **channel** of one type of charge carrier in a semiconductor material. FETs are sometimes called unipolar transistors to contrast their single-carrier-type operation with the dual-carrier-type operation of bipolar junction transistors (BJT).

All FETs have a *gate*, *drain*, and *source* terminal that correspond roughly to the *base*, *collector*, and *emitter* of BJTs. With the exception of JFET, all FETs also have a fourth terminal called the *body* (or also *base*, *bulk*, or *substrate*), which serves to bias the transistor into operation. The body terminal is normally used only in complicated circuit designs; its presence is important in integrated circuits, but not in our laboratory.

No current flows through the gate electrode, thus the gate is essentially insulated from the source-drain channel. Because no current flows through the gate, the input impedance of the FET is extremely large (in the range of 10^{10} – $10^{15}\Omega$). The large input impedance of the FET makes them an excellent choice for amplifier inputs.

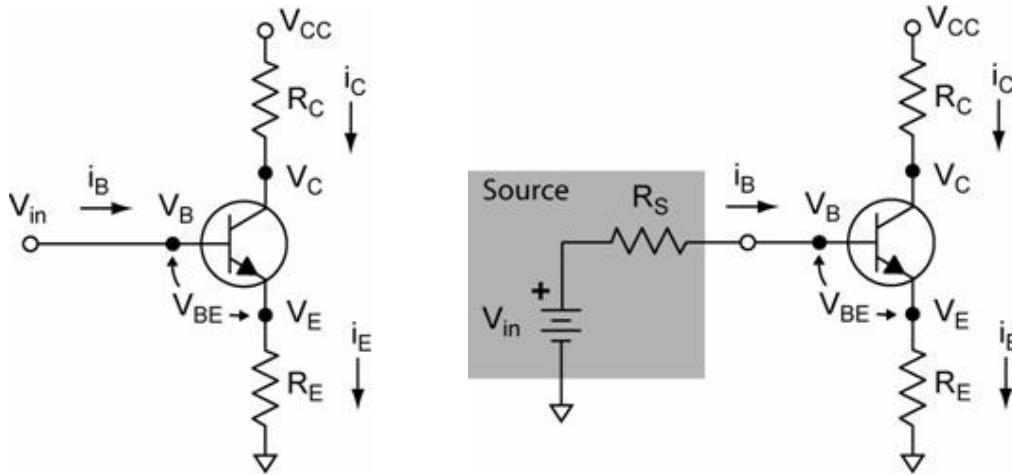


Figure 5: Left: Setup to calculate the input impedance of a BJT. Right: example of an emitter follower (common collector).

The two common families of FETs, the junction FET (JFET) and the metal oxide semiconductor FET (MOSFET) differ in the way the gate contact is made on the source-drain channel. In the JFET the gate-channel contact is a reverse biased pn junction. The gate-channel junction of the JFET must always be reverse biased otherwise it may behave as a diode. All JFETs are depletion mode devices, i.e. they are on when the gate bias is zero ($V_{GS}=0$).

In the MOSFET the gate-channel contact is a metal electrode separated from the channel by a thin layer of insulating oxide. MOSFETs have very good isolation between the gate and the channel, but the thin oxide is easily damaged (punctured!) by static discharge through careless handling. MOSFETs are made in both depletion mode (on with zero biased gate, $V_{GS} = 0$) and in enhancement mode (off with zero biased gate).

4 MOSFET

The metaloxidesemiconductor field-effect transistor (MOSFET) is by far the most common transistor in both digital and analog circuits. Figure 6 shows the structure of a MOSFET.

If the MOSFET is an n -channel or nMOS FET, then the source and drain are $n+$ regions and the body is a p region. With sufficient gate voltage, above a threshold voltage value, electrons from the source (and possibly also the drain) enter the inversion layer or n -channel at the interface between the p region and the oxide. This conducting channel extends between the source and the drain, and current is conducted through it when a voltage is applied between source and drain.

For gate voltages below the threshold value, the channel is lightly populated, and only a very small subthreshold leakage current can flow between the source and the drain.

The operation of a MOSFET can be separated into three different modes, depending on the voltages at the terminals. In particular, for an enhancement-mode, n -channel MOSFET, the three operational modes are:

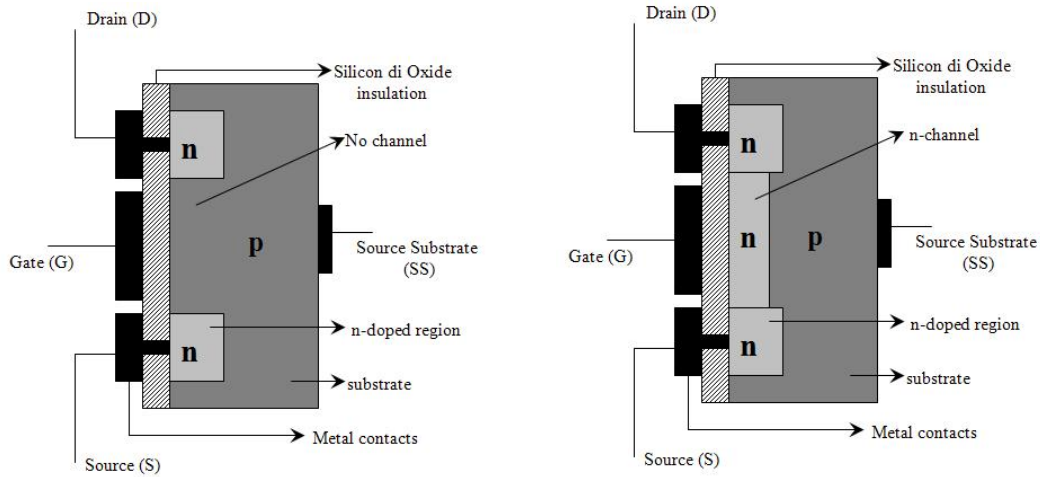


Figure 6: Structure of the MOSFET transistor without (left) and with (right) an active channel.

- **Cutoff:** for $V_{GS} < V_{th}$. The transistor is turned off, and there is no conduction between drain and source. In reality, there is a weak-inversion current I_D , sometimes called subthreshold leakage.
- **Triode mode** (or linear or ohmic mode): for $V_{GS} > V_{th}$ and $V_{DS} < (V_{GS} - V_{th})$. The transistor is turned on, and a channel has been created.
- **Saturation** (or active mode): for $V_{GS} > V_{th}$ and $V_{DS} > (V_{GS} - V_{th})$. The switch is turned on, and a channel has been created, which allows current to flow between the drain and source. Since the drain voltage is higher than the gate voltage, the electrons spread out, and conduction is not through a narrow channel but through a broader, two- or three-dimensional current distribution extending away from the interface and deeper in the substrate. The onset of this region is also known as pinch-off to indicate the lack of channel region near the drain.

Figure 7 shows a summary of the regions of operation.

The **body effect** describes the changes in the threshold voltage by the change in the source-bulk voltage.

5 JFET

In the JFET the gate-channel contact is a reverse biased pn junction. The gate-channel junction of the JFET must always be reverse biased otherwise it may behave as a diode. All JFETs are depletion mode devices they are on when the gate bias is zero ($V_{GS} = 0$). Regions of operation:

- **Cut-off region:** The transistor is off. There is no conduction between the drain and the source when the gate-source voltage is greater than the cut-off voltage. ($I_D = 0$ for $V_{GS} > V_{GS,off}$)

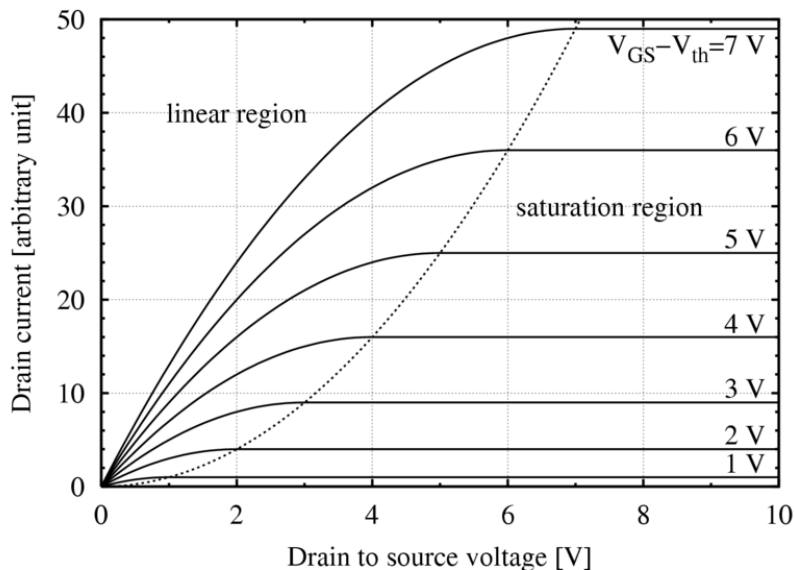


Figure 7: Regions of operation for a MOSFET. The cut-off region has $I_D \sim 0$.

- Active region (also called the Saturation region): The transistor is on. The drain current is controlled by the gate-source voltage (V_{GS}) and relatively insensitive to V_{DS} . In this region the transistor can be an amplifier.

$$I_D = I_{DSS}(1 - V_{GS}/V_{GS,off})^2 \quad (4)$$

- Ohmic region: The transistor is on, but behaves as a voltage controlled resistor. When V_{DS} is less than in the active region, the drain current is roughly proportional to the source-drain voltage and is controlled by the gate voltage.

$$I_D = I_{DSS}[(1 - V_{GS}/V_{GS,off})\frac{V_{DS}}{-V_{GS,off}} - (\frac{V_{DS}}{V_{GS,off}})^2] \quad (5)$$

where I_{DSS} is the drain current in the active region for $V_{GS} = 0$ (i.e. I_D source shorted to gate), and $V_{GS,off}$ is the minimum V_{GS} where $I_D = 0$; $V_{GS,off}$ is negative for n -channel and positive for p -channel.

6 Darlington pair

The Darlington transistor, also called a *Darlington pair*, is a compound structure consisting of two bipolar transistors connected in such a way that the current amplified by the first transistor is amplified further by the second one, as shown in figure 8. This configuration gives a much higher current gain (called β or h_{FE}) than each transistor taken separately and, in the case of integrated devices, can take less space than two individual transistors because they can use a shared collector.

A similar configuration but with transistors of opposite type (nnp and $pnnp$) is the **Sziklai pair**, also called the “complementary Darlington”.

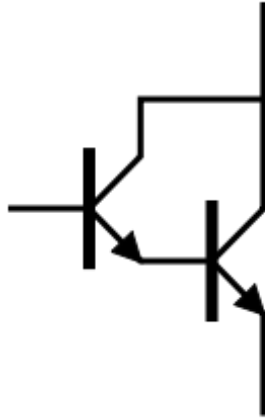


Figure 8: A Darlington pair.

A Darlington pair behaves like a single transistor with a high current gain (approximately the product of the gains of the two transistors).

A general relation between the global current gain and the individual gains is given by:

$$\beta_{\text{Darlington}} = \beta_1 \cdot \beta_2 + \beta_1 + \beta_2 \approx \beta_1 \cdot \beta_2 \quad (6)$$

where the approximation is valid if β_1 and β_2 are high enough (hundreds). A typical modern device has a current gain of 1000 or more, so that only a small base current is needed to make the pair switch on.

However, this high current gain comes with several drawbacks:

- The base-emitter voltage is approximately doubled, since there are two pn junctions between the base and emitter of the Darlington transistor.
- The saturation voltage is increased. The output transistor is not allowed to saturate (i.e. its base-collector junction must remain reverse-biased) because its collector-emitter voltage is now equal to the sum of its own base-emitter voltage and the collector-emitter voltage of the first transistor, both positive quantities in normal operation; in symbols, $V_{CE2} = V_{BE2} + V_{CE1}$, so $V_{C2} > V_{B2}$ always. The saturation voltage of a Darlington transistor is one V_{BE} (about 0.65 V in silicon) higher than a single transistor saturation voltage, which is typically 0.1 - 0.2 V in silicon. For equal collector currents, this drawback translates to an increase in the dissipated power for the Darlington transistor over a single transistor.
- The switching speed is reduced, because the first transistor cannot actively inhibit the base current of the second one, making the device slow to switch off.
- The Darlington pair has more phase shift at high frequencies than a single transistor and hence can more easily become unstable with negative feedback (i.e., systems that use this configuration can have poor phase margin due to the extra transistor delay).

Darlington pairs are available as integrated packages or can be made from two discrete transistors, as we do in the laboratory; Q1 (the left-hand transistor in the diagram) can be a low power type, but normally Q2 (on the right) will need to be high power.

References

- [1] E. Cortina, notes on 'Circuits with transistors', Univ. Genève, 2007
- [2] <http://en.wikipedia.org/wiki/MOSFET>
- [3] Horowitz, The art of electronics
- [4] http://www.nhn.ou.edu/~bumm/ELAB/Lect_Notes/BJT_FET_transistors_v1_1.html
- [5] http://en.wikipedia.org/wiki/Darlington_transistor