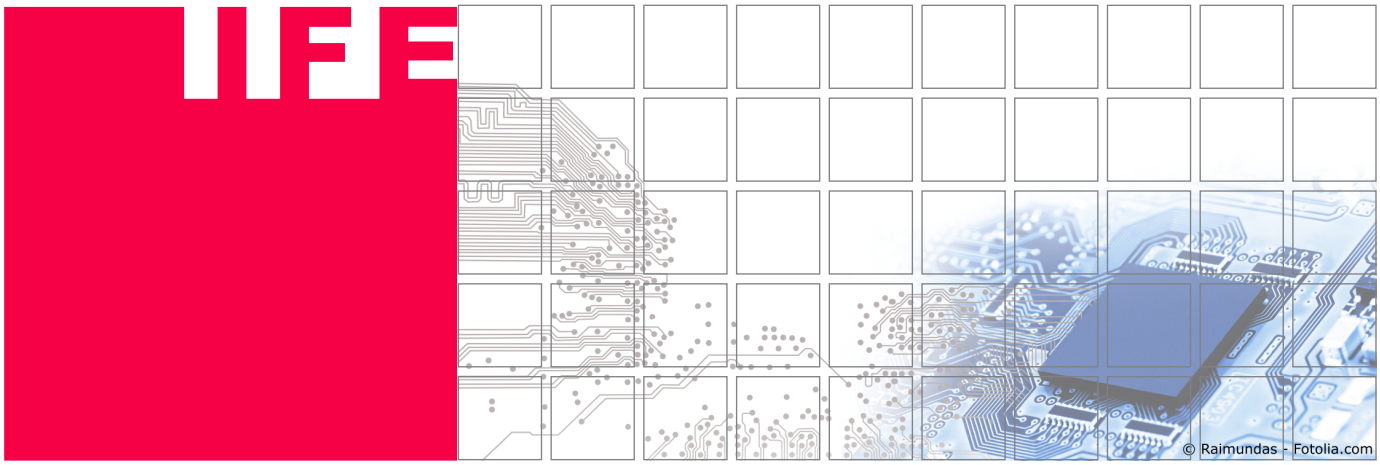
Institute of Electronics



Physics of a Bipolar Junction Transistor

Calculation of the Current-Voltage Relation

Patrick SCHREY

January 27, 2020

Lecture Notes

Contents

1 What is a bipolar junction transistor?

Remember the diode, which is a single p-n junction? According to the well known saying ”The more, the merrier“, let’s add another junction to a diode. We immediately see that we can add this second junction to either side of the diode. We could add a n-doped region to the p-doped side of the diode, or a p-doped region to the n-doped side of the diode. The two resulting devices are called Bipolar Junction Transistor (BJT). They are further distinguished according to their sequence in doping and referred to as npn or pnp transistors. The npn and pnp transistors are dual devices, i.e. they are the same but also the opposite of each other.
A cross-section of a npn and pnp can be seen in Figure 1.

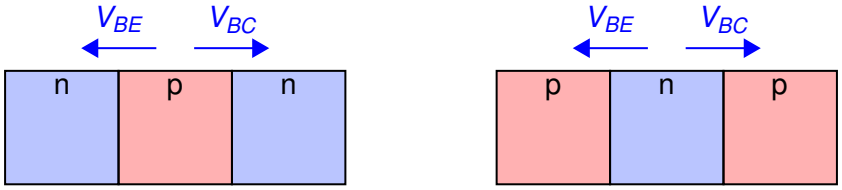


Figure 1: Simplified cross-section of the two principal types of BJT.
Left: A npn BJT. Right: A pnp BJT.

The first working bipolar junction transistor was set into operation around Christmas eve in 1947. The transistor was made of a slab of germanium contacted by two point contacts from the top. These point contacts were made by a gold plated triangular plastic wedge. With a razor edge, the gold was cut at the lower tip of the plastic wedge to form two closely spaced contacts. A simplified representation of the setup is shown on the left-hand side of Figure 2.

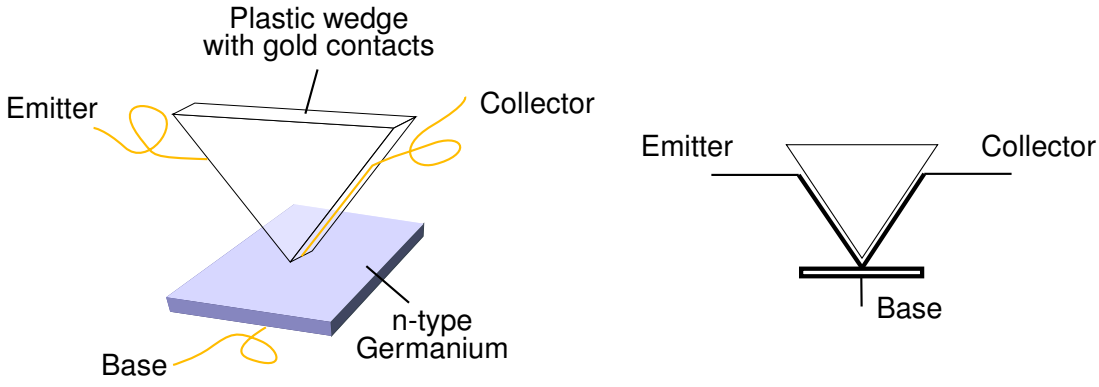


Figure 2: The first bipolar junction transistor was formed by a triangular plastic wedge pressing two gold point contacts on a slab of n-type germanium. Left: Simplified representation of original experimental setup. Right: In one of the engineers notebooks, an electric circuit diagram showed a schematic picture of the setup. This picture is very similar to the symbol for a BJT used today.

The base formed the mechanical base for the whole transistor, hence the name. The emitter emits charge carriers into the base, while the collector collects them back up again.

2 How does a bipolar junction transistor work?

At a first glance, the BJT looks like two diodes in an anti-serial connection as shown in Figure 3.

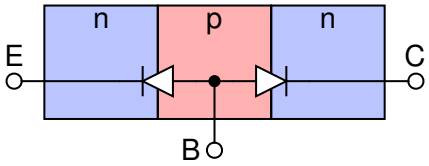


Figure 3: Simplified cross-section of a npn BJT showing the anti-serial diodes.

This arrangement does not seem to be able to conduct current between collector and emitter at all. In fact, without any current into the base, no current flows between collector and emitter. But once we apply a small base current, the transistor starts conducting. How can this be?
Well, the representation above isn't quite appropriate to understand why a transistor behaves the way it does. A better representation is offered by the depletion layers in the BJT as shown in Figure 4.

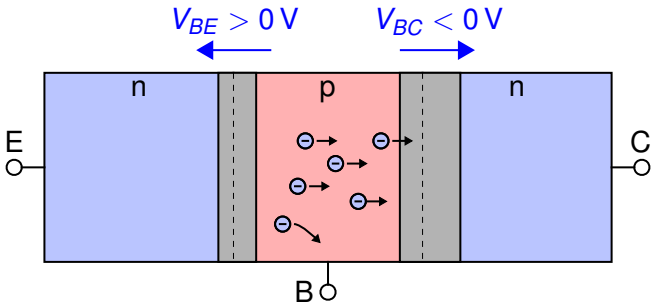


Figure 4: Two depletion layers are inside a BJT, one at the base-emitter junction and one at the base-collector junction.

Typically, the base-emitter junction is forward-biased and the base-collector junction is reverse-biased. Electrons are injected from the emitter into the base. There they diffuse towards the collector. At the base-collector depletion layer, the in-built electric field forces the electrons to move from the base to the collector. Thus, the electrons travelled from the emitter to the collector. Only a small fraction of electron leave the base via the base terminal.
But this is only a qualitative description of the BJTs operation. In order to get a formula for the current-voltage relation, we have to do some math.

2.1 Charge carrier concentration

The most crucial region in a BJT is the base. Each and every current in a BJT has to flow through the base. The current in the base is governed by the minority charge carrier density. The upcoming discussion will focus on a npn BJT but the same calculations can be performed for a pnp BJT.

We define the coordinate x inside the base of the BJT as shown in Figure 5.

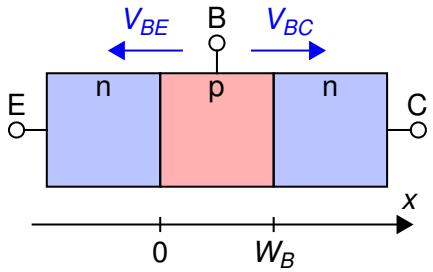


Figure 5: For the calculation of the current-voltage relation, we are mainly interested in the base. The position x starts at the left-hand side. At the right-hand side, x is equal to the base width W_B.

The minority charge carrier concentration in the p-doped base of a npn BJT is n_p and this is what we are after. I think you can imagine what's coming now. Behold the dreaded continuity equation!

$$\frac{\partial n_p}{\partial t} = D_n \cdot \frac{\partial^2 n_p}{\partial x^2} + \mu_n \cdot \vec{\mathbf{E}} \cdot \frac{\partial n_p}{\partial x} + G - R \tag{1}$$

Calculating the current through a BJT is very similar to the one through a diode, which is to be expected as we built the BJT from a diode.

Like we did for the diode, we restrict the continuity equation to one dimension x. Furthermore, a few assumptions are necessary to simplify the problem. Units will be omitted throughout the calculation.

We are mainly interested in the static current-voltage relation, i.e. we set $\frac{\partial n_p}{\partial t} = 0$.

The entire electric field is assumed to be located in the depletion layers. The p- and n-doped regions are free of an electric field $\vec{\mathbf{E}} = 0$.

The number of generated charge carriers in the diode are considered to be negligible, i.e. the generation rate $G = 0$.

Last but not least, the recombination rate is assumed to be proportional to the concentration of excess charge carrier. Excess charge carriers are there in addition to the ones in thermal equilibrium. The total concentration is n_p and the thermal equilibrium concentration is n_{p0}. This results in an excess charge carrier concentration of $n_p - n_{p0}$. Every charge carrier recombines after its life time τ_n has passed. Thus, the recombination rate becomes $R = \frac{n_p - n_{p0}}{\tau_n}$.

Summarizing the assumptions:

- Steady state $\frac{\partial n_p}{\partial t} = 0$
- No electric field in the p-doped region $\vec{\mathbf{E}} = 0$
- No generation $G = 0$
- Recombination rate equals $R = \frac{n_p - n_{p0}}{\tau_n}$

The above assumptions simplify the continuity equation in Equation 1 on the previous page to

$$D_n \cdot \frac{d^2 n_p}{dx^2} - \frac{n_p - n_{p0}}{\tau_n} = 0 \tag{2}$$

$$D_n \cdot \frac{d^2 n_p}{dx^2} - \frac{n_p}{\tau_n} = -\frac{n_{p0}}{\tau_n} \tag{3}$$

This inhomogeneous differential equation solved exactly like we did for the diode. And our inhomogeneous solution is

$$n_p = k_1 \cdot e^{+\frac{x}{L_n}} + k_2 \cdot e^{-\frac{x}{L_n}} + n_{p0} \tag{4}$$

This function for n_p solves the continuity equation. But there are two missing pieces. What about k_1 and k_2?
To uniquely determine the charge carrier concentration, we have to use boundary conditions. And in this boundary conditions, the BJT differs from the diode.

2.1.1 Boundary conditions

Boundary conditions allow to uniquely define the charge carrier concentration. As we have still two unknown quantities k_1 and k_2, we will need two boundary conditions.
For the diode, there was only one side of the p- and n-doped region facing the depletion layer. In the base of a BJT, there is a depletion layer on both sides. We can re-use the boundary condition at the depletion layer interface of the diode. The boundary conditions for the base become

$$n_p(x = 0) = n_{p0} \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}}$$
$$n_p(x = W_B) = n_{p0} \cdot e^{\frac{q \cdot V_{BC}}{k \cdot T}}$$

Applying the boundary conditions to the inhomogeneous solution in Equation 4 and calculating the charge carrier densities yields

$$n_p(x) = \frac{n_{p0}}{2 \cdot \sinh\left(\frac{W_B}{L_n}\right)} \cdot \left\{ \left(e^{\frac{q \cdot V_{BE}}{k \cdot T}} - 1\right) \cdot \left(e^{\frac{W_B - x}{L_n}} - e^{\frac{x - W_B}{L_n}}\right) + \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - 1\right) \cdot \left(e^{\frac{x}{L_n}} - e^{-\frac{x}{L_n}}\right) \right\} + n_{p0} \tag{5}$$

Where sinh() is the hyperbolic sine function.

2.2 Current densities and currents

To determine the current densities at the base-emitter and base-collector junction, we use the drift diffusion equation.

$$\vec{J}_n = q \cdot n_p \cdot \mu_n \cdot \vec{\mathbf{E}} + q \cdot D_n \cdot \nabla n_p \tag{6}$$

As the electric field is present in the depletion layers only, the charge transport through the base region is governed by diffusion. To get the emitter current density, we calculate the diffusion current density at $x = 0$.

$$\begin{aligned} \vec{J}_{nE} &= q \cdot D_n \cdot \left.\frac{dn_p}{dx}\right|_{x=0} \\ \vec{J}_{nE} &= -\frac{q \cdot D_n \cdot n_{p0}}{L_n} \cdot \coth\left(\frac{W_B}{L_n}\right) \cdot \left[\left(e^{\frac{q \cdot V_{BE}}{k \cdot T}} - 1\right) - \operatorname{sech}\left(\frac{W_B}{L_n}\right) \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - 1\right)\right] \end{aligned} \tag{7}$$

Where coth() is the hyperbolic cotangent function and sech() the hyperbolic secant function.
To get the collector current density, we calculate the diffusion current density at $x = W_B$.

$$\begin{aligned} \vec{J}_{nC} &= q \cdot D_n \cdot \left.\frac{dn_p}{dx}\right|_{x=W_B} \\ \vec{J}_{nC} &= -\frac{q \cdot D_n \cdot n_{p0}}{L_n} \cdot \operatorname{cosech}\left(\frac{W_B}{L_n}\right) \cdot \left[\left(e^{\frac{q \cdot V_{BE}}{k \cdot T}} - 1\right) - \cosh\left(\frac{W_B}{L_n}\right) \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - 1\right)\right] \end{aligned} \tag{8}$$

Where cosech() is the hyperbolic cosecant function and cosh() is the hyperbolic cosine function.
Both current densities are negative. This makes sense, as we expect the current to flow from collector to emitter, but x in Figure 5 on page 5 points into the opposite direction.
To get the emitter and collector currents, I_E and I_C, respectively, we have to multiply with the cross-section area of the individual junctions. For the sake of simplicity, we assume that both junctions have the same area A. Also, we define in which direction a current passing through the junctions is counted positive. This is done by defining a normal vector $\vec{e}_n$ in the opposite x-direction, i.e. pointing from collector to emitter. The direction of the different vectors is emphasized in Figure 6 on the following page.

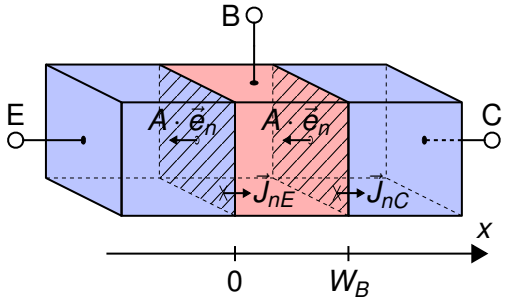


Figure 6: The calculated current densities $\vec{J}_{nE}$ and $\vec{J}_{nC}$ point in x direction. The normal vector $\vec{e}_n$ of the two junctions point in the opposite direction.

Calculating the currents gives

$$I_E = \vec{J}_{nE} \cdot A \cdot \vec{e}_n$$

$$I_E = \frac{q \cdot A \cdot D_n \cdot n_{p0}}{L_n} \cdot \coth\left(\frac{W_B}{L_n}\right) \cdot \left[\left(e^{\frac{q \cdot V_{BE}}{k \cdot T}} - 1\right) - \operatorname{sech}\left(\frac{W_B}{L_n}\right) \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - 1\right)\right]$$

$$I_C = \vec{J}_{nC} \cdot A \cdot \vec{e}_n$$

$$I_C = \frac{q \cdot A \cdot D_n \cdot n_{p0}}{L_n} \cdot \operatorname{cosech}\left(\frac{W_B}{L_n}\right) \cdot \left[\left(e^{\frac{q \cdot V_{BE}}{k \cdot T}} - 1\right) - \cosh\left(\frac{W_B}{L_n}\right) \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - 1\right)\right]$$

These are some pretty complicated equations. But there are a few assumptions that will make things easier. First, we assume a reverse biased base-collector junction $V_{BC} < 0\,\text{V}$, i.e. the collector is at a higher potential than the base. Second, we assume a forward biased base-emitter junction $V_{BE} > 0\,\text{V}$. This means the exponential with V_{BE} becomes very large and we can neglect the -1. For both currents I_E and I_C, the term in squared brackets can be approximated by

$$\left[...\right] \approx e^{\frac{q \cdot V_{BE}}{k \cdot T}}$$

The emitter and collector currents become

$$I_E \approx \frac{q \cdot A \cdot D_n \cdot n_{p0}}{L_n} \cdot \coth\left(\frac{W_B}{L_n}\right) \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}}$$

$$I_C \approx \frac{q \cdot A \cdot D_n \cdot n_{p0}}{L_n} \cdot \operatorname{cosech}\left(\frac{W_B}{L_n}\right) \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}}$$

If the base width W_B is much larger than the diffusion length L_n, the electrons no longer reach the depletion layer at the base-collector junction and simply recombine in the base region. One could say that “the junctions are no longer able to communicate with each other”. This is by the way the reason why two back-to-back diodes do not form a BJT, as the parasitic diodes in Figure 3 on page 4 would suggest.

For base width W_B smaller than the diffusion length L_n, the hyperbolic cotangent and cosecant functions approach each other. This means the emitter and collector current become approximately the same $I_E \approx I_C$ and the base current is negligible $I_B \approx 0\,\text{A}$.

For very small W_B, the recombination in the base can be neglected. This means we can re-write the differential equation from Equation 2 on page 6 to a much simpler form.

$$D_n \cdot \frac{d^2 n_p}{dx^2} = 0$$

This means we are looking for a function $n_p(x)$, which has a vanishing second derivative. The simplest function with this property is a linear function. So we choose the ansatz:

$$n_p(x) = k_1 \cdot x + k_2 \tag{9}$$

The boundary conditions are the same as before.

$$n_p(x = 0) = n_{p0} \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}}$$
$$n_p(x = W_B) = n_{p0} \cdot e^{\frac{q \cdot V_{BC}}{k \cdot T}}$$

This results in a charge carrier concentration of

$$n_p(x) = n_{p0} \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - e^{\frac{q \cdot V_{BE}}{k \cdot T}} \right) \cdot \frac{x}{W_B} + n_{p0} \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}}$$

Again, we use the drift-diffusion equation to get the current densities $\vec{J}_{nE}$ and $\vec{J}_{nC}$.

$$\vec{J}_{nE} = \frac{q \cdot D_n \cdot n_{p0}}{W_B} \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - e^{\frac{q \cdot V_{BE}}{k \cdot T}} \right)$$
$$\vec{J}_{nC} = \frac{q \cdot D_n \cdot n_{p0}}{W_B} \cdot \left(e^{\frac{q \cdot V_{BC}}{k \cdot T}} - e^{\frac{q \cdot V_{BE}}{k \cdot T}} \right)$$

Unsurprisingly, the currents are the same as we neglected recombination in the base region. We again assume a reverse-biased base-collector junction, i.e. $V_{BC} < 0\,\text{V}$. Thus, the exponential with V_{BC} becomes very small and can be neglected.

$$\vec{J}_{nE} = \vec{J}_{nC} \approx -\frac{q \cdot D_n \cdot n_{p0}}{W_B} \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}} \tag{10}$$

We get the current by multiplying with the junction area A. As shown in Figure 6 on page 8, we define the current flow direction by the areas normal vector $\vec{e}_n$. Thus, the current through the BJT becomes

$$I_C = \frac{q \cdot A \cdot D_n \cdot n_{p0}}{W_B} \cdot e^{\frac{q \cdot V_{BE}}{k \cdot T}} \tag{11}$$

To make the equation even more elegant, we can now define the saturation current I_S

$$I_S = \frac{q \cdot A \cdot D_n \cdot n_{p0}}{W_B} \tag{12}$$

and the thermal voltage V_T.

$$V_T = \frac{k \cdot T}{q} \tag{13}$$

Finally, we are able to write the collector current as

$$I_C = I_S \cdot e^{\frac{V_{BE}}{V_T}} \tag{14}$$

This results in the same equation for a BJT operating in the linear regime as can be found in text books on electronic circuit design [1, 3].

References

[1] H. Hartl, E. Krasser, W. Pribyl, P. Söser, G. Winkler. *Elektronische Schaltungstechnik: mit Beispielen in PSpice*. Pearson Deutschland GmbH, 2008.

[2] S. M. Sze, K. K. Ng. *Physics of Semiconductor Devices*. 3rd ed. John Wiley & Sons, 2007.

[3] U. Tietze, C. Schenk. *Halbleiter-Schaltungstechnik*. 12th ed. Springer, 2002.

Legal Notice