Contributions to the Demythisation of the Bipolar Junction Transistor

Dragoş NICOLAU

Institutul Național de Cercetare-Dezvoltare în Informatică – ICI București
B-dul Mareșal Alexandru Averescu nr. 8-10, 011455, București, România
dragos.nicolau@ici.ro

Abstract: The Bipolar Junction Transistor (further referred as BJT) remains an interesting scientific topic. Explanations dedicated to this device generally resort to mathematic or circuit models, thus offering less substantial information based on physical phenomena. Aiming to correct what we consider to be a deficiency, this paper wishes *both* to present some intuitive explanations on the functional features of the Bipolar Junction Transistor *and* to emphasize some concepts and denominations that the author considers rather misleading or imprecise.

Keywords: Bipolar Junction Transistor, "amplification", intuitive explanation, physical explanation.

1. Introduction

Concerning the working principle of a Bipolar Junction Transistor, we think that clearer explanations and explanatory clarifications on how a BJT operates might be welcome, at least for any studious individual tormented by the questions "how ?" and "why ?". In this spirit, the present material is striving *not only* to cast light to what happens basically inside a BJT *but also* to correct what appears to be famous misconceptions, leaving aside the scholastic approach of usual abundance of mathematical formula and instead focusing on an intuitive explanation relying on a cause and effect chain based on common sense physics.

First and foremost, we shall now shift from the comfort of mythology to the severity of reality, presenting a few myths, traditionally governing common knowledge on the BJT.

Myth: A BJT is an amplifier.
Reality: No, never. It is definitely not an amplifier, but a voltage-current transducer. It is simply a semiconductor-based, non-linear circuit element whose main current is controlled by a small voltage.

Myth: A BJT is a current source.
Reality: No, *per se* - but yes, in a specific circuit configuration. When powered and biased correctly, the BJT can be seen as a (of course, non-ideal) current source.

Myth: A BJT is a current controlled device.
Reality: No, it is a device (consisting of semiconducting materials) that can accept voltage control, not current control.

Generally, various kinds of humans will answer the question "what a BJT is?" saying that a BJT is "an amplifier", or "a current source", of "a something controlled something" practically never saying that a BJT is nothing but a three-port electronic device consisting of semiconducting materials in a specific arrangement. Why people answer what something does when asked what something is?

To summarize, a BJT is a three-port electronic device consisting of semiconducting materials in a specific arrangement. Together with some passive components (resistors, capacitors) and a DC power source used for polarization, it can be used to produce large voltage variations at the output – caused by smaller voltage variations at the input. This operation is nicknamed "amplification".

2. Odd try on triodes

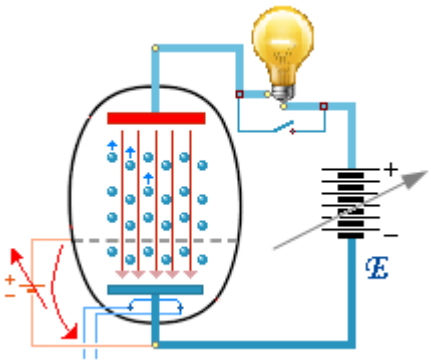


Figure 1. Vacuum Triode with control circuit and Power Circuit

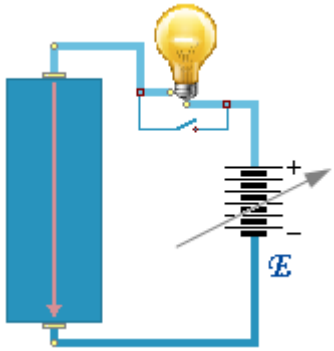


Figure 2. Instead of a Triode we place an n± doped semiconductor

Let us have a look at the most elementary circuit with a vacuum triode, represented in Figure 1.

We have a vacuum triode, whose main current (blue bubbles, i.e. electrons in Figure 1) is controlled by adjusting the small voltage intercalated right in the way of the electrons flow.

Why would we not try to put a semiconductor device instead of a Triode? This is exactly what is represented in Figure 2, in which the magenta vertical arrow represents the electric field established by the adjustable voltage source "E".

But how could we emulate the situation presented in Figure 1 using the elements of Figure 2?

Is there any way in which we could set the "power" current such that it could be adjusted by a tiny variation of a command voltage? What would be the best idea with this respect?

Well, how about setting a PN junction inside the n± doped semiconductor?

Let us have a look at Figure 3 and Figure 4.

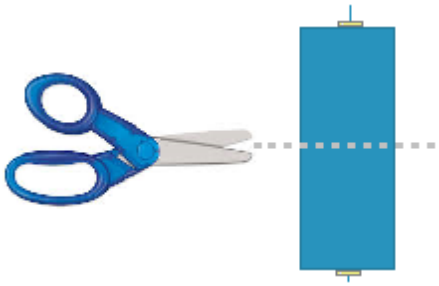


Figure 3. Imaginary split in the n± doped semiconductor

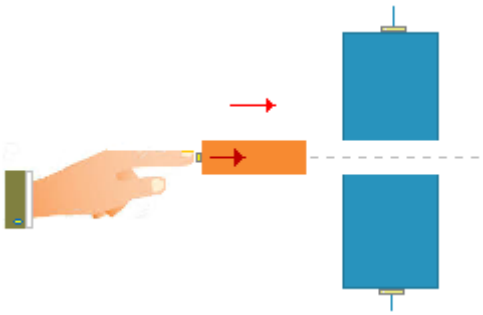


Figure 4. Imaginary insertion of an p-doped layer

They all suggest the insertion of a p-doped layer in order to create a PN junction inside the n± doped semiconductor (In fact, we have to acknowledge the inevitable birth of two pn junctions: between the p-doped layer and the lower slice as well as between the same p-doped layer the upper n-region). This sequence (Figure 3 and Figure 4) does refer *neither* to how the transistor was invented *nor* to how a BJT is manufactured. Instead, this sequence tries to suggest a "transition" from the vacuum triode to the BJT, keeping in place the functional similarities as much as possible.

After the insertion we have obtained 2 PN Junctions that are created spontaneously, as represented in Figure 5. Now, let as reconstitute the Triode circuits, just like in Figure 6.

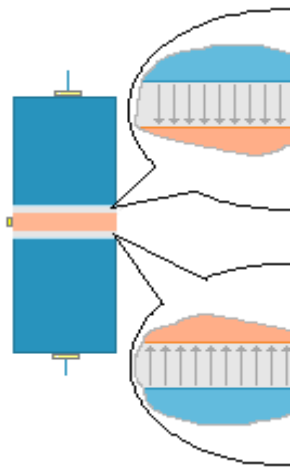


Figure 5. Two PN Junctions are created spontaneously

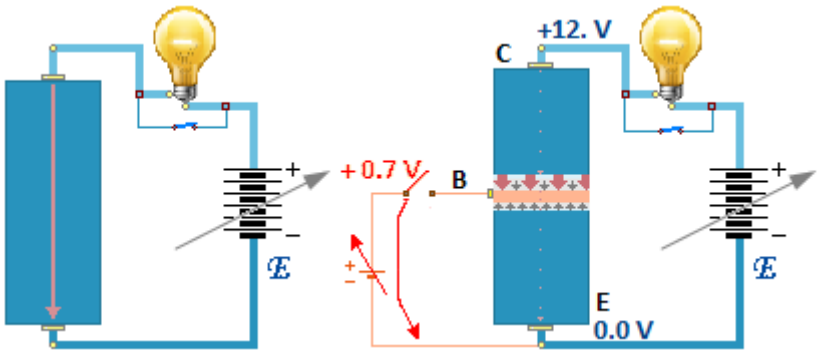
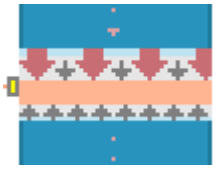


Figure 6. Emulating the Triode circuits: without (left) and with (right) the p-doped slice

In the main circuit, we have the "characters" listed in the next Table:

	• the two junctions, each with the respective "adversative" inner electric field (gray short arrows); each one is an impeder (obstructer) to the intended current. Here it is important to mention that automatically a "voltage barrier" is created across the pn-junctions (*diffusion** voltage , ~ 0.6 to 0.7V). That is the reason we need an external voltage to work against this barrier. • the electric field; it is the motive (driving) force (magenta thick arrows);
	the pool of electrons; they are ready to pour (to spill out) inside the **B**ase (and maybe further), but at the same time they are hampered by the inner field of the B-E junction.

***** ***Diffusion*** is spontaneous movement of substance from a high concentration zone to a low concentration zone.

Let us take a closer look at Figure 6, considering two situations.

(1) The red intermediary small voltage source in the B-E circuit in not yet connected.

Do we expect a current (flow) from **E** to **C**, in these condition? (see Figure 6, focusing on the right image).

So let us see: we might be tempted to think that the electric field "magenta", the motive (driving) force (quite strong), would be supposed to bias forward the B-E junction, whose opposing voltage is about ~ 0.7 V. But it will not, as its whole strength is being kept busy with over-widening the C-B junction that it biases in reverse (Figure 6, image on the right).

After the insertion of the p-doped slice, the distribution of the electrical field is inflicted a dramatic change, visually suggested as follows: the long, thin magenta arrow in the left image becomes the group of short, thick arrows interleaved between the two timid dotted arrows (Fig. 6, image on the right). This suggests that the whole u_{CE} drops practically only on the CB junction, its value outside the CB junction being extremely low (expectable, since the 3 doped regions are all pretty good electrical conductors). Both pn junctions can be seen as a series combination of two back-to-back diodes – however only one (B-E) readily *would* conduct, while the other one (C-B) firmly interdicts any current through the whole C-E path. Consequently, the circuit will yield NO current.

Still, we have an opportunity:

(2) To connect the intermediary small external voltage, i.e. the red circuit of the **B**ase (just "close" the red mini circuit-breaker in Figure 6). A FORWARD voltage will drop DIRECTLY on the B-E junction, narrowing it, thus reducing its resistance, hence offering to the "nascent" current a chance to exist (offering the standing-by electrons a chance to flow) - see Fig. 7 and 8.

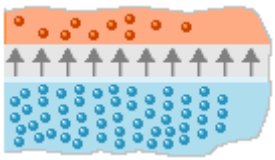


Figure 7. Carriers from the Base and from the Emitter are ready to diffuse into each other but the inner field of the BE junction firmly prevents them to do so

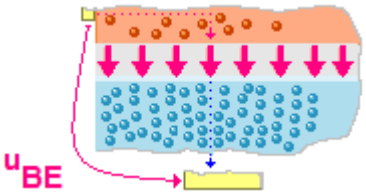


Figure 8. An external voltage is the chance to counteract the barrier

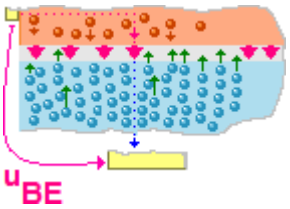


Figure 9. An increasing E-field allows more and more electrons to move toward the Base

In other words, at this point the internal diffusion voltage comes into the game: the external voltage counteracts the barrier and the increasing E-field allows more and more electrons to move (spill out) into the junction area (see Figure 9).

A flow indeed will occur, in the form of a diffusion current, as charge carriers in both regions (the blue **E**mitter and the orange **B**ase - as shown in Fig.6, the image at the right) can hardly wait to rush (pour, spill-out) into each other due to the charge gradient existing between these 2 territories.

We remind that our wish is that as many as possible electrons *can* reach the C-B junction where the rather strong electrical field awaits to grab them and drag them toward the Collector terminal, thus generating the desired "power" current. Nevertheless, 2 problems *emerge*: *(a)* how can we be sure that among the moving electrons *really exists* an important fraction able to overcome the risk of recombining inside the **B**ase (this risk can deteriorate the success of a "go-by" flow toward the Collector); *(b)* how can we be sure that the current will not *rather* deviate to the control circuit (the close one, on the left side, depicted in red) *instead of* choosing the path of the **C**ollector?

The only way to persuade the "power" current to head toward the **C**ollector at the greatest extent possible is to take the following 2 counter-measures:

- the **B**ase must be extremely thin (fractions of a micron to some dozens of microns), so that the junction (as wide as ~ 15÷20% of the base) be also extremely thin, so that the blizzard of electrons spilling out by diffusion from Emitter to Base have good odds to traverse the Base "in a big hurry" and reach the CB junction where the strong magenta electric field is awaiting to drag them further to the Collector (Figure 10 bis). The magenta field, though objectively widening the gap of the C-B junction, is nonetheless the sole entity that drags the electrons inside the "desert" of the CB junction. The majority of electrons in the junction area "feels" more and more that there is a larger positive voltage at the C node, thus being driven toward the Collector;
- the **B**ase must be scarcely doped, so that the electrons traveling from E to C have a very low chance of recombining.

And that is all. Basically, by imposing these manufacturing peculiar solutions, we ensure that about 98 ÷ 99 ÷ 99.8 % of the current departing from the E is retrieved at the **C** terminal. It is just a small (nearly fixed) fraction of the emitter current that does not reach the **C** node. This residue forms the base current, which should be as small as possible (it cannot totally be made zero, in reality having the value of 0.2% ÷ 1% of i_C). How has the Triode functionality been reconstructed with semiconductors? We shall see in the paragraphs bellow.

3. Out of respect for the Triode. Not so late for the Early effect

All right, now, about 98÷99÷99.8 % of the current departing from the **E**mitter is retrieved at the **C** terminal. Everybody is doing its job: *(a)* the **E**mitter offers a pool of standing-by electrons; *(b)* the BE external voltage biases forward the B-E junction thus rendering possible the flow, i.e. permits the electrons to spill out (pour, diffuse) into the very thin and scarcely doped **B**ase and reach in a "big hurry" the strongly reverse biased C-B junction were the magenta field awaits to drag them - this is a diffusion current; *(c)* the magenta field drags them - this is a drift current - through the "desert" toward the **C** terminal (while being transited, the **B**ase has just successfully lured some electrons with the chance of a one micro/nano-second stand recombination).

The C-B junction is deprived of charge carriers, but contrary to intuition has an inner field that helps - not hampers - the flow. The current traversing the C-B is a sui-generis, a "foreign" current.

But how can we control this "power" current? How can we emulate the old good Triode?

In order to answer this question, it is necessary to know what a control voltage does to a junction; *(a)* if reverse biased by the control voltage, a junction gets widened; *(b)* if forward biased by the control voltage, a junction gets narrowed, just like in Figure 10. Again we emphasize that it is the influence of the diffusion barrier (gray arrows) that u_{BE} must overcome, thereby allowing an exponential current increase: $i_C \sim exp\ (\ u_{BE}\ /\ V_T\ )$.

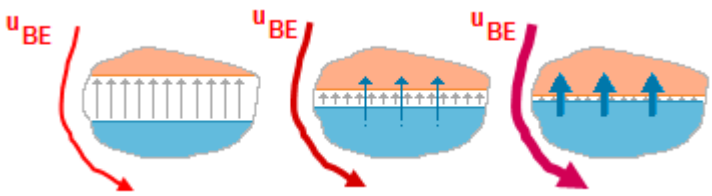


Figure 10. The higher the uBE voltage, the narrower the B-E junction, consequently the higher the diffusion current poured

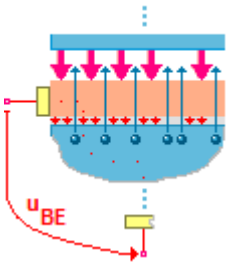


Figure 10 bis. ~ 99% of the electrons traverse the Base "in a big hurry" and reach the CB junction

So, in the end, how does a BJT amplify? We repeat, it does not! It uses a low power voltage input to control a "power" current.

- When $\mathbf{u_{CE}}$ **is being kept constant**, any small variation of the $\mathbf{u_{BE}}$ voltage widens / squeezes (narrows) the BE junction, thus inflicting a dramatic increase / decrease to the resistance of the junction, thus dramatically diminishing / augmenting the "power" current. So, basically, everything is at the good will of the B-E junction (see Figure 7, 8, 9, 10);
- When $\mathbf{u_{BE}}$ **is being kept constant**, any variation of the $\mathbf{u_{CE}}$ voltage, no matter how important, does nothing with respect to changing the value of the power current. Why? Well, who sets the value of the i_C current? It is exactly the overspill coming from the Emitter, whose mass is established by the doping level of the **E**mitter and whose speed is established by (a) the charge gradient between **E**mitter and **B**ase and (b) by the aperture (opening) of the E-B junction. And that is all. It does not matter whether this yielded current is driven totally toward the **B**ase terminal (hypothetically missing **C**ollector) or it is directed to the maximum possible extent to be absorbed by the a reverse electric field and further captured into the **C**ollector terminal: this current (this electro-kinetic "harvest") remains the very same, regardless of the u_{CB} voltage. This voltage does nothing but simply taking over the output that is being offered by the Emitter and drifting it away, on a tiny distance (CB depletion zone), further to the **C**ollector. In addition, the CB junction is a reversed biased one, about which we know that no matter how wide be the range in which the reverse voltage fluctuates, the reverse current remains the same. Usually, the reverse current is tiny because minority carriers are extremely few. So, no wonder that this *unusual* reverse current is way bigger than *usual* reverse currents, as the

Emitter happens to supply a ready copious load of electrons. However, rigorously speaking, the voltage $\mathbf{u_{CE}}$ has indeed a small influence on the useful current (this is the Early effect, discovered in the 1950's by James Early). When $\mathbf{u_{CE}}$ increases, the depletion area between C and B increases slightly – and the remaining junction area between **B** and **E** decreases correspondingly, hence augmenting the forward biasing electric field, hence yielding a undesired small increase in $\mathbf{i_C}$ with rising $\mathbf{u_{CE}}$.

That is the BJT: a voltage - controlled circuit element, whose point is to adjust a "strong" current by a small voltage.

4. I_B, who?

What about I_B? Where is it? Where does it appear in this happenstance? Nowhere, obviously. All we have heard about are a pool of electrons ready-to-move, an electric field ready to drag them and 2 impeders (a superable one, i.e. the C-B junction and a decisive one, i.e. the B-E junction, ***whose*** adjustment performed by $\mathbf{u_{BE}}$ will establish at what value will the current be permitted to settle).

So, first and foremost, the idea that $i_B = f(\boldsymbol{u_{BE}})$ is *per se* important must be abandoned and the reality that $i_C = f(\boldsymbol{u_{BE}})$ is *per se* important must prevail.

Still, there is a base current. But how does it appear in the picture? What generates it? Does it play any role in any action? Does it bear any importance at all? Is it noxious? Is it innocuous? Is it spontaneous? Is it set on purpose? We shall see in the paragraphs bellow. Let us look at Figure 11. It represents a BJT (the one in Figure 6, image on the right side, but rotated 90° clockwise) working in the Active Mode [1], the main "characters" being explicited in the Table annexed alongside the right of Figure 11.

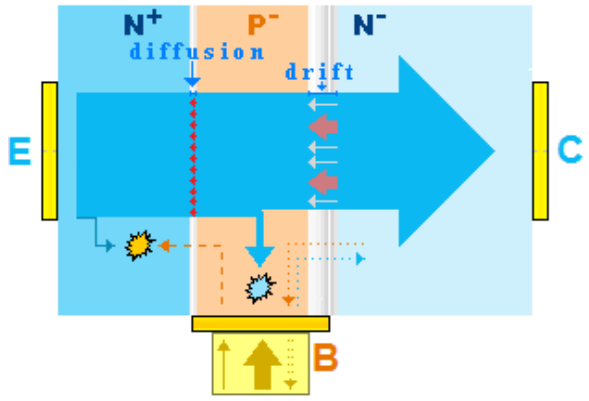


Figure 11. Currents in the Active Mode

	It starts as a *Diffusion* current; it is the movement of electrons being poured by the charge gradient from Emitter to the Base, under the permission of u_{BE}. This current enters the CB junction where is being grabbed by the strong magenta field and retransmitted *by drift* to the territory of the Collector.
	It is the fraction (of the useful current) *inevitably lost* by recombination with the holes in the base. It generates the spontaneous absorption of the *diffusion current* (as nature always hates to be electrified and always loves to be electrically neutral, it makes charge carriers roll down from higher concentration to lower concentration), under the formula: $\boldsymbol{j} \cong grad\,(Q)$. This accidental fraction is the main wrongful argument of the misleading and erroneous opinion that a BJT is a current controlled device.

	Diffusion current; it is the movement of holes being poured by the charge gradient. This current progressively "dissolves" itself (fades away) inside the electrons mass, inside the Emitter. It is a very small current.
	Extra current absorbed spontaneously by the Emitter to compensate the small current from the row above.
	Drift current; it is the movement of minoritary holes being dragged by the motive electric field (magenta horizontal arrow heading to the left) from the Collector to the Base. It has an extremely low value.
	Drift current; it is the movement of minoritary electrons dragged by the motive electric field (magenta horizontal arrow heading to the left) from the Base to the Collector. It has an extremely low value.

Does anybody perceive any connection of any sort between these components (i.e. the **B**ase current) and the "power" (main, strong) current that is adjusted by the u_{BE} voltage? As obviously, we do not. If yes, could you name just one? We could not. Because there are none. All these are simply accidental parasitic currents, being neglected in many calculations. They are merely an unavoidable undesired *side* effect. This confirms that a BJT is a voltage controlled quadripole, not an amplifier (if it were an amplifier, the strong entity would have been controlled by a co-generic entity, i.e. current-to-current, voltage-to-voltage etc.). As well, at this point, it should be mentioned that the base current is responsible for the finite input resistance at the base node (disadvantage if compared with the FET).

5. The controversial impostor

It is Beta. The illustrious Beta = I_C/I_B, or di_C/di_B, or i_c/i_b is, phenomenologically, devoid of any significance. It rather expresses the ability of the BJT to minimize a spontaneous inevitable loss, but unfortunately and misleading, it is common practice to say that the current ratio beta gives the "current amplification".

Why would it be empowered with such great importance?

Perhaps because humans prefer a-dimensional measures when quantitatively assessing a conversion.

- efficiency = (useful power) / (useful power + looses) × 100
- interest = (money/money) × 100
- profit rate = (money - money) / money × 100
- voltage ratio for electric. transformers = (voltage / voltage) × 100
- speed ratio for toothed gears = (rpm / rpm) × 100
- etc.

Or perhaps it is convenient in small signal circuit calculations. Who knows?

What would be the correct assessment? An average slope, the differential ratio (slope of the $i_C = f(u_{BE})$ curve) that is called transconductance: $g_m = di_C / du_{BE}$.

6. However, a Darlington pair obviously means current control, right?

Wrong. The forward bias voltage between the **B**ase of Small and the **E**mitter of LARGE must be - as it is known - approximately **1.4** volts. As it is obvious, a Darlington pair is a voltage divider consisting of the two B-E junctions. This means: it is not the **B**ase current of LARGE that "produces" the corresponding ~ **0.7** V drop on the BE junction of LARGE, but, on the contrary, the ~ **0.7 V** drop on the BE junction of Small is the necessary precondition *for the existence* of the **B**ase current of LARGE (exactly the "power" current of Small).

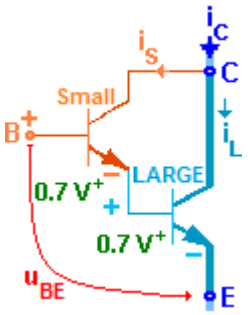


Figure 12. A Darlington configuration still confirms the voltage control

7. I am saturated: I want to eat another big steak

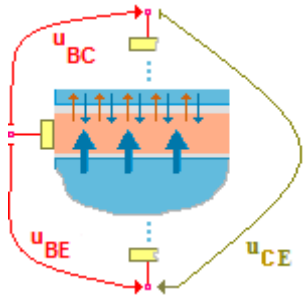


Figure 13. Both junctions are forward biased

Let us focus on Figure 13.

$u_{BC} = u_{BE} - u_{CE} \cong$ **0.7** V - u_{CE} . Quite elementary.

The reverse field (the motive force) that drives electrons away to the Collector resides in the CB junction (not presented in Figure 13, but presented in Figure 6, image on the right side, as the thick magenta arrows).

In the "saturation" region (where u_{CE} ranges from 0 to 0.7 V ÷ ~1.+V) u_{BC} will range from ~0.7 V to ~0. V (see the formula above), case in witch the field generated by this voltage biases forward the CB junction, having as consequence the diffusion current (*the electrons* coming from the Collector to spill out into the **B**ase - thin blue downward arrows *and the holes* coming from the **B**ase to spill out into the **C**ollector - thin orange upward arrows) that counters (opposes) *the "power" desired* i_C *current* - thick blue upward arrows, as suggested in Figure 13. Within this functional zone, the augmentation of the main current strongly depends on u_{CE}, but as much as u_{CE} augments u_{BC} diminishes ("saturation" zone is gradually abandoned) and consequently i_C re-becomes independent of u_C, as suggested in Figure 14.

The result is that if we open widely and widely the BE barrier (i.e. if we keep augmenting u_{BE}) at a given u_{CE} (in the range 0 ÷ 0.7 V ÷ ~1.+ V) the power current (i_C) will not increase at all as expected (it will be as much as ~30 % ÷ 40 % of the value reached in case of a "decent" u_{CE}). This is being illustrated by the well known commencement zone of the output curves, as suggested in Figure 14.

Physically, there cannot be any saturation, because at low values of u_{CE} a sufficient increase in the input voltage (u_{BE}) can generate correspondingly a main current of high value (that can transform easily the BJT into a "barbecue") even if not reaching the expected level. Saturation would have meant an extremely severe (if not total) limitation of the main current. Unlike "*saturation*", denominations such as: "*subnutrition*", "*under-supply*", "*low-biasing*" etc. would offer a much more appropriate description.

* * *

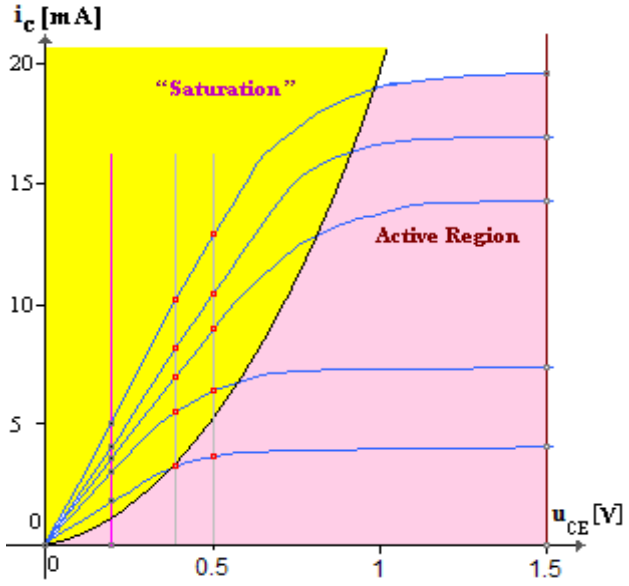


Figure 14. Saturation? Where is it?

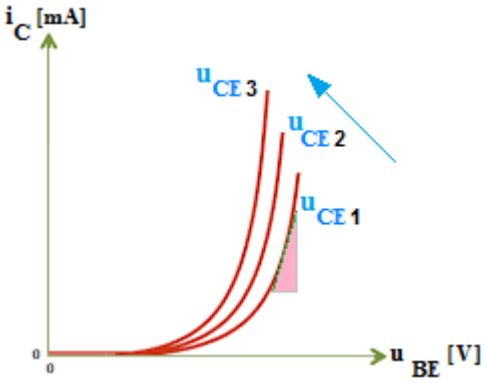


Figure 15. The input voltage controls the main current

If it looks like a duck, swims like a duck and quacks like a duck, then it surely is a pig

If you think that the phrase above is usually pronounced in lecture theater at the Agronomic or Veterinary Science Academy, then you are wrong. This phrase describes exactly how people teach (or are being taught) the Bipolar Junction Transistor.

When discussing "saturation", we notice that in addition to the imprecision pertaining to denominations, there is an another one, pertaining to graphical representation: the notorious blue curves in Fig. 14 have a sheer problem of consistency between two concepts and their respective visual incarnation.

Saturation is the state of fact in which satisfaction ceases to grow, after having achieved a "filling up" value, regardless of the on-going growth of the supply. In other words, the effect becomes indifferent to further augment of the cause. With respect to the *visual representation*, saturation as *effect = f (cause)* is rendered in the form of an initially "growing" curve that flattens after reaching the turning point (corresponding to the "filling up" value) .

Figure 15 represents the input characteristics. Now, just look carefully at Figure 14 and 15, in which the vertical lines of Figure 14 correspond respectively to the burgundy exponential curves of Figure 15. Do you perceive anything that can suggest visually the idea of saturation? As obviously, we do not. Ironically, ridiculously and troubling, yes, there is however something looking like saturation, namely exactly the flat zones of the blue curves.

Let us look at the grow zone of the blue curves in Figure 14. Shockingly, this portion is precisely the legendary "saturation" zone. Do you think it looks like saturation? As obviously, we do not. Moreover so, just look at the flattening zone of the same blue curves. Do you guess how is this zone popularized as? No matter how grotesque it be, it is renowned as the linear zone. Obviously, some individuals will argue that "linear" in Figure 14 pertains to the rapid-grow zones of the curves (quasi - slopes) in Figure 15. We shall strongly disagree, otherwise we must conclude that *scrutinizing* the curves of a BJT *is based* on hocus-pocus tricks that induce optical illusions. So, the 2 zones of the blue curves both bear *not only* utterly outragingly anti-intuitive, *but also* exactly inverted names.

8. Still, why a BJT cannot possibly be controlled by current?

Let us suppose that somehow we manage to inject a load of holes in the **B**ase (with a current source, or with a syringe, or by pouring a bucket of holes or does not matter how). What is going to happen?

All that is going to occur will be nothing but a re-absorption/redistribution of the charge, more or less close to the BE junction. The junction will still remain in place, its inner electrical field implacably playing further its role of obstruct to the desire of the charge carriers from both sides to diffuse (spill out) into each other. This inner electrical field is a blocking force that can be mitigated (counteracted) only by another force, hence by an external voltage.

9. Conclusions

Generally using mathematical mechanisms, *teaching* the inner phenomena that govern the functioning of the BJT still *retains* unanswered questions on concrete physical behavior.

The material herein presented is striving to fill the above mentioned lacunas, proposing a model based on intuitive explanations that cast light to what happens basically inside a BJT.

Our paperwork commences this "journey" by underlining that a BJT is a semiconductor-based version of an extremely important, classical valve (the vacuum triode), then proceeds by providing intuitive constructions based on physical phenomena and by proposing appropriate corrections to what appears to be famous misconceptions, and ends by drawing attention on what we consider to be misleading or imprecise concepts / denominations.

This material does not resort to mathematical formula, nor to circuit models, instead focusing on intuitive explanations relying *on a cause and effect chain* based on common sense physics.

Acknowledgement

This article is the result of the fruitful conversations that the author has had with Prof. Lutz von Wengenheim (Bremen, Germany), in Apr-May 2019. We wish to express our gratefulness for Prof. von Wangenheim's generosity and promptitude in offering his body of knowledge for the writing of this material.

REFERENCES

1. ZEKRY, Abdelhalim - "Electronic Devices", Cairo 1996; Fig. 4.4, pg. 141.

Dragoș NICOLAU is a graduate of the Faculty of Electrical Engineering at the Politehnica University of Bucharest, since 1991. Currently he holds the position of scientific researcher III at ICI, Bucharest. He has skills in developing web, desktop and network applications. He is strongly passionate about the less-explored areas of object-oriented programming. As a field of interest, Dragoș Nicolau also mentions: securing networks, securing JavaScript codes and semiconductor physics. Mr. Nicolau loves to study and deploy software based on execution threads, encryption / compression algorithms, image analysis, and network communications. He has published more than 30 articles in the country and abroad. Between 1997-2002 he performed Academic teaching activity at the Faculty of Electrical Engineering of UPB. Dragoș Nicolau is speaking Italian, French and English.