Chapter 1

Bipolar Transistors

John D. Cressler and Katsuyoshi Washio

COPYRIGHTED MATERIAL

1.1 Motivation

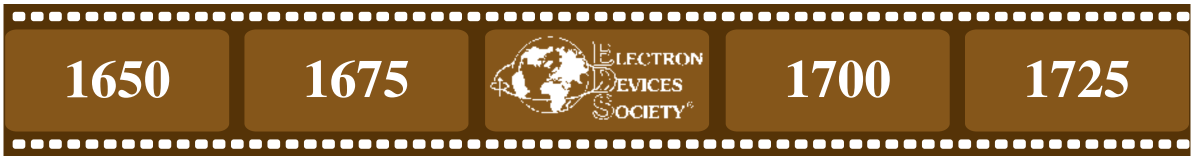
In terms of its influence on the development of modern technology and hence, global civilization, the invention of the point contact transistor on December 23, 1947 at Bell Labs in New Jersey by Bardeen and Brattain was by any reckoning a watershed moment in human history [1]. The device we know today as a bipolar junction transistor was demonstrated four years later in 1951 by Shockley and co-workers [2] setting the stage for the transistor revolution. Our world has changed profoundly as a result [3].

Interestingly, there are actually seven major families of semiconductor devices (only one of which includes transistors!), 74 basic classes of devices within those seven families, and another 130 derivative types of devices from those 74 basic classes (Figure 1.1) [4]. Here we focus only on three basic devices: (1) the *pn* homojunction junction diode (or *pn* junction or diode), (2) the homojunction bipolar junction transistor (or BJT), and (3) the special variant of the BJT called the silicon-germanium heterojunction bipolar transistor (or SiGe HBT). As we will see, diodes are useful in their own right, but also are the functional building block of all transistors.

Surprisingly, all semiconductor devices can be built from a remarkably small set of materials building blocks (Figure 1.2), including [4]:

- the metal–semiconductor interface (e.g., Pt/Si; a "Schottky barrier")
- the doping transition (e.g., a Si *p*-type to *n*-type doping transition; the *pn* junction)
- the heterojunction (e.g., *n*-AlGaAs/*p*-GaAs)
- the semiconductor/insulator interface (e.g., Si/SiO_2)
- the insulator/metal interface (e.g., SiO_2/Al).

Guide to State-of-the-Art Electron Devices, First Edition. Edited by Joachim N. Burghartz.
© 2013 John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.



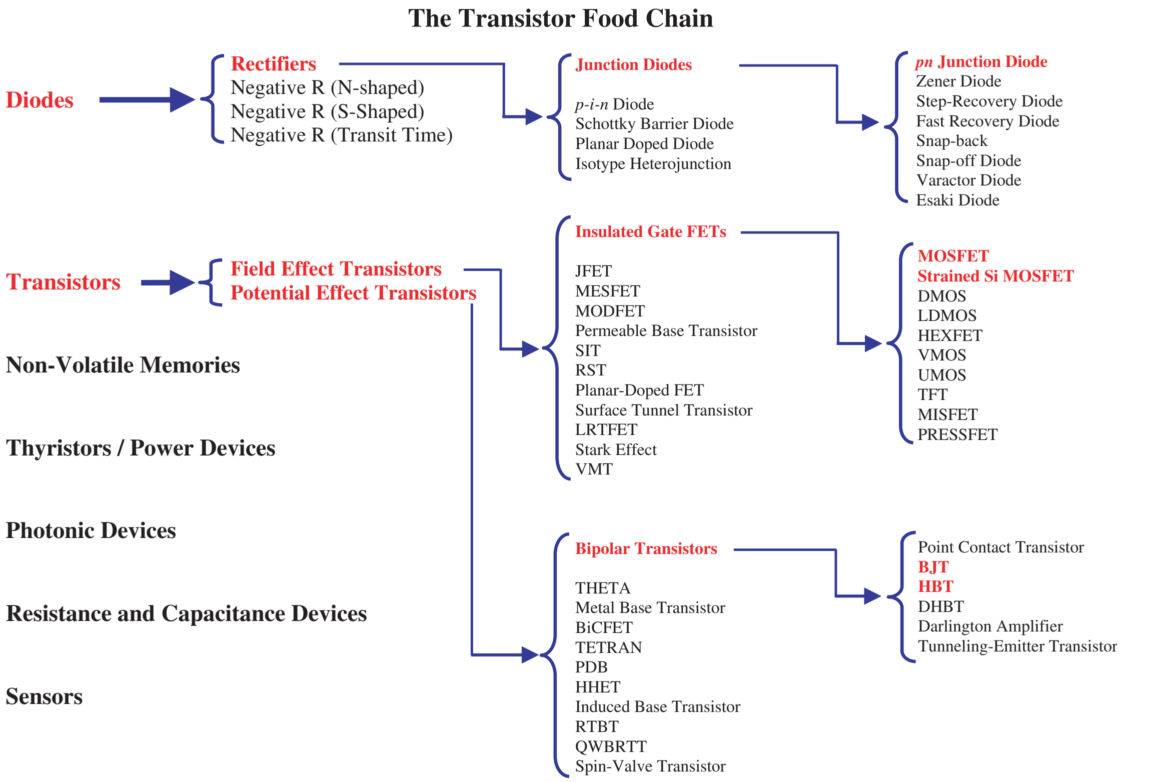
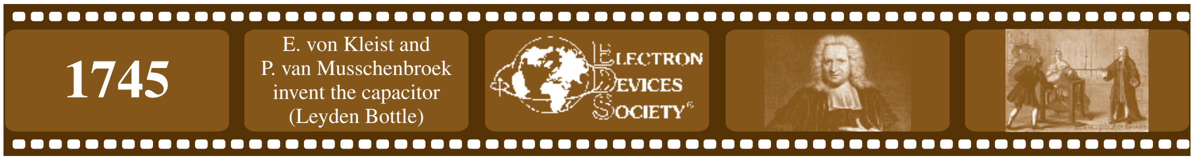


Figure 1.1 The transistor ''food chain'' showing all major families of semiconductor devices. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press

Why do we actually need transistors in the first place? Basically, because nature attenuates all electrical signals. By this we mean that the magnitude of all electrical signals (think "1s" and "0s" inside a computer, or an EM radio signal from a cell phone) necessarily decreases as it moves from point A to point B, something we call "loss". When we present an (attenuated) input signal to the transistor, the transistor is capable of creating an output signal of larger magnitude (i.e., "gain"), and hence the transistor serves as a "gain block" to "regenerate" (recover) the attenuated signal in question, an essential concept for electronics. In the electronics world, when the transistor is used as a source of signal gain, we refer to it as an "amplifier." Amplifiers are ubiquitous to all electronic systems.

Naming of the Transistor

The name "transistor" was actually coined by J.R. Pierce of Bell Labs, following an office betting pool which he won. He started with a literal description of what the device actually does electronically, a "transresistance amplifier," which he first shortened to "trans-resistor," and then finally "transistor" [3].



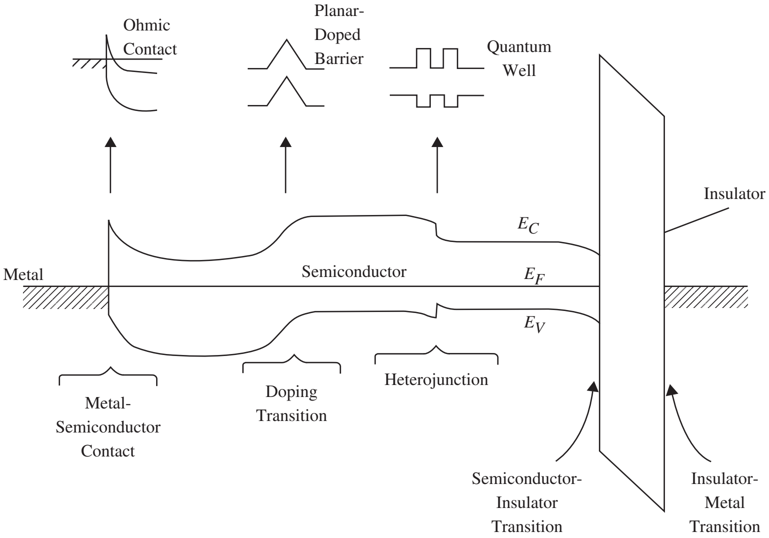


Figure 1.2 The essential building blocks of all semiconductor devices. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press

Not only can the transistor serve as a wonderful nanoscale sized amplifier, but importantly it can also be used as a tiny "regenerative switch"; meaning, an on/off switch that does NOT have loss associated with it. Why is this so important? Well, imagine that the computational path through a microprocessor requires 1 000 000 binary switches (think light switch on the wall – on/off, on/off) to implement the complex digital binary logic of a given computation. If each of those switches even contributes a tiny amount of loss (which it inevitably will), multiplying that tiny loss by 1 000 000 adds up to unacceptably large system loss. That is, if we push a logical "1" or "0" in, it rapidly will get so small during the computation that it gets lost in the background noise. If, however, we implement our binary switches with gain-enabled transistors, then each switch is effectively regenerative, and we can now propagate the signals through the millions of requisite logic gates without excessive loss, maintaining their magnitude above the background noise level.

In short, the transistor can serve in one of two fundamental capacities: (1) an amplifier or (2) a regenerative switch. Amplifiers and regenerative switches work well only because the transistor has the ability to produce gain. So a logical question becomes, where does transistor gain come from? To answer this, first we need to understand *pn* junctions.

1.2 The *pn* Junction and its Electronic Applications

Virtually all semiconductor devices (both electronic and photonic) rely on *pn* junctions (a.k.a., "diodes", a name which harkens back to a vacuum tube legacy) for their functionality. The simplest embodiment of a *pn* junction is the *pn* "homojunction", meaning that within a single piece of semiconductor (e.g., silicon – Si) we have a transition between p-type doping and n-type doping (e.g., p-Si/n-Si). The opposite would be

1750 1775 Electron Devices Society 1782 Alesandro Volta
develops the condenser

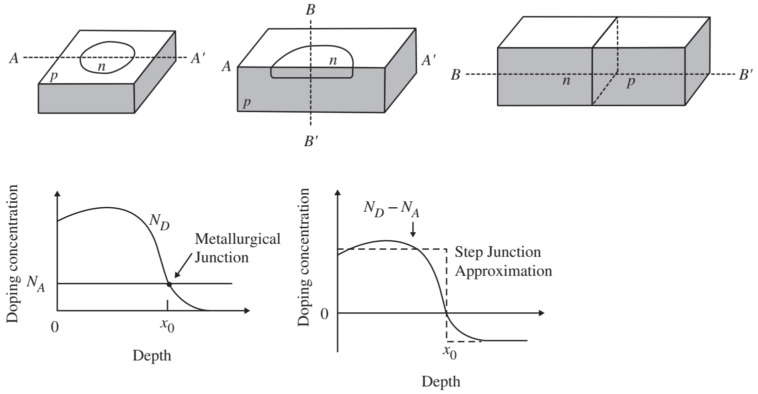


Figure 1.3 Cartoons of a pn junction, showing doping transition from n-type to p-type. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press

a *pn* heterojunction, in which the p-type doping is within one type of semiconductor (e.g., p-GaAs), and the n-type doping is within another type of semiconductor (e.g., n-AlGaAs).

As shown in Figure 1.3, to build a *pn* junction we might, for instance, ion implant and then diffuse *n*-type doping into a *p*-type wafer. The important thing is the resultant "doping profile" as one moves through the junction ($N_D(x) - N_A(x)$, which is just the net doping concentration). At some point in the doping transition, $N_D = N_A$, and we thus have a transition between net n-type and net p-type doping. This point is called the "metallurgical junction" (x_0 in Figure 1.3) and all of the important electrical action of the junction is centered here. To make the physics easier, two simplifications are typically made: (1) Let us assume a "step junction" approximation to the real *pn* junction doping profile, which is just what it says, an abrupt change (a step) in doping occurring at the metallurgical junction (Figure 1.3). (2) Let us assume that all of the dopant impurities are ionized (one donor atom equals one electron, etc., an excellent approximation for common dopants in silicon at 300 K).

So, how does a *pn* junction actually work? The operation of ALL semiconductor devices is best understood at an intuitive level by considering the energy band diagram, which plots electron and hole energy as a function of position as we move physically through a device. An n-type semiconductor is electron rich (i.e., majority carriers), and hole poor (i.e., minority carriers). Conversely, a p-type semiconductor is hole-rich and electron-poor. If we imagine bringing an n-type and p-type semiconductor into "intimate electrical contact" where they can freely exchange electrons and/or holes from *n* to *p* and *p* to *n*, the final equilibrium band diagram shown in Figure 1.4 will result. Note, that under equilibrium conditions, there is no NET current flow across the junction.

We might logically wonder what actually happened inside the junction to establish this equilibrium condition. When brought into contact, the *n*-type side of the junction is electron rich, while the *p*-type side is electron poor. That is, there is a large driving force for electrons to diffuse from the *n* region to the *p* region. Recall, that there are in fact two ways to move charge in a semiconductor: (1) drift, whose driving force is the electric field (voltage/length), and (2) diffusion, whose driving force is the carrier density gradient (change in carrier density per unit distance). The latter process is what is operative here.

1799

The first battery was
demonstrated by
Alessandro Volta

ELECTRON
DEVICES
SOCIETY

1800

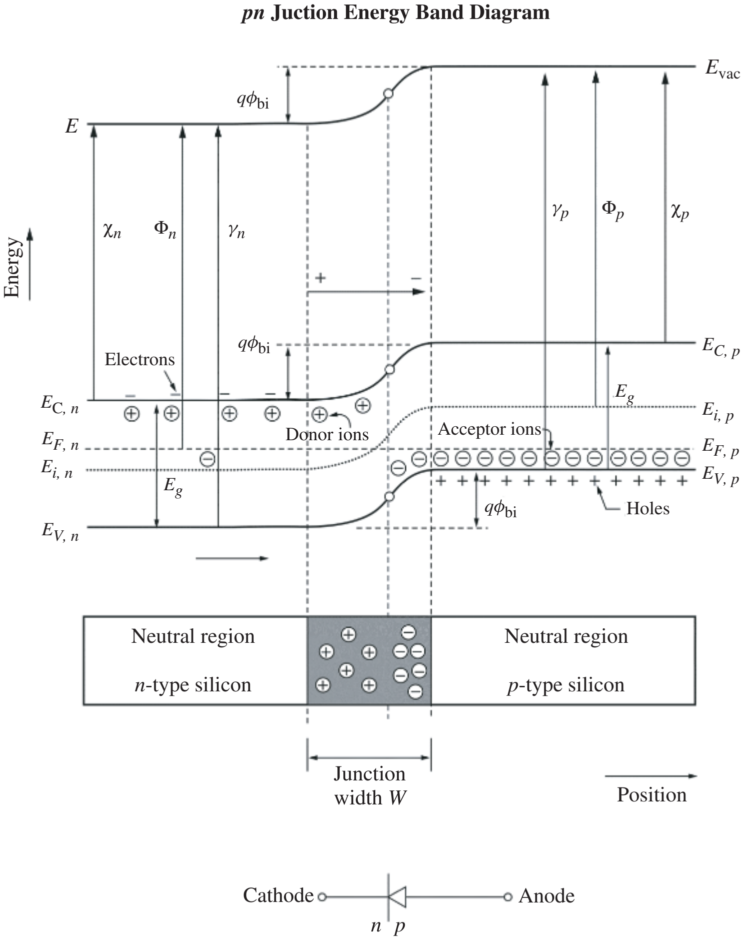


Figure 1.4 Energy band diagram of a pn junction at equilibrium. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution;* 2009, Cambridge University Press

Once in electrical contact an electron moves from the *n*-side to the *p*-side, leaving behind a positively charged donor impurity (N_D^+). Note, that far away from the junction, for each charged donor impurity there is a matching donated electron, hence the semiconductor is charge neutral. Once the electron leaves the *n*-side, however, there is no balancing charge, and a region of "space charge" results. The same thing happens on the *p*-side. Hole moves from *p* to *n*, leaving behind an uncompensated acceptor impurity (N_A^-) behind. This resultant charge "dipole" produces an electric field, pointing from + to − (to the right in this case). How does that induced field affect the diffusion-initiated side-to-side transfer of charge just described? It opposes the diffusive motion of both electron and holes via Coulomb's law. Therefore, in a *pn* junction the diffusion gradient moves electrons from *n* to *p* and holes from *p* to *n*, but as this happens a dipole of space charge is created between the uncompensated ionized dopants, and an induced electric field opposes the further diffusion of charge. When does equilibrium in the *pn* junction result? When the diffusion and the drift processes are perfectly balanced and the net current density is zero.

The *pn* junction in equilibrium consists of a neutral *n* region and a neutral *p* region, separated by a space charge region of width *W*. This structure forms a capacitor (conductor/insulator/conductor), and *pn* junctions have built-in capacitance which will partially dictate their switching speed. The electric field in the space charge region (for a step junction) is characteristically triangular shaped, with some peak value of electric field present. There is a built-in voltage drop across the junction, and, thus, from the energy band diagram we see that there is a potential barrier for any further movement of electrons and holes from side-to-side. This barrier to carrier transport maintains a net current density of zero, and the junction is by definition in equilibrium.

If one wanted to get current flowing again across the junction, how would this be done? Well, we must unbalance the drift and diffusion mechanisms by lowering the potential barrier to the electron and hole transport, and we can do this trivially by applying an external voltage to the *n* and *p* regions such that the *p* region (anode) is more positively biased than the *n* region (cathode). As shown in Figure 1.5, this effectively lowers the side-to-side barrier, drift no longer balances diffusion, and the carriers will once again start diffusing from side-to-side, generating useful current flow. This is called “forward bias”. What happens if we apply a voltage to the junction of opposite sign? (i.e., *p* region more negatively biased than the *n* region). Well, the barrier the carriers experience grows, effectively preventing any current flow, a condition called “reverse bias” (Figure 1.5).

The *pn* junction thus forms a solid-state switch (a.k.a. the “diode”). Consider: Apply a voltage of one polarity and current flows. Apply a voltage of the opposite polarity and no current flows; an on/off switch. Shockley shared the Nobel Prize with Bardeen and Brattain largely for explaining this phenomenon, and of course by wrapping predictive theory around it which led to the demonstration of the BJT. The result of that particularly elegant derivation is the celebrated “Shockley equation” which governs the current flow in a *pn* junction

$$I = qA\left\{\frac{D_n n_i^2}{L_n N_A} + \frac{D_p n_i^2}{L_p N_D}\right\}\left(e^{qV/kT} - 1\right) = I_S\left(e^{qV/kT} - 1\right) \tag{1.1}$$

where A is the junction area, V is the applied voltage, $D_{n,p}$ is the electron/hole diffusivity ($D_{n,p} = \mu_{n,p}\, kT$), $L_{n,p}$ is the electron/hole diffusion length, and I_S is the junction “saturation current” which collapses all of these factors into a single (measurable) parameter.

Observe, that all of the parameters in the Shockley equation refer to the minority carriers. If we build our junction with the *n* and *p* doping the same, then the relative contributions of the electron and hole minority carrier currents to the total current flowing will be comparable (to first order). Let us look closer at the operation of the junction. Under forward bias, electrons diffuse from the *n*-side to the *p*-side, where they become minority carriers. Those minority electrons are now free to recombine and will do so, on a length scale determined by L_n, and thus as we move from the center of the junction out into the neutral *p*-region, the minority electron population decreases due to recombination, inducing a concentration gradient as we move to the *p*-side, which drives a minority electron diffusion current. The same thing is happening with holes on the opposite side of the junction, and these two minority carrier diffusion currents add to produce the total forward bias current flow. What is the actual driving force behind the forward bias current in a *pn* junction? Recombination in the neutral regions, since recombination induces the minority diffusion currents. Alas, simple theory and reality are never coincident, and there a finite limits to the voltages that

1831

Michael Farady and Joseph Henry formulate “Faraday's law of induction”

ELECTRON DEVICES SOCIETY

1833

K.F. Gauss and W.Weber devise electromagnetic telegraph

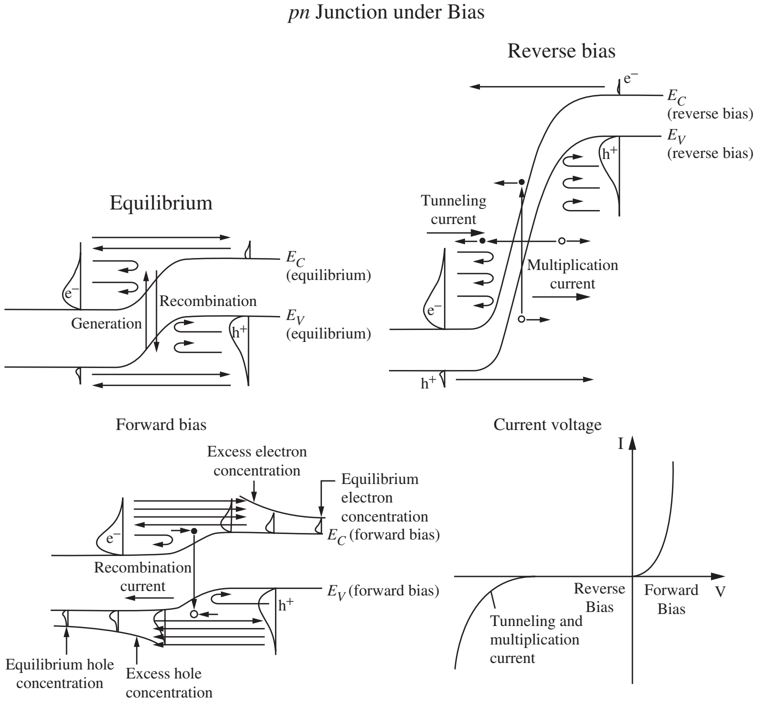


Figure 1.5 The pn junction under both forward and reverse bias, showing the resultant current–voltage characteristics. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press

can be applied to the diode, and how much current can be passed through it and how much voltage can be applied across it [3].

So, what makes the junction so useful? Well, as stated, it makes a nice on/off switch with low loss when forward biased, and it can provide very good electrical isolation when reverse biased. In power electronics the diode would be said to provide a “blocking” voltage, not allowing current flow in reverse bias up to some finite, and often huge, applied reverse voltage (hundreds to even thousands of volts). This is very useful. The diode can also function as a wonderful solid-state “rectifier”. Rectifiers are ubiquitous in power generation, conversion, and transmission, (e.g., to turn AC voltage into DC voltage). Finally, the diode can also emit and detect light, which is also extremely useful as a transducer for converting optical to electrical energy, and vice versa (see Chapters 16 and 20).

All of this said, however, the diode does NOT possess gain, and, thus, is insufficient for realizing complex electronic systems. From a transistor perspective, however, the *pn* junction can be used to make a tunable minority carrier injector, which, if cleverly employed, can indeed produce gain when carefully implemented

1836

Inductor coil is invented by Nicholas Callan

ELECTRON DEVICES SOCIETY

within a transistor. Importantly, one can trivially skew the relative magnitudes of the minority carrier injection from side-to-side in a *pn* junction by making the doping levels on one side of the junction much more heavily doped than on the other side. Let us imagine that the *n*-doping is far larger than the *p*-doping. Fittingly, this is referred to as a "one-sided" junction. In this scenario, it can be easily shown that electrons make up most of the total current flow in forward bias in such a junction. If we wanted to use a *pn* junction under forward bias to enhance the "forward-injection" of electrons into the *p*-region, and suppress the "back-injection" of holes into the n-region, we could simply use an $\mathrm{n}^{++} - \mathrm{p}^{-}$ junction as an "electron injector"! This will lead us directly to the BJT, a transistor with gain.

BJT versus FET

At a deep level, the BJT and the FET are closely related devices. Both have two *pn* junctions which are integral to their functionality. In an FET, a "gate" electrode is capacitively coupled (through the gate oxide) to the charge conduction path, altering the current flow from source to drain. In the BJT, the "base" electrode is directly tied to the charge conduction path, altering the current flow from emitter to collector. Thus, the differences between BJTs and FETs lie with the how the control terminal is electrically tied to the charge conduction path.

1.3 The Bipolar Junction Transistor and its Electronic Applications

The *pn* junction, as a two-terminal object, can be made to serve as an efficient minority carrier injector, but it does NOT possess inherent gain. This is the fundamental reason why we do not build microprocessors from diode-resistor logic. Diodes make excellent binary switches, but without a gain mechanism to overcome Nature's preference for attenuation, complex functions are not going to be achievable in practice. Let us imagine, however, that we add an additional third terminal to the device which somehow controls the current flow between the original two terminals. Let terminal 1 = the input "control" terminal, and terminals 2 and 3 have high current flow between them when biased appropriately by the control terminal. Then, under the right bias conditions, with large current flow between 2 and 3, if we could somehow manage to suppress the current flow to/from 1, we'd be in business. That is, small input current (1) generates large output current (from 2 to 3), and hence we have gain!

How do we do this in practice? Let us use two *pn* junctions, placed back-to-back, such that the control terminal (our #1; which we will call the "Base" terminal – B) is in the central *p* region, and the two high current flow path output terminals (our #2 and #3, which we will call the "Emitter" and "Collector" terminals – E, and C), are the two outside *n* regions (see Figure 1.6). Since the two central *p* regions are shared by both diodes, those can be coincident. That is, an *n* region separated from another *n* region by an intermediate *p* region actually contains two *pn* junctions.

1837

Wheatstone and Cooke file patent on electric telegraph

This telegraph only transmitted 20 of the 26 letters of the english alphabet …

… leaving out C, J, Q, V, X and Z

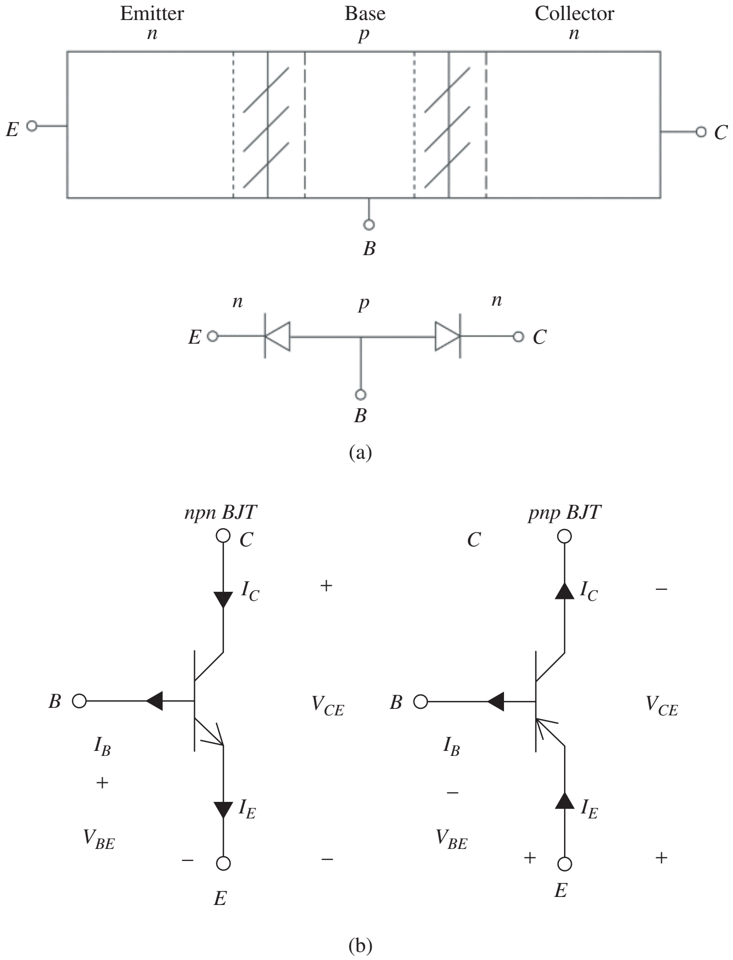


Figure 1.6 (a) Schematic of the two back-to-back pn junctions that form a bipolar junction transistor; (b) the circuit symbol of both doping polarity types are also shown. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press

Let us imagine forward biasing the emitter–base junction, and reverse biasing the collector–base junction, and then adding two more puzzle pieces: (1) We must dope the emitter very heavily with respect to the base, such that when we forward bias the emitter–base junction we have large electron flow from E to B and simultaneously suppress the hole flow from B to E (this is our tunable minority carrier injector!). (2) We must make the central base region VERY thin. Why? Well, if we don't, then the electrons injected from E to B will simply recombine in the base before they can reach the collector (to be collected and

1839

Photovoltaic effect discovered by Alexandre-Edmund Bequerel, …

ELECTRON DEVICES SOCIETY

… the foundation of modern solar cell technology

to generate the required large output current flow from E to C). Recall that the rough distance a minority carrier can travel before it recombines is given by the diffusion length ($L_{n,p}$). Clearly, we need the width of the *p*-type base region to be much, much less than this number; in practice, a few hundred nm is required for a modern BJT. The final result? We have created the *npn* BJT! (One could of course swap the doping polarities *n* to *p* and *p* to *n* and achieve the same result – a *pnp* BJT. We thus have two flavors of BJT, and this is often VERY handy in electronic circuit design.

Consider now how the BJT actually works: (1) The reverse-biased CB junction has negligible current flow. (2) The forward-biased EB junction injects (emits) lots of electrons from E to B, that diffuse across the base without recombining (because it is thin) and are collected at C, generating large electron flow from E to C (current). BUT, due to the doping asymmetry in the EB junction, while a large number of electrons get injected from E to B, very few holes flow from B to E. Forward electron current is large, but reverse hole current is small. That is: small input base current; large output collector. Gain! This is otherwise known in electronics as “current gain” (or β).

How do we make the BJT? Well, as might be imagined it is more complex than a *pn* junction, but even so, the effort is worth it. Figure 1.7 shows the simplest possible variant. Figure 1.7 also superposes both the equilibrium and forward-active bias energy band diagrams, with the carrier minority and majority carrier distributions, to help connect the *pn* junction physics to the BJT operation. Within the band diagram context, here is intuitively how the BJT works. In equilibrium, there is a large barrier for injecting electrons from the emitter into the base. Forward bias the EB junction and reverse bias the CB junction, and now the EB barrier is lowered, and large numbers of electrons are injected from E to B. Since B is very thin, and the CB junction is reverse biased, these injected electrons will diffuse across the base, slide down the potential hill of the CB junction, and be collected at C, where they generate a large electron current flow from E to C. Meanwhile, due to the doping asymmetry of the EB junction, only a small density of holes is injected from B to E to support the forward bias EB junction current flow. Hence, I_C is large, and I_B is small. Gain! A different visualization of the magnitudes of the various current contributions in a well-made, high gain, BJT, are illustrated in Figure 1.8.

Shockley’s theory to obtain an expression for β is fairly straightforward from basic *pn* junction physics (although you have two different ones to contend with obviously), provided you make some reasonable assumptions on the thickness of the base (base width $W_b \ll L_{nb}$). For the output and input currents under forward-active (amplifier) bias, we obtain:

$$I_C \cong qA\left\{\frac{D_{nb}n_i^2}{W_bN_{Ab}}\right\}e^{qV_{BE}/kT} = I_{CS}\,e^{qV_{BE}/kT} \tag{1.2}$$

$$I_B \cong qA\left\{\frac{D_{pe}n_i^2}{L_{pe}N_{De}}\right\}e^{qV_{BE}/kT} = I_{BS}e^{qV_{BE}/kT} \tag{1.3}$$

where the “b” and “e”, or “B” and “E”, subscripts stand for base and emitter, respectively. Interestingly, the current gain does not to first-order depend on bias voltage, the size of the junction, or even the bandgap! We finally obtain,

$$\beta \cong \frac{I_C}{I_B} = \frac{I_{CS}}{I_{BS}} \cong \left\{\frac{D_{nb}L_{pe}N_{De}}{D_{pe}W_bN_{Ab}}\right\} \tag{1.4}$$

1844 Morse's telegraph first used for transmission from Baltimore to Washington

Electron Devices Society

1850 Mica used as insulator in capacitors

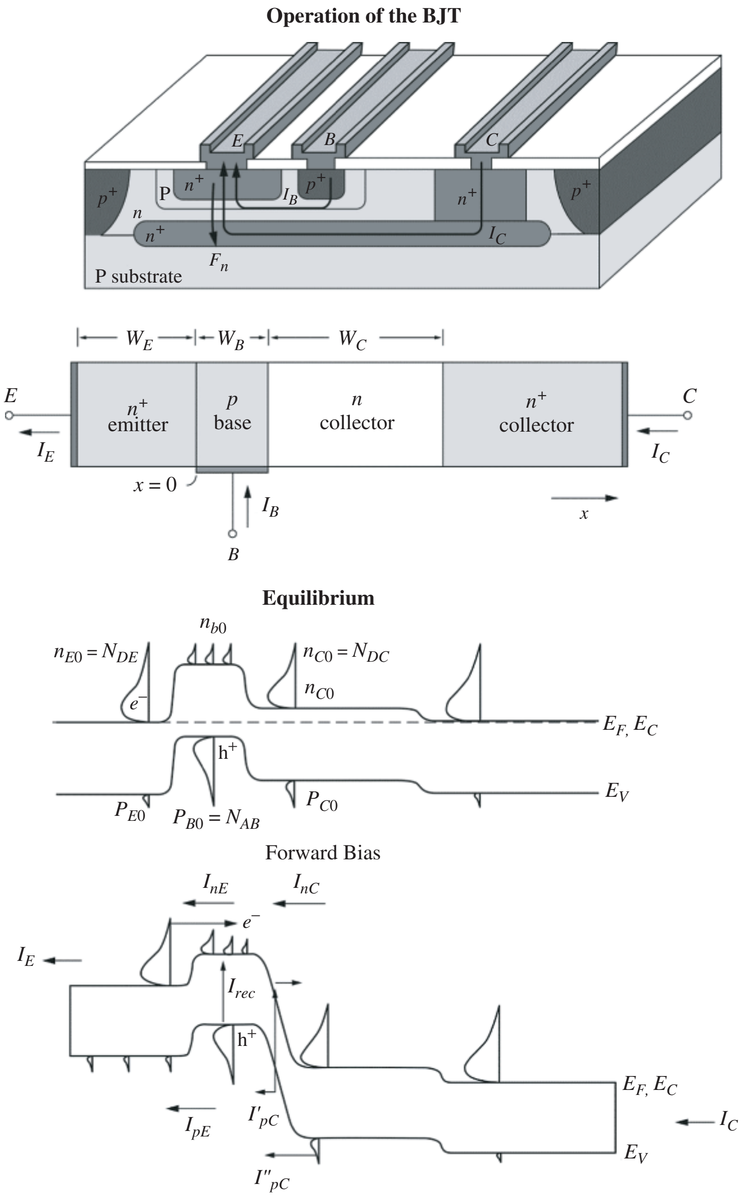
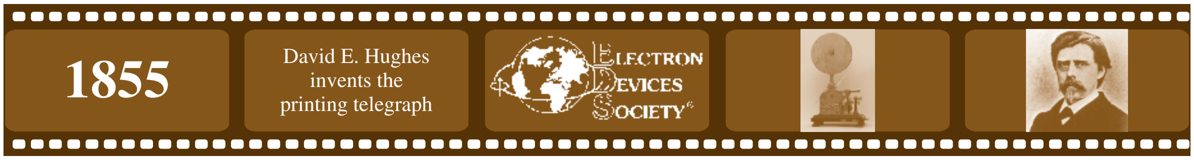
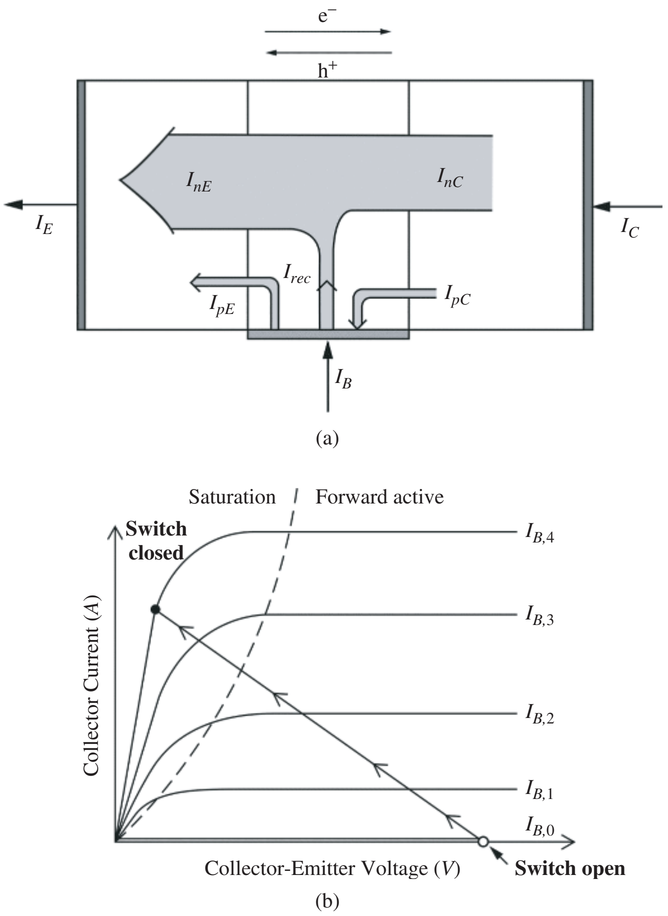


Figure 1.7 Basic structure and operational principles of the bipolar transistor. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press


1855

David E. Hughes
invents the
printing telegraph

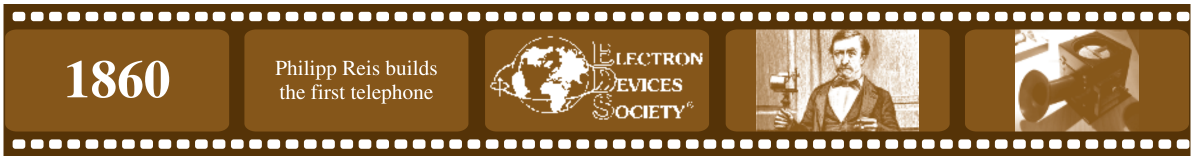
Clearly, the current gain is a tunable parameter, giving us great flexibility in design. A common way to plot the BJT current–voltage characteristics is shown in Figure 1.8, where linear I_C is plotted versus linear V_{CE}, as a further function of I_B. Since I_C is larger than I_B, the gain is implicit here. This plot is known as the output "family" or "output characteristics". We use the output family to define the three regions of operation of the BJT: (1) "forward-active" (EB junction forward-biased; CB junction reverse-biased); (2) "saturation" (both EB and CB junctions forward-biased), and (3) "cut-off" (both EB and CB junctions reverse biased). As indicated, forward-active bias is typically for amplifiers, and as we will see, switching between cutoff and saturation will make an excellent regenerative digital switch!



(a)

(b)

Figure 1.8 Sketch of (a) the relative current contributions of the bipolar transistor and (b) the resultant current–voltage characteristics. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press



How fast can transistors switch states (on to off)? The current speed record for a bipolar transistor digital switch is less than 10 picoseconds (0.000000000010 seconds – 10 trillionths of a second!). What limits that speed? Intuitively, the speed is limited by the time it takes the electrons to be injected from the emitter, transit (diffuse across) the base, and then be collected by the collector. In other words, a transistor can't be faster than it takes the charge to move through it. In most transistors, step two is the limiting one, and the so-called "base transit time" (τ_b) sets the fundamental speed limit on how fast the BJT can switch. A first-order base transit time expression can be easily derived,

$$\tau_b \cong \frac{{W_b}^2}{2D_{nb}} \tag{1.5}$$

Hence, the smaller τ_b is, the faster the BJT can switch. Clearly, making W_b as small as possible gives us a double benefit. It helps increase the current gain, yes, but even more importantly, it makes the transistor faster – quadratically!

So what does the BJT do for us? Let's restate some points for clarity. This beautiful three-terminal semiconductor device, if constructed correctly, will exhibit a (tunable) gain. Gain is the key to success in building any electronic system; hence the deserved fame of the BJT. This intrinsic gain will allow us to create a wide variety of amplifiers for use in a myriad of electronics applications. Amplifiers that take: (1) A small input current and turn it into a large output current (a.k.a., a "current amplifier"); (2) a small input voltage and turn it into a large output voltage (a.k.a., a "voltage amplifier"); (3) a small input current and turn it into a large output voltage (a.k.a., a "transconductance amplifier"); and (4) a small input voltage and turn it into a large output current (a.k.a., a "transimpedance amplifier"). Transconductance (g_m) in the electronics world just means the incremental change in current divided by the incremental change in voltage. As a real-world example of amplifiers-in-action, at the input of your cell phone you have a hand-crafted voltage amplifier that takes the tiny little RF signals and boosts them to a level sufficient to manipulate and decode them (see Chapter 14). In a receiver for a fiber optic link, you have a hand-crafted transimpedance amplifier that interfaces with the input photodetector, to change the in-coming photonic signals into electronic signals for processing (see Chapter 20).

In addition to building amplifiers, gain also allows us to construct nice regenerative binary switches. As can be seen in Figure 1.8, if the input base current I_B (or input voltage V_{BE}) is zero, the output current I_C is zero, the on/off switch is now open and the output voltage V_{CE} is thus high. Let us call that state a logical "1". Conversely, if the input current I_B (or input voltage V_{BE}) is large enough to turn on the transistor, the output current I_C is large, output voltage V_{CE} drops to a low value, and the on/off switch is now closed. Let's call that state a logical "0". A regenerative binary switch!

1.4 Optimization of Bipolar Transistors

There are two typical performance metrics (or figures-of-merit: FoM), which indicate how fast or how high a frequency a bipolar transistor can operate. The first is the so-called "cutoff frequency" (f_T), the frequency at which the AC (alternating current) current gain becomes unity. The f_T is simply given by the inverse of total transit time (τ_{ec}) from the emitter to the collector ($f_T = 1/2\pi\tau_{ec}$) and, thus, gives an estimate of the speed-limit of the BJT switch and is a good FoM for digital circuits. As described above, to improve f_T (make it larger) major attention must be paid to make the base width as narrow as possible.

1866 | Alfred Nobel invents dynamite | Electron Devices Society | 1869 | Opening of Suez Canal

Here, heavily doped polysilicon is introduced to form the emitter region. This idea is widely used even in the modern BJTs and is called a "poly-emitter". The poly-emitter is utilized to form a shallow out-diffusion for the emitter impurities, and thereby allows both a thin base and emitter design. The poly-emitter has an additional advantage; namely, that the very thin native interface oxide which naturally occurs between the polysilicon and the single-crystal silicon acts as an effective barrier to prevent the minority carrier (hole) back-injection from B to E. It is necessary to increase the base doping concentration for a narrower base (to avoid the disappearance of the neutral base, so-called device "punchthrough") but the emitter doping concentration already reaches its maximum value (limited by solid solubility), so the current gain in a scaled BJT naturally decreases due to the low emitter injection efficiency (the ratio of the injecting electrons from E to B to the injecting holes from B to E). Therefore, the poly-emitter interfacial oxide helps to increase the current gain and is a very useful secondary by-product. However, as the emitter scaling progresses, the interfacial oxide causes the problem of the high emitter resistance and, thus, must be carefully optimized.

The second important BJT FoM is the so-called "maximum oscillation frequency" ($f_{\max}$), the frequency at which the unilateral power gain becomes unity. The unilateral power gain is the forward power gain in a feedback amplifier, so it is a suitable index for many analog and RF circuits. The $f_{\max}$ is approximately given by, $f_{\max} = \sqrt{f_T/8\pi C_c r_b}$, where C_c is collector capacitance and r_b is base resistance. The critical difference between f_T and $f_{\max}$ is as follows. The f_T is a FoM determined from the one-dimensional (vertical) structure, but the $f_{\max}$ is a FoM which includes the two-dimensional (planar) structure of the device, because the parasitic C_c and r_b appear in the equation. This means, to improve $f_{\max}$, it is essential to minimize the parasitic capacitances and resistances of the planar structure. As can be seen in Figure 1.7, the intrinsic region for BJT is a one-dimensional structure just under the emitter. The other areas of the transistor structure are provided mainly to lead the base and collector current to their electrodes, so they are non-essentially the operation of the device. To improve lateral parasitics, several important transistor structures and process sequences (e.g., the so-called "self-aligned transistor structure" or "self-aligned fabrication process") have been developed. Figure 1.9 shows a typical self-aligned BJT structure formed by using a self-aligned fabrication process. To reduce C_c, it is very important to reduce the junction area between the base and collector. Therefore, in this self-aligned transistor, the base electrode constructed by a polysilicon film is placed on a thick oxide layer, minimizing C_c in the extrinsic base region. To reduce r_b, the narrow (typically 100 nm wide or less) space between the polysilicon emitter and the polysilicon base is defined by the thickness of the insulator which is formed on the side of emitter or base polysilicon. This is the origin of the usage of the term "self-aligned", that is, the edge of the emitter and base is automatically defined by

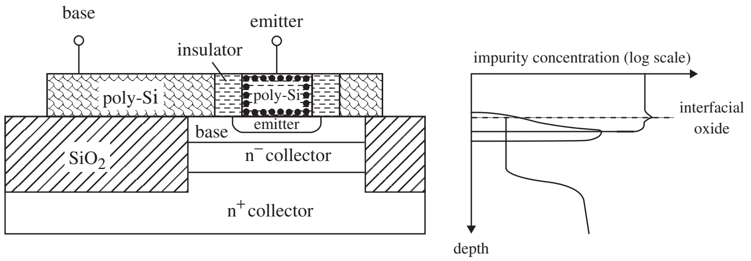
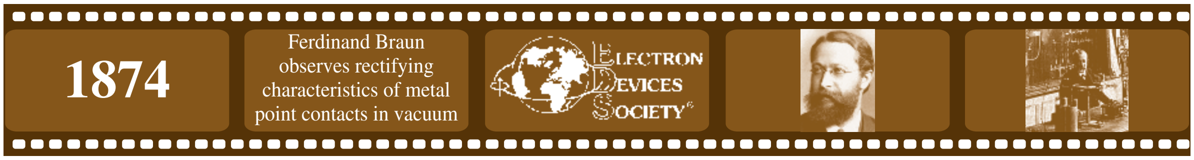


Figure 1.9 Self-aligned bipolar transistor structure and impurity profile under the emitter


1874

Ferdinand Braun observes rectifying characteristics of metal point contacts in vacuum the structure, independent of the lithography used. In the case of a non-self-aligned planar configuration defined by lithography, as shown in Figure 1.7, the base current must flow along a long (about 1 μm long or more) path, so it is difficult to decrease r_b. On the other hand, in a BJT formed using a self-aligned process, the distance separating from the emitter and base is very small, so this can be used to effectively reduce r_b.

Finally, the breakdown voltages (which set the maximum useful operating voltage of the BJT) are key transistor parameters for improving the high-speed and high-frequency characteristics of BJT. There is a fundamental trade-off between the speed (f_T) and the breakdown voltage (BV_{CEO}, the breakdown voltage between the collector and emitter when the base is open-circuited), often termed the "Johnson limit" [6]. The Johnson limit is derived only from considering fundamental issues associated with carrier transport, and predicts an achievable $f_T \cdot BV_{CEO}$ product of 200 GHzV, though in practice this value is significantly higher. The concept of a constant $f_T \cdot BV_{CEO}$ product in a BJT is useful for designing the collector region of the BJT, since it captures the tradeoff between achievable speed and operating voltage.

1.5 Silicon-Germanium Heterojunction Bipolar Transistors

The basic concept of the "heterojunction" bipolar transistor (HBT) was proposed by Shockley in the original BJT patent (refer to the history in [3]), and the basic theory of the HBT was published by Kroemer in 1957 [7]. Figure 1.10 shows the equilibrium energy band diagram, with the minority and majority carrier distributions, of the wide bandgap emitter HBT. The wide bandgap emitter creates a large barrier for injecting holes from the base into the emitter, thus increasing the current gain. Many III-V compound semiconductors (e.g., GaAs or InP) have been successfully applied in HBTs by virtue of their compositionally-adjustable growth technology which can tailor the bandgap for a specific need (called "bandgap engineering"). III-V HBTs benefit from this approach and can provide a large advance in performance over BJTs (see Chapter 14). However, bandgap engineering did not extend into the world of Si-based technologies for many decades, even though the basic idea was envisioned early-on for HBTs based on silicon-compatible silicon-germanium (SiGe) alloys. The lattice constants between Si and Ge differ by roughly 4.2%, so the SiGe films grown on Si are compressively strained. The criterion giving the stability of such pseudomorphically grown strained SiGe films on Si indicates a maximum "critical thickness"

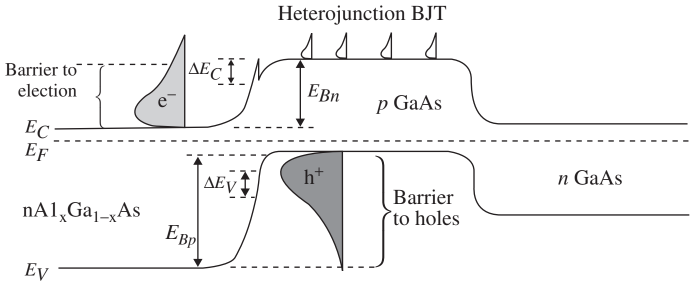
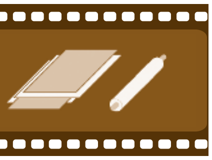


Figure 1.10 Basic idea behind the wide bandgap emitter heterojunction bipolar transistor. Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press

1875 1876 ELECTRON DEVICES SOCIETY Paper becomes used as insulator in capacitors



of SiGe film for a given Ge content [8]. SiGe films which are device quality, meaning the SiGe films remain stable after thermal processing, were first epitaxially grown in the mid-1980s, and shortly thereafter the first SiGe HBTs were demonstrated [5].

The bandgap of Ge (0.66 eV) is smaller than that of Si (1.12 eV), so the SiGe HBT has a narrow bandgap base, differing from the wide bandgap emitter HBT. The compressive strain associated with sandwiched SiGe base layer between Si emitter and collector layers produces an additional bandgap shrinkage. As a result, a bandgap reduction of about 70–80 meV for each 10% of Ge content can be utilized in device engineering. Figure 1.11 shows the basic structure and forward-active bias energy band diagram of a SiGe

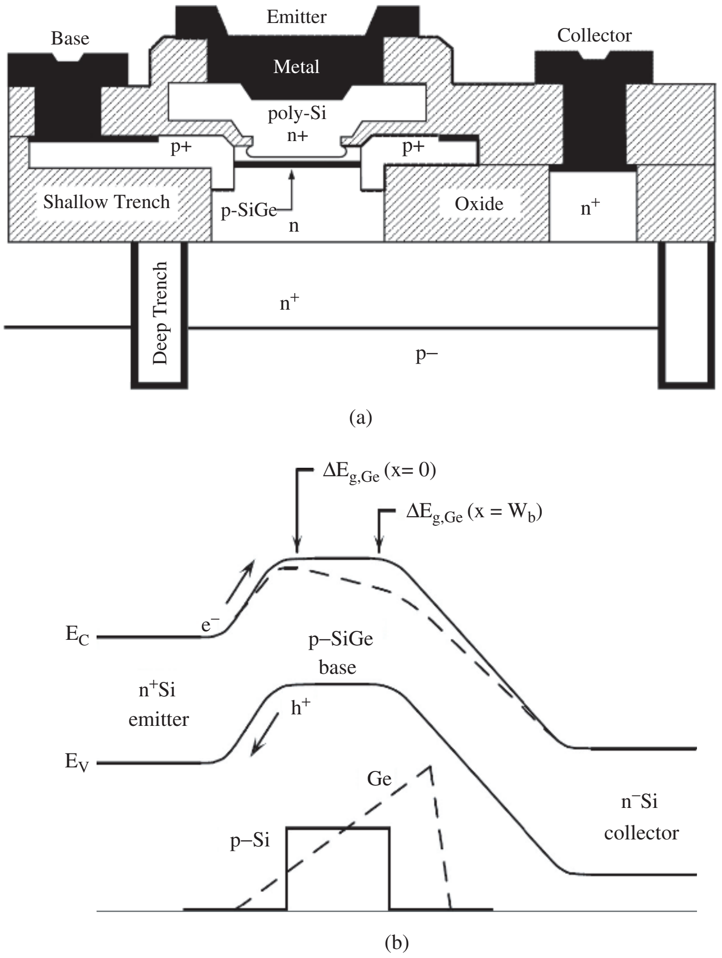
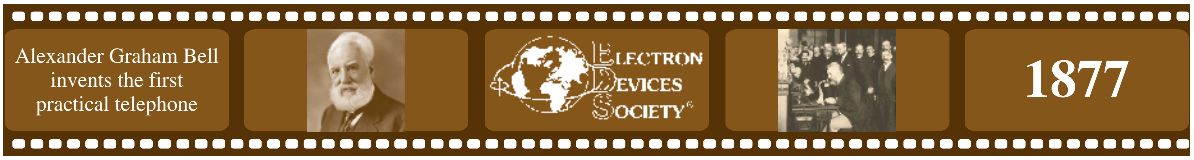


Figure 1.11 Sketch of (a) the basic structure and (b) the band structure and doping profile of the silicon-germanium heterojunction bipolar transistor (SiGe HBT). Reproduced with permission from Cressler, J. D.; *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*; 2009, Cambridge University Press



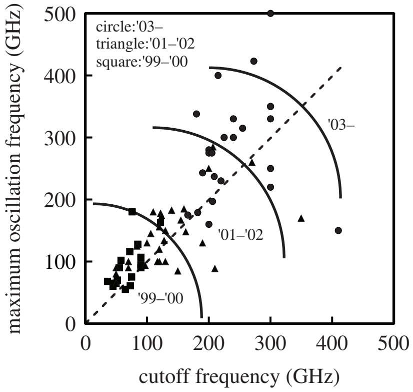


Figure 1.12 Evolutionary improvement in cutoff frequency and maximum oscillation frequency from 1999–2011 for SiGe HBTs

HBT. Similar to the wide bandgap emitter HBT, the emitter injection efficiency effectively increases due to the Ge-induced band offset occurred in the valence band.

After the first demonstration of functional SiGe HBT in 1987 [9], the development of SiGe HBTs evolved rapidly and their performance has dramatically improved from the mid-1990s to present. For Si BJTs, the peak f_T is limited to approximately 50 GHz. However, using a SiGe HBT, both f_T and $f_{\max}$ go rise above 300 GHz, as shown in Figure 1.12. In the early stages of evolution, the SiGe HBT had a non-self-aligned structure, so only f_T was improved by the shrinkage of the base width and bandgap engineering. However, SiGe HBTs soon incorporated self-aligned transistor structures, with rapid improvement in transistor $f_{\max}$. The schemes to fabricate self-aligned SiGe HBTs are roughly categorized into two types, depending on the SiGe epitaxial growth technologies used: selective or blanket epitaxial growth [10]. Recently, attention has been placed on achieving ultra-high $f_{\max}$ due to the emerging applications such as terahertz wireless systems.

One of the most important aspects of SiGe HBTs is that it can be easily combined with Si CMOS on the same wafer to enable highly-integrated systems. So-called SiGe BiCMOS (SiGe HBT + Si CMOS) technologies can be constructed using well-established Si-based processes and are 100% silicon manufacturing compatible. This represents a fundamental difference between SiGe HBTs technology and III-V HBTs (see also Chapter 14). The wide-spread application of SiGe HBTs in high-speed digital and RF/analog integrated circuits offer ample evidence to this crucial advantage enjoyed by SiGe BiCMOS technology (see examples in [10]).

References

[1] J. Bardeen and W. H. Brattain, "The transistor, a semiconductor triode", *Physical Review*, vol. 74, pp. 230–231, 1948.

[2] W. Shockley, M. Sparks, and G. K. Teal, "p-n junction transistors", *Physical Review*, vol. 83, p. 151, 1951.

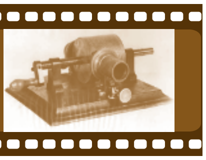
[3] J. D. Cressler, *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*, New York, NY, Cambridge University Press, 2009.

[4] K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Edn, New York, NY, John Wiley & Sons, Inc., 2002.

1878

Thomas Alva Edison invents the phonograph

Electron Devices Society



[5] J. D. Cressler (ed.), *Silicon Heterostructure Handbook: Materials, Fabrication, Devices, Circuits, and Applications of SiGe and Si Strained-Layer Epitaxy*, Boca Raton, FL, CRC Press, 2006.

[6] E.O. Johnson, “Physical limitations on frequency and power parameters of transistors”, *RCA Rev.*, vol. 26, pp. 163–177, 1965.

[7] H. Kroemer, “Theory of a wide-gap emitter for transistors”, *Proc. IRE*, vol. 45, pp. 1535–1537, 1957.

[8] J. W. Matthews and A.E. Blakeslee, “Defects in epitaxial multilayers– I:misfit dislocations in layers”, *J. Cryst. Growth*, vol. 27, pp. 118–125, 1974.

[9] S. S. Iyer, G. L. Patton, J. M. C. Stork, *et al.*, “Silicon-germanium base heterojunction bipolar transistors by molecular beam epitaxy”, *Tech. Dig. IEEE Int. Elect. Dev. Meeting*, pp. 874–876, 1987.

[10] K. Washio, “Silicon-germanium (SiGe) heterojunction bipolar transistor (HBT) and bipolar complementary metal oxide semiconductor (BiCMOS) technologies”, Chapter 18 in, *Silicon-Germanium Nanostructures* (eds Y. Shiraki and N. Usami), Cambridge, Woodhead Publishing, 2011.

1879

Edison invents
the electric light bulb

ELECTRON
DEVICES
SOCIETY

1880

Jaques and Pierre Curie
discover the piezo electric
effect in crystals