The SiGe bipolar transistor: History, present, future

Ted Johansson
Infineon Technologies Sweden AB, SE-164 81 KISTA, Sweden

Abstract
The history, the present state, and the future trends of the SiGe bipolar transistor are reviewed. The development of the SiGe transistor is described from a historical perspective, starting with the invention of the first (bipolar Ge) transistor, the planar process and the integrated circuit. I^2L is discussed as an example of the BIP IC vs. MOS IC competing developments, and finally the transformation of the implanted-base Si transistor into the SiGe transistor is described. The present state is reviewed using SiGe transistor theory and state-of-the-art device structures, leading to SiGe BiCMOS and PNP device integration. Design and optimization of SiGe power amplifiers for different communication
systems are reviewed as example of real-world devices. Finally, future trends of the SiGe transistor are discussed.

1. Introduction

The transistor is the heart of all electronic circuits. It is the active device that amplifies current or voltage, and makes it possible to perform a large number of electronic functions. Examples from the analog world are radio building blocks, like amplifiers, mixers, and filters. Examples from the digital world are logic circuits, from the simplest two-transistor inverter to the largest product-specific designs, which can today contain up to billions of transistors on a single chip.

The bipolar transistor was developed in parallel with the advances in silicon material and process technology from the start in the late 1940s. In the 1970s, before the MOSFET had taken over as the workhorse for digital applications, the bipolar transistor formed the basis for the integrated circuit market and was foreseen a bright future. However, the MOSFET turned out to have a higher potential for high yield and a still unbroken record of miniaturization and complexity scalability, leading to the billion-transistors CMOS ICs we have today [1].

Silicon bipolar technologies matured considerably during the 1980s, mainly as the active device in high-performance low-density low-complexity circuitry. The advances in wireless and optical telecommunication during the last decades of the 20th century motivated the industry to spend greater efforts than ever on improving and understanding semiconductor devices, especially their high-speed properties. The silicon-germanium (SiGe) bipolar transistor was developed as a gradual improvement of the conventional implanted base bipolar transistor and reused many of the features of these devices.

The SiGe bipolar transistor became even more successful with the merge of MOS into BiCMOS technology and thus was able to support the increased complexity and integration level of modern system-on-chip solutions such as one-chip transceivers for mobile phones or wireless local networks. The SiGe BiCMOS technology is today commercially available at all large semiconductor manufacturers as well at silicon foundries, to be used even by the smallest fabless design house in high-performance wireless products.

The SiGe bipolar transistor has extended the conventional silicon bipolar transistor well into domains previously only been occupied by III-V-based and similar technologies, while maintaining every advantageous aspect of industrial-grade silicon technology such as scalability, manufacturability and year-by-year gradual improvements. Today, with $f_T$
values reaching well over 200 GHz for commercial-class devices, SiGe technology is seriously challenging III-V technologies in its own backyard.

As we write in 2005, it can be interesting to look back and follow the path of the development of this device, check the status of the present devices, and speculate about what is expected to be achieved in the future for the SiGe bipolar transistor.

2. History

2.1. The invention of the bipolar transistor

The SiGe bipolar transistor’s history is in many aspects the history of the semiconductor technology and the history of the bipolar transistor.

On December 16, 1947, J. Bardeen and W. Brattain at Bell Telephone Laboratories, Murray Hill in New Jersey, USA, almost by mistake pushed two electrodes into a half inch piece of germanium. The electrical current coming out of the germanium was hundred times larger than what went in!

In fact, the invention was the result of a long and systematic work performed at Bell Labs to find a replacement for the vacuum tube. Already by the early 1930’s, the engineers realized that the practical limit of the vacuum tube had been reached. Invented by Lee de Forest in 1906 as the “audion”, it was based on a phenomenon discovered by Thomas Edison in 1883. Edison noted that a current leaked through the empty bulb when he applied a positive voltage to tiny metal plate into the glass envelope of a light bulb. John A. Fleming exploited the effect to build a rectifying device, which was suited as a detector of radio waves. de Forest introduced a third electrode, called the “grid”, and was able to control the current through the tube by applying different voltages to the grid. In 1915, the tube was taken in use for long-distance telephone system, and by the mid-1930s, the quality of the tubes had improved to last ten years, requiring only a tenth the power of the original tubes.

In 1945, a rapid expansion of telecommunication in the post-war era was foreseen. However, several areas of major concern that would pose serious limitations on future systems were identified. One area was the use of high-frequency signals in the systems. A reliable method for rectifying and amplifying signals was needed, but the slow speed of the electro-mechanical relays and the high power dissipation and poor reliability of vacuum tube amplifiers could not fulfill the new demands.

In 1925, J. E. Lilienfield patented the field-effect principle [2]. An electric field applied through the surface of a semiconductor could modify the density of charge in the body of the material and, thereby, change its conductivity. It was like adding a third electrode to a crystal rectifier to amplify signal, just as de Forest had added the grid to Fleming’s rectifying
tube. If the theory proved to be correct, the new semiconductor device would work as an amplifier, but using much less power and being much smaller than the tube. This was exactly what was needed for the new systems. But to that point, all attempts to make such a device had been unsuccessful.

Bell Labs established a group to focus on understanding semiconductors, in particular to explore the feasibility of making an amplifier. Two critical decisions for the group’s priorities were made. The first was to focus on crystals of silicon and germanium and to ignore more complex materials frequently used in prior investigations. The second decision was to follow the ideas set forth in field-effect theories as being the most promising way to a working device.

By the fall of 1947, the work had focused on trying to confirm Bardeen’s theory regarding surface state traps using metal probes on the surface of germanium crystals. The theory seemed to be correct. For the first time there was some insight why the attempts to build a field-effect device did not work. In the course of their work, they tried to modify the surface states with electrolytes surrounding the metal contacts to the germanium surface. Brattain replaced the electrolyte with an evaporated gold spot adjacent to the point contact.

Finally, they replaced both contacts by an arrangement of two strips of gold foil separated by just a few mils and pressed onto the germanium surface (see Figure 1). With one gold contact forward biased and the other reverse biased, power gain was observed. This was December 16, 1947. Almost immediately, Bardeen and Brattain understood, in a general way, what was going on, as apparent from their lab notebooks.

Figure 1. The first transistor: the point-contact semiconductor amplifier. (a) Photo of Bardeen’s and Brattain’s amplifier. (b) Cross-sectional diagram of the amplifier. (It would be interesting for readers to figure out how the device works.)
On December 23, 1947, the new device was demonstrated to top Bell Labs management. For this demonstration, the transistor was incorporated in an oscillator circuit, which is a critical test for the existence of power gain. The circuit was also used to amplify speech, with no noticeable degradation in the sound quality. In Figure 1(a) is a photo of the first point-contact transistor setup from 1947, and in Figure 1(b), a cross-sectional diagram of the device is shown.

The invention was announced six months later, on June 30, 1948 at Bell Labs’ headquarters in New York [3]: “We have called it the Transistor, T-R-A-N-S-I-S-T-O-R, because it is a resistor or semiconductor device which can amplify electrical signals as they are transferred through it from input to output terminals. It is, if you will, the electrical equivalent of a vacuum tube amplifier.” The transistor was invented and the information age began!

The reason for the delay of the public announcement was that time was needed to gain an understanding of the device and to investigate the patent position. The first device was a point-contact transistor [4]. The fundamental physical principles behind the invention were not initially understood. Brattain and Bardeen thought it was a surface or field-effect mechanism. Shockley disagreed and was frustrated with himself that he had not personally played a greater role in the discovery. He was convinced that the amplification was not due to a surface effect but rather was related to the bulk material. Shockley worked feverishly over the next few weeks and by late January of 1948, he had completed a thorough formulation of positive-negative (p-n) junction theory and the role played by the injection of minority carriers in forward bias and their collection in reverse bias. In his analyses, Shockley also suggested the design for an improved version, the n-p-n junction transistor, which would prove to work better for volume production and later for integration, avoiding the whiskers on the semiconductor surface [5]. It was soon to replace the point-contact transistor when the manufacturing techniques evolved and silicon was taken in use.

Because of Shockley’s disappointment for not immediately understanding or foreseeing the principle for the bipolar transistor, a chasm broke out between Bardeen/Brattain and Shockley when the patent applications to protect the invention were to be defined. It was made worse when photos were to be taken for the newspapers and magazine writing articles about this new device. In all photos, Shockley took the central position, while Bardeen and Brattain had to act more as an audience.

The problem was now to experimentally verify Shockley’s theories and to build better devices. Later in 1948, G. K. Teal and J. B. Little succeeded in growing a single crystal of germanium by slowly pulling a seed crystal
out of a melt of high-purity material. Using crystals grown this way, experiments showed that useful devices could be made according to Shockley’s junction transistor theory, and that the surface conduction did not play any essential role for the operation of the device. However, it took another two years to actually build a working junction transistor.

The potential of the transistor was soon realized. Shortly after Shockley’s patent was issued in September 1951, Western Electric began licensing the rights to manufacture transistors for a 25 thousand US dollar fee. Six months later, forty companies had paid a similar fee and over one hundred participants were gathering at Bell Labs for a Transistor Technology Symposium to learn the transistor manufacturing art. In this way, the technical contributions to solve material and fabrication problems as well as continued research in the areas were to be shared by many. The rapid advances made in transistor manufacturing from 1952 to 1960 were probably the result of an open sharing of technology developed at various industry and university laboratories and presented at regular open technology symposiums. The first transistors used germanium and were very sensitive to changes in temperature, and silicon junction transistors, which were first fabricated in 1954 by a group headed by Teal, who later joined Texas Instruments, soon replaced them.

Texas Instruments also created the first mass-market applications for the transistor, the portable transistor radio, which came on the market in October 1954 and soon became a huge success. The military, to which cost was of no concern, offered the first larger-scale use of the transistor. In 1954, the first fully transistorized computer was built, using 700 point-contact transistors and more than 10 000 germanium crystal rectifiers in its circuits.

In 1955, Shockley left Bell Labs for California, where he wanted to earn his fortune by fabricating and selling semiconductor devices. He decided to set up his office and labs in Palo Alto, close to Stanford University to “attract outstanding technical personnel for our group and permit close association with the University.” This very much contributed to the birth of Silicon Valley.

Bardeen, Brattain and Shockley were awarded the 1956 Nobel Prize in physics for the invention of the transistor.

2.2. The planar process and the invention of the integrated circuit

During the 1950s, a number of important inventions were made which have become the bases for fabricating ULSI-scale integrated circuits. Advancements in material quality as well as process steps such as
diffusion, epitaxy and especially the planar process were important milestones.

In the summer of 1959, a very important patent was filed for the “IC”, the Integrated Circuit, a small block of semiconductor material containing several different types of isolated electronic devices like transistors, resistors and capacitors [6]. They were directly connected using a thin layer of aluminum metallization over an oxide layer to provide interconnects and to form a circuit. The foundation for this structure was the planar process [7], which made the silicon transistor structure much easier to manufacture and miniaturize, compared to the existing mesa isolated transistor. At that time, all connections were made by hand, similar to bonding techniques. The flat surface of the planar transistor enabled electrical connections to be made directly on the surface, using evaporated and plated metal layers. Although not as significant scientific breakthrough as the transistor, the invention of the IC revealed the potential for extending the cost and operating benefits of transistors to every mass-produced electronic circuit, including today’s computers and electronics for telecommunication. Figure 2 shows the original drawings from the 1959 patent application.

The patent was filed by Robert Noyce, at that time working for Fairchild Semiconductors in California, but soon to become one of the three founders of Intel, the largest semiconductor company today.

However, just a few months before Noyce, Jack Kilby from Texas Instruments (TI) had filed another, very similar patent [8], also describing the Integrated Circuit. Using germanium substrate and mesa-shaped isolated regions, small circuits such as oscillators and flip-flops had been fabricated with wire bonding interconnects. But Kilby and TI had been in a hurry when filing the patent application, afraid of rumors that RCA already was filing a patent on a similar device. The question about interconnection

---

**Figure 2.** The integrated circuit according to Noyce, showing a planar structure with evaporated aluminum interconnections adherent to the surface [6].
was not yet solved, although Kilby was also thinking of providing some metal on top of the semiconductor. Instead, the drawing in the patent shows bondwires to interconnect the devices; a method not particularly suited for mass fabrication and scaled dimensions. Figure 3 shows the original IC drawing from Kilby’s patent 1959.

Both companies soon realized this to be very important patents, and several long disputes in court followed. Noyce and Fairchild finally won in the U.S. Supreme Court (“Kilby vs. Noyce”), but Fairchild and TI settled an unofficial agreement before the final court decision and cross-licensed the patents. The patents were then licensed to the other companies in the IC business, because now it was 1970 and the IC was already a multi-billion dollar industry. The patents on the IC have generated a lot of money for Fairchild and TI.

In fact, when the complete stories were revealed, both Noyce and Kilby independently had invented the IC. Kilby could show by his notebooks that he was the first to have integrated devices on a chip, while Noyce had concentrated on the metallization interconnects on his chip, and thus both have been given the honor of having invented the IC.

Jack Kilby was in 2000 awarded the Nobel Prize in physics for the integrated circuit. Unfortunately, Robert Noyce died in 1990; he would certainly have shared the prize with Kilby.

![Figure 3](image-url)

**Figure 3.** The integrated circuit according to Kilby: germanium flip-flop using mesa transistors, bulk resistors, diffused capacitors, and air isolation of the components [6].

### 2.3. BIP IC vs. MOS IC: The I²L

During the 1970s, bipolar and MOSFET integrated circuit technology were competing and for a long time, it looked like the bipolar would win the race. Integrated injection logic or I²L is a low-power bipolar technology suitable for integration. Figure 4(a) shows a schematic of the two tightly integrated devices that forms the gate and in Figure 4(b) is shown a device cross-section of an early I²L gate implementation. The I²L consists of a lateral
PNP transistor (Q1) and an inverted vertical NPN transistor (Q2) with multiple collector contacts.

The I\(^2\text{L}\) has an interesting history [9]. It was developed unknowingly in parallel by Hart and Slob [10] at Philips in Holland and by Berger and Wiedmann [11] at IBM in W. Germany in the early 1970s. Both teams’ findings were presented at the 1972 IEEE Solid-State Circuits Conference, but the approaches were quite different.

The IBM team had developed the device from deep device knowledge and gradual improvements while trying to improve the devices for bipolar ICs and demonstrated this new technology using an eight-transistor adder chip. The Philips team had started the development a year later than the IBM team, with an approach to reduce the power consumption using light, then turned the transistor upside down, and finally replaced the light source with a hole injector, arriving at same structure as IBM. For the conference, they had already built four chips, one of them containing over 1000 gates, forming a battery-operated 8-digit pocket calculator.

The I\(^2\text{L}\) became very popular during the 1970s; it really revitalized bipolar technology, which already was suffering from the competition with the emerging MOS devices. The unique features of I\(^2\text{L}\) were suitable for applications that combined both digital and analog circuit on a single die and could offer a low-power solution for applications that did not need high operating speed. One large application for I\(^2\text{L}\) was circuits for color television and Philips produced a huge amount of such chips during 1970s and onwards. In fact, the I\(^2\text{L}\) was considered during the 1980s to be superior to BiCMOS.

In a more modern perspective, the I\(^2\text{L}\) suffers from a relatively poor dynamic performance since the gate delay is primarily determined by stored charge in parasitic diodes associated with the extrinsic base regions.
of the I²L gate. Today, high-performance bipolar logic circuits are usually realized using emitter-coupled logic (ECL). However, ECL is rather limited for large circuits since it has relatively low packing density and high power dissipation.

Interestingly, there has been a renewed interest in I²L motivated by the impressive progress of the SiGe bipolar transistor [12]. The SiGe technology offers the possibility of bandgap engineering to improve the deficiencies in the stored charge in the I²L gate. Figure 5 shows a device cross-section of such a modern SiGe-based I²L device structure. The device structure has a vertical P-N-p(SiGe), which attempts to improve the poor dynamic performance using the heterojunction to limit the hole injection back into the base of the P-N-p injector transistor and to the substrate. The SiGe collector of the injector transistor is merged with the base of the N-p(SiGe)-N switching transistor, and the N substrate forms the emitter of the switching transistor. The resulting main advantages are high beta, reduced hole injection into the base of the injector which reduces charge storage and improves switching time, and the use of MBE or LPCVD for forming the critical layers which results in good reproducibility of critical parameters.

![Figure 5. Modern surface-fed I²L (C- I²L) gate featuring SiGe.](image)

### 2.4. The dawn of the SiGe HBT

The evolution of the bipolar transistor continued during the 1970s by utilizing new techniques developed to create smaller devices and more precise doping profiles. Ion implantation created the possibility of improved emitter and base profile design [13], while polysilicon emitters and self-aligned device structures further improved the device performance and led to the double-poly implanted base transistor in 1981 [14] which a majority of the high-performance commercial bipolar processes today, Si as well as SiGe, are based on.
The SiGe heterojunction bipolar transistor (HBT) emerged during the late 1980s to overcome fundamental limitations with all-silicon implanted base transistors [15]. Interestingly, germanium (Ge) was the preferred semiconductor material during the 1950s, but the narrow bandgap (0.66 eV) causes high junction leakage current, especially at raised junction temperature. Furthermore, passivation layers on the semiconductor surface, necessary for fabrication of integrated circuits, does not readily exists with germanium. GeO$_2$, germanium oxide, is difficult to form, is water soluble, and dissociates at 800 °C. What is more, the resistivity of undoped germanium is only 47 Ω cm, which precludes fabrication of devices with high breakdown voltage, like rectifiers, high-voltage transistors and thyristors.

Instead, another material, better suited for fabrication of diodes and transistors, was required. Silicon (Si), with its 1.1 eV bandgap and thermally grown or deposited oxide (SiO$_2$) which is easy to form, and thermally and chemically very stable, replaced germanium for microelectronic components. Silicon devices can be operated up to 150 °C junction temperature without large leakage currents and with long-term reliability.

Already in 1951, Shockley suggested to use an abrupt heterojunction as efficient emitter-base junction in a bipolar transistor [16] and Kroemer analyzed in the same year a graded heterojunction to be used as a wide-bandgap emitter [17]. Bandgap engineering is the technique to utilize the bandgap properties of different semiconductor materials to achieve superior device characteristics. The most common use of bandgap-engineered devices is for high-speed devices. By using two different materials for e.g. base and collector, high gain be obtained with thin layers while still keeping the doping in the layers high which minimizes parasitics. Not only can different materials be combined, additional advantages are achieved by grading the materials over the device regions, which will create a built-in field that will further increase the speed (see section 3.1).

The first SiGe HBT was demonstrated in 1987 [18] after significant advances in epitaxial growth of SiGe layers, and it took off rather quickly with the first commercially devices available in 1996 [19]. A mature first generation device had typically an $f_T$ of 35-50 GHz, which is about the limit for a bipolar device with implanted base with comparable performance and intrinsic parasitics. A second-generation device, first demonstrated in 1990 [20] had typically an $f_T$ of 75 GHz, and soon 100+ GHz was achieved [21]. In 1992, the first SiGe-BiCMOS integration was demonstrated [22] at the 0.5 µm process technology node. Today, all large RF-oriented semiconductor manufacturers use SiGe-BiCMOS for the more demanding high-frequency applications.
Figure 6. 100+ GHz SiGe transistor: (a) Non-self-aligned transistor structure, (b) SIMS profile of 0-25 % SiGe HBT with an in-situ doped polysilicon emitter. [21] © 1993 IEEE.

Figure 6 shows the first 100+ GHz device structure (which also includes such features as shallow trench isolation, n+ subcollector, and pedestal collector). A 0-25 % ramp of Ge together with an in-situ doped emitter was used for the emitter/base/collector.

IBM has dominated the development of commercial SiGe devices and the internal development history at IBM is well documented [23-24].

3. Present

In this section, only a few important features of the present technology of the SiGe bipolar transistor will be discussed. For extensive treatments of the physics of bipolar and SiGe bipolar devices, see the textbooks by Ashburn [25-26] and Roulston [27]. Furthermore, in [28], a good overview of bipolar device design and various device structures is given, which combined with another chapter in the same book [29] gives a good view of the basics of the SiGe bipolar transistor.

3.1. SiGe bandgap engineering

The SiGe bipolar transistor can be seen as a variant of the modern integrated Si bipolar transistor extended with bandgap engineering, which results in a new degree of freedom in device design. The SiGe-base transistor offers improved device characteristics for analog and high-frequency applications. The main improvements are the higher current gain, shorter base transit time, higher Early voltage, and lower 1/f noise.

Si and Ge are located in the same group in the periodic system and can be combined to form a stable alloy. The lattice constant however is about 4 % smaller in Si and SiGe alloys can only be used in thin layers with a small fraction of Ge if the layer will be used on a Si substrate. If a thin enough SiGe layer is grown on a Si substrate, the lattice will adopt the
structure of the underlying crystal and resulting in a SiGe layer containing compressive strain. Since the layer must be thermodynamically stable to withstand high-temperature processing during the device fabrication, typically 100-200 nm layer thickness with no more than 20-30 % of Ge can be fabricated [30], otherwise defects in the layers very quickly appear, causing device leakage and degraded performance.

The first SiGe layers were grown by molecular beam epitaxy (MBE) [18], but this method has a too low throughput for commercial production. The main growth techniques for thin SiGe layers that are used in industrial manufacturing are different variants of chemical vapor deposition (CVD). For commercial production, low-pressure UHV/CVD [31] has mainly been used by IBM, while many other companies use CVD-processes at somewhat higher pressure, such as RPCVD. Ge can also be added by ion implantation [32].

The commonly used profile, with the SiGe base layer sandwiched between a Si emitter and Si collector with a linearly graded Ge profile as shown in Figure 7, has been widely adopted. The bandgap of the base layer narrows gradually toward the base-collector junction. The electric field created from this bandgap reduction due to the increasing Ge concentration is a key factor for the speed increase. The collector current, which depends on the base properties, will be higher for a SiGe transistor, while the base current, which depends on the emitter properties, will be the same as for the Si transistor. Thus the current gain is increased due to the lower bandgap of the emitter-base junction. Furthermore, a quasi-electric field will delay base depletion when the supply voltage ($V_{CB}$) is increased, thus the Early voltage will be improved.

**Figure 7.** Schematics showing the energy bands and Ge profile of a typical SiGe-base transistor together with the energy bands of Si-base transistor.
Assuming uniform doping profiles and a linearly graded Ge profile across the base, the gain enhancement for such a device can be written as [33]:

\[
\beta_{\text{SiGe}} = \frac{(N_e N_i D_{ab})_{\text{SiGe}} \Delta E_{G(\text{grade})}}{(N_e N_i D_{ab})_{\text{Si}} kT} \exp(-\frac{\Delta E_{G(0)}}{kT}) \frac{1}{1 - \exp(-\frac{\Delta E_{G(\text{grade})}}{kT})}
\]

(1)

where \( \Delta E_{G(0)} \) is the Ge-induced bandgap narrowing at the emitter end of the base, \( \Delta E_{G(\text{grade})} = \Delta E_{G(WB)} - \Delta E_{G(0)} \) is the grading over the base, \( \Delta E_{G(WB)} \) is the Ge-induced bandgap narrowing at the collector end of the base and \( D_{nb} \) is the average diffusivity of electrons in the graded SiGe base.

The ratio of the base transit time is given by:

\[
\frac{\tau_{B,\text{SiGe}}}{\tau_{B,\text{Si}}} = \frac{2kT}{\Delta E_{G(\text{grade})} (D_{ab})_{\text{SiGe}}} \left[1 - \frac{kT}{\Delta E_{G(\text{grade})}} \exp(-\frac{\Delta E_{G(\text{grade})}}{kT}) \right]
\]

(2)

which is less than unity. The term \( D_{ab}/(D_{ab})_{\text{SiGe}} (> 1) \) accounts for the strain-enhancement of the minority electron mobility with increasing Ge content. Since \( \tau_B \) is often limiting the \( f_T \) in bipolar transistors, the transit time is decreased with SiGe in the base. The Ge-grading across the base will also influence the emitter delay in a similar way, which will further increase \( f_T \).

The enhancement in the Early voltage \( (V_A) \) is given by:

\[
\frac{V_{A,\text{SiGe}}}{V_{A,\text{Si}}} \approx \exp\left(\frac{\Delta E_{G(\text{grade})}}{kT}\right) \left\{1 - \exp\left(-\frac{\Delta E_{G(\text{grade})}}{kT}\right) \right\} \frac{\Delta E_{G(\text{grade})}}{kT}
\]

(3)

which results in an improved output conductance.

By adding substitutional carbon into Si and SiGe, the boron outdiffusion from the base layer can be minimized [34], and very sharp base profiles can be used which leads to increased high-frequency properties [35].

Ge diffusivity is lower than the B diffusivity [36] and the Ge profile can therefore be tuned precisely to further enhance the performance by creating fields that accelerate the carriers in the base and reduce the transit time even further.
3.2. The SiGe HBT transistor

The early device structures were of *mesa* type, where the entire emitter/base/collector profile was grown by epitaxy (often *MBE*), followed by a series of etching steps to isolate and contact the base and collector regions [18, 37]. Although the device results by this approach have been impressive [38], the structure is less optimal for IC processes, especially BiCMOS integration.

Modern device structures can be divided into Non-Self-Aligned (NSA) and Self-Aligned (SA) device structures [39]. The SA structure has lower parasitics, therefore higher $f_{\text{max}}$, which is a better figure-of-merit for circuits such as emitter-coupled logic and power amplifiers (see section 3.5) than $f_T$, and is therefore the most common structure for high-performance structures.

A self-aligned high-performance SiGe bipolar technology, developed by Infineon Technologies [40-41], will now be used to illustrate state-of-the-art SiGe bipolar technology as of 2005. The highest performing transistors have $f_T$ of 225 GHz, $f_{\text{max}}$ of 300 GHz and a ring oscillator gate delay time of 3.3 ps. Using this technology, a static frequency divider was operating up to 102 GHz input frequency [42]. The results represent state-of-the-art SiGe bipolar technology, yet the process has a demonstrated maturity for production with a low spread of device parameters. Millimeter-wave applications up to 100 GHz, such as 77 GHz automotive radar systems and 60 GHz WLAN, which only could be realized in III-V technologies in the past, are now feasible in silicon technology.

The transistors have a double-polysilicon self-aligned (SA) emitter-base configuration with a SiGe:C base, integrated by selective epitaxial growth (SEG), which is very effective for ultra-shallow base formation with low parasitics. Figure 8 shows the main steps in the fabrication of the emitter/base and in Figure 9 a TEM cross-section of a fabricated device is shown.

Device isolation consists of shallow and deep trench isolation. 0.3 μm lithography is used together with 80 nm TEOS spacers to obtain a minimum effective emitter width of 0.14 μm, Figure 8 (d). The pedestal implant is made after the nitride spacer formation, Figure 8 (a), thus reducing the implanted area and the associated base-collector capacitance. Three types of NPN transistors with different collector dopings are used, as well as two types of poly resistors, a TaN resistor, a MIM capacitor, and four layers of Cu metallization.

Figure 10 (a) shows the emitter-base doping profile used for this process, and in Figure 10 (b) the data extraction to determine the $f_T$ and $f_{\text{max}}$ values is shown.
Figure 8. Fabrication of the emitter-base complex for a double-polysilicon self-aligned emitter-base structure with SEG SiGe:C base. a) nitride spacer formation, pedestal collector implantation, b) self-aligned formation of $p^+$-polysilicon overhangs, c) selective SiGe base deposition, nitride spacer removal, d) formation of emitter-base spacer and emitter processing [40] © 2003 IEEE.

Figure 9. TEM cross-section of the emitter-base complex of a 225 GHz $f_T$ transistor [43].
In the neutral base a boron spike with a concentration of $5 \times 10^{19}$ cm$^{-3}$ is grown to enable a very low base sheet resistance. The boron peak is surrounded by carbon to suppress undesirable boron diffusion. The thermal budget for the emitter drive-in is very limited; this leads to a very small emitter depth and very little boron diffusion of the base. The Ge content is linearly graded across the base and the maximum Ge fraction is 25%. The purpose of the 5% Ge-plateau is to make the collector current insensitive to inhomogeneities of the emitter drive-in. The transistors have a monocrystalline emitter contact which leads to a smaller emitter resistance compared to a polycrystalline contact. This contact is also more reproducible due to the missing interface oxide which is difficult to control, but the drawback is lower beta. In comparison to transistors with a polycrystalline contact the standard deviation of the base current over the wafer could be reduced by more than a factor of 2. Also the matching constants for the base current and current gain could be cut by more than half.

The demonstrated best-combined performance reported by mid-2005 is by IBM [44], which has developed a SiGe HBT technology with $f_T=300$ GHz, $f_{max}=350$ GHz and gate delay below 3.3 ps. This is the highest reported speed for any Si-based transistor with a combined performance of $f_T$ and $f_{max}$. The device has a very thick extended “raised” base connection, see Figure 11, reducing parasitics important to reach a high $f_{max}$.

3.3. SiGe BiCMOS

There are many benefits of combining MOS and bipolar transistors on a single chip. CMOS has for a long time been the preferred choice for digital design. It has low power consumption, small design features, and
small overall area for a given building block and it is easy to design in. However, a bipolar device usually outperforms CMOS on parameters such as current drive ability and high-frequency performance. The BiCMOS has therefore become a popular solution for high-performance mixed-signal or high-frequency circuits when the need of integration level is high. One such important class of circuitry is GSM transceivers. The radio blocks are realized mostly in bipolar and the CMOS is used for digital circuitry such as digital interfaces, signal processing, and etc.

The SiGe bipolar transistor can with some additional effort replace the Si bipolar transistor in BiCMOS integration schemes. The first demonstration was made in 1992 [22] at the 0.5 µm process technology node. The integration requires some additional attention compared to an implanted base process because of the limited thermal budget the deposited SiGe is allowed to experience. The need of a high drive-in temperature cycle for the (n-type) S/D often requires that the FET is done before the HBT [45] in high-performance processes. Figure 12 shows a process integration sequence for a 0.2 µm SiGe BiCMOS with a 0.3 µm CMOS on SOI substrates with shallow and deep trench isolation [46]. The CMOS module was inserted after the formation of the trench isolation.

3.4. The SiGe vertical PNP

To further extend the BiCMOS, the addition of a PNP bipolar transistor, to obtain complementary bipolar devices, is useful for the design of linear analog circuits with symmetrical architectures [47]. In bipolar IC and BiCMOS processes, substrate and lateral PNP transistors can easily be
created without using additional process steps but the performance of most device parameters for such devices is low. For highest performance, the simultaneous use of vertical device structures for both NPN and PNP is mandatory. However, the different dopant properties make processing and process integration somewhat complex.

By adding SiGe to the PNP structure, similar advantages as for the NPN are earned. Since the base of the PNP device consists of an n-doped layer, the base sheet resistance will be about half compared to an NPN device because of the higher mobility of the majority carriers. The SiGe is therefore mainly used to compensate for the lower gain and transit time.

The first SiGe PNP device was demonstrated in 1988 [48] in the same time frame as the first NPN SiGe base transistors [23]. The base was formed using MBE with antimony as dopant, and a single-crystal boron-doped emitter was deposited using UHV/CVD with a minimum heat cycle to minimize boron outdiffusion.

A recent implementation is found in [49] where deep trench isolation on an SOI substrate is used to create a commercial 5V complementary SiGe BiCMOS technology for ultra-high speed precision analog applications, see Figure 13.

3.5. Real-world devices: The SiGe power amplifier

The space in this review article is too limited to discuss all possible SiGe discrete and integrated circuits applications and how device design parameters are best selected for maximum performance. Instead, I will concentrate on SiGe transistors for high-frequency power amplifiers (PAs).
Figure 13. 5V Complementary-SiGe BiCMOS with vertical NPN and PNP [49] © 2003 IEEE.

PAs are core components in the high-growth wireless communications industry. Depending on supply voltage, which usually translates to output power levels, devices for high-frequency PAs are divided into two classes. High-voltage includes 12 V, 18 V, 25-28 V supply voltages while Low-voltage is defined as around 6 V supply and lower. Since many device parameters scale with doping levels, which are used to set breakdown voltages for operation at a certain supply voltage, a device optimized for high voltage will not perform well at a lower voltage without re-optimization. High-voltage devices are used in the stationary systems, with maximum output level in excess of 100 W. For a large device, high supply voltage is advantageous since the power per device area can increase almost linearly with the supply voltage, while the impedance stays almost constant. The low-voltage devices are typically used in portable systems where power supply comes from one or several battery cells. The output level is typically in the 0.1-5 W range. As a general observation, systems that operate at higher frequency bands use lower output power, but as the power added efficiency (PAE) for PAs drops at higher frequency, the power consumption is not scaled accordingly.

In conventional high-speed SiGe HBTs a low base doping concentration in conjunction with a low Ge content of a trapezoid shape is commonly used to obtain a high current gain. Alternately, the SiGe induced bandgap narrowing permits a high base doping concentration with more Ge content to be employed in the base region for a given current gain. The
high-speed characteristics of high-performance SiGe HBTs (such as the devices discussed in section 3.2) are generally obtained with the sacrifice of breakdown voltages and for such devices, the emitter-to-collector transit time $\tau_{EC}$ is mainly limited by the emitter transit time $\tau_E$ [50]. For PAs, high breakdown voltage SiGe HBTs is needed, and the collector space charge layer delay $\tau_{CSCL}$ instead is the dominant time delay component. Parameters such as Ge content and Ge profile in the base will have less impact on $f_T$ for such devices, but the reduced base resistance can significantly enhance $f_{max}$, which results in high power gain, advantageous for PAs.

In a famous paper published in 1965, E. O. Johnson of RCA pointed out the fundamental limits for high-voltage, high-speed devices [51]. In its basic form, the product of the cutoff frequency, $f_T$, and the maximum allowable applied voltage $V_m$ (breakdown voltage, usually $BV_{ceo}$) is constant:

$$V_m f_T = \frac{E v_s}{2 \pi}$$  \hspace{1cm} (4)

where $E$ is the breakdown field strength and $v_s$ the material-dependent, saturation drift velocity, which takes the value of $\sim 2 \times 10^{11}$ cm/s for silicon.

While the Johnson limit is a valuable concept, real power transistors are instead often operated in a mode equivalent to base-emitter connection through a resistor, with the breakdown value $BV_{cer}$, which for reasonable resistance values approaches the much higher $BV_{cbo}$. Instead of $f_T$, the maximum oscillation frequency, $f_{max}$, or power-gain cutoff frequency, defined as

$$f_{max} = \sqrt{\frac{f_T}{8 \pi R_b C_{bc}}}$$  \hspace{1cm} (5)

where $R_b$ is the base resistance, and $C_{bc}$ is the collector-to-base capacitance, is more useful as a figure of merit for the amplifier’s high-frequency performance. The RF power, efficiency and gain deteriorate as the operation frequency increases. The gain is directly related to $f_{max}$ since at the frequency $f$, the power gain $G_p$ is given by [52]:

$$G_p \approx \frac{G_0}{\sqrt{1 + G_0^2 \left(\frac{f}{f_{max}}\right)^4}}$$  \hspace{1cm} (6)

where $G_0$ is the gain at low frequency ($h_{FE}$). Since for a SiGe transistor, the $f_{max}$ is less tightly connected to the Johnson limit than $f_T$ ($R_b$ can be lowered
by higher base doping without affecting the BV), there is a larger degree of freedom regarding device optimization compared to other device applications [53].

For high-voltage (high output power) devices for applications such as cellular base stations or cable TV transmitters, the bipolar transistor had been the workhorse until around ten years ago. The market is now totally dominated by the LDMOS, the main advantages being higher gain, linearity and thermal stability, and easier integration. However, bipolar power transistors do offer better performance with regards to output power per area, linearity in some cases and long-term reliability, and therefore studies of SiGe transistors for high output power applications have been published. A high $BV_{CBO}$ is generally desired which leads to a much thicker collector than for small-signal transistors. Hobart et al. demonstrated in 1995 SiGe transistors with $f_T$ and $f_{max}$ values of 10 and 22 GHz, respectively, for transistors with $BV_{CBO}$ of 40 V [54]. The transistors structures were grown by MBE and fabricated in a double-mesa process. The results were later used to fabricate a 230 W device for pulsed radar at 2.8 GHz [55].

For cellular base station applications using typically 26 V supply, which requires the $BV_{CBO}$ to be >55 V since the device will experience $\sim 2*V_{cc}$ in the usual PA configuration, this author has demonstrated 20 W output power at 2 GHz [56] using the device structure shown in Figure 14. Large size devices were processed using an existing poly-Si emitter RF power BJT technology with Au metallization and some necessary modified steps for the SiGe implementation. The base layers with designed Ge and B profiles were deposited either by MBE or CVD.

Figure 14. Device structure for 2 GHz cellular base stations PA, (a) Schematic cross-sectional drawing, showing part of repeated e–b structure (left) and device isolation (right), (b) SEM micrograph of the processed SiGe RF power HBT, showing repeated e–b structure only [56]. Reprinted with permission from Elsevier.
The devices showed very high $BV_{CBO} (>80 \text{ V})$ with very low leakage currents. The current gain was very stable over a wide collector current range and weakly influenced by the environmental temperature. At 2 GHz, the CW output power of 20 W (at 25 V supply) was obtained with an efficiency of 68 % in class AB operation. The long-term temperature stability was excellent. SiGe transistors could be operated at full output power for an extended time without any external temperature bias compensation, which is virtually impossible with conventional Si-BJTs.

A well-known problem with Si-BJTs is the thermal instability caused by the increased beta with temperature. For large-area multi-finger transistors, emitter ballasting \cite{57} is used to equal and limit the current between fingers, but the ballasting will degrade output power, power gain, and PAE. With SiGe transistors, it is possible to design the SiGe profile such that the beta will have a negative temperature coefficient \cite{33}, and an emitter-ballasting-resistor-free PA at 15 V supply, 900 MHz, 5 W output power has been demonstrated \cite{58}.

The temperature dependence of the $h_{FE}$ value for devices with three different designed base profiles was therefore also studied for the previously discussed device (25 V, 2 GHz, 20 W). The results are plotted in Figure 15. For the reference Si-BJT, there is a strong increase of $h_{FE}$ when increasing operation temperature. In contrast, the SiGe transistors with

![Figure 15](image.png)

**Figure 15.** Current gain values of the SiGe power transistor, normalized to one, measured at room temperature, are plotted as a function of the operation temperature. For a comparison, data for a conventional Si RF power transistor are also shown \cite{56}. Reprinted with permission from Elsevier.
different profile selections showed improved thermal properties. The transistor with a box-like Ge profile (“profile #3”) demonstrated a negative temperature coefficient, i.e. $h_{FE}$ decreased as the temperature was increased. This is an expected result, since the valence band offset at the e–b junction interface contributes to the gain with a factor that decreases with increasing temperature.

For low-voltage PAs, Si BJT, SiGe HBT, and GaAs (and related materials) HBT technologies are competing as the technology of choice. Low-voltage SiGe PA performance was demonstrated in 1995 [59] with a PAE of 33 % at 5.7 GHz although the output power was only in the 0.1 W range. The transistor structure was fabricated using a double mesa process with self-aligned base contacts and the base/emitter layers were deposited by MBE. In 1996, operation in the 0.9-1.9 GHz bands was investigated [60] using an improved mesa process [61]. 1 W output power with 44 % PAE in class A operation at 1.9 GHz was reached.

BiCMOS-compatible PAs, especially for GSM, have recently been extensively studied [62]. The driving force is the higher integration level, lower cost, and better reliability compared to today’s GaAs HBTs. The SiGe device design is mainly constrained by the ruggedness requirements (ability to withstand high-voltage excursions, quantified by voltage standing-wave ratio (VSWR), which is primarily addressed by collector engineering [63], without compromising PA performance.

As previously discussed, SiGe HBTs suitable for wireless PA applications do not fully benefit from high-performance HBT scaling trends because reduced intrinsic base width does not significantly improve $f_t$ unless collector thickness also is scaled. This results in an optimization conflict when very high-performance devices are to be used in the same technology as an optimal device for PA applications for which ruggedness is a critical consideration, since the two device types demand the opposite trends in collector design. However, technology scaling does benefit SiGe PAs through improved lithographic tolerances and optimized intrinsic and extrinsic base processes, both of which increase $f_{max}$ [62]. The scaling trend is shown in Figure 16, which plots the $f_t^*BV_{CEO}$ and $f_{max}^*BV_{CEO}$ products for high breakdown SiGe HBTs, from both high-performance and low-complexity process approaches and from lithographic generations from 0.5 to 0.13 µm. The increased $f_{max}^*BV_{CEO}$ product can be translated to increased gain at a fixed frequency or increased frequency of operation, opening possibilities of operation at new, higher band.

All PAs discussed so far have been for the popular communication systems in the 1-2 GHz bands. A 300 mW SiGe PA operating at 18 GHz has recently been demonstrated [64]. The $BV_{CBO}$ was 26 V and the device operated at 7 V supply voltage.
Figure 16. Observed speed–ruggedness tradeoff for four lithographic generations of high-breakdown SiGe HBTs, from both production and experimental technologies. These results suggest that technology scaling will enhance PA device performance primarily through improvements in $f_{\text{max}}$, although improvements in the $f_T^*BV_{CEO}$ metric are observed at the 0.13 $\mu$m generation [62] © 2004 IEEE.

4. Future

4.1. Future trends

Of course it will be risky in a review paper to “review” future trends, but as the SiGe bipolar transistor is by 2005 a rather mature device, first demonstrated in 1988, commercialized in 1996, and resting on the shoulders of the silicon technology with its additional decades of development, what can go wrong with a summary of trends and some speculations about the future...?

The performance of the SiGe bipolar transistor has continuously increased over the years. Important improvements are the formation of very shallow bases and self-aligned device concepts to reduce the parasitics and scale the horizontal dimensions, e.g. the emitter size. The SiGe bipolar transistor can now compete with the most advanced III–V-based devices. Figure 17 shows some recent published $f_T$ values for different device types. In fact, Si-based bipolar transistors have begun to outperform their III–V counterpart heterojunction bipolar transistors (HBTs) in terms of $f_T$.

The SiGe bipolar transistor has now in most high-performance applications replaced the implanted base Si BJT transistor. The reason why they are used in such a wide range of applications is simply because they are superior or equal to conventional BJT in every important performance metric. Since the cost penalty is in the order of 10 % when a simple SiGe
module is added to an existing BiCMOS process with implanted base, there is no reason not to. Even the marketing department may be willing to support this cost, since not having a SiGe process on the roadmap today may be much more costly! We can therefore assume that SiGe will be employed in even more application areas in the future.

The SiGe continues to improve its performance [66]. The principal advantage of Si devices comes from its extremely aggressive scaling and extensively optimized structure, which are enabled by the state-of-the-art Si technology (see section 3.2). While the CMOS is driving the technology scaling according to Moore’s law, the (CMOS-compatible) SiGe transistor benefits from the process module development, the large-scale integration, and the high yield, leading to a strong cost efficiency for all Si-based technology [67-68]. Although CMOS is maturing for RF/analog/mixed-signal applications, the main advantages being integration (system-on-chip) and cost, SiGe HBT is still preferred for high-performance applications because of its advantages in transconductance, 1/f noise, device matching, and power performance. The solution is the BiCMOS process, which is available from major semiconductor companies; two examples are given in Refs. [69, 70]. To extend the usability even further, LDMOS has been integrated into SiGe:C-BiCMOS [71], and thin-body SOI CMOS and high-performance SiGe HBTs have been integrated in a common process flow [72]. Soon, we will see sub-100 nm strained MOS and SiGe HBT in a common process flow as well.

Finally, there are application domains where SiGe bipolar transistors have not yet been used [65]. These are usually found in really high-frequency bands where III-V devices still dominate and where cost limits

Figure 17. Trend of the best-achieved $f_T$ for various transistors types [65]. (The numbers indicate the corresponding references in [65].) © 2004 IEEE.
The usability of the applications, or in new frequency bands where the reduced cost because of the availability of Si technology will make new applications much more interesting from a commercial point of view, such as 77 GHz automotive radar systems [73].

4.2. Summary and conclusions

In this chapter, the history, present state, and future of the SiGe bipolar transistor have been reviewed. From a historical perspective, we have seen how the bipolar transistor was developed and refined, and how it transformed into the SiGe bipolar transistor that has been dominating the commercial research and process development for more than a decade now.

As always with semiconductor technology in general, and with silicon technology in particular, the question is how much the technology can be further developed, performance further increased, and the relative costs further reduced.

A study of nano-scale SiGe HBT design for beyond THz \( f_T \) has already been published [74], and with the words of Rieh [65]: “With no visible limitation for further enhancement of device speed at hand, the march toward terahertz band with Si-based technology will continue for the foreseeable future”, I will end this review. The sky is the limit!

References


43. J. Böck, personal communication.


