From The Lab to The Fab: Transistors to Integrated Circuits

Howard R. Huff
International SEMATECH
2706 Montopolis Drive
Austin, TX 78741

Abstract. Transistor action was experimentally observed by John Bardeen and Walter Brattain in n-type polycrystalline germanium on December 16, 1947 (and subsequently polycrystalline silicon) as a result of the judicious placement of gold-plated probe tips in nearby single crystal grains of the polycrystalline material (i.e., the point-contact semiconductor amplifier, often referred to as the point-contact transistor).The device configuration exploited the inversion layer as the channel through which most of the emitted (minority) carriers were transported from the emitter to the collector. The point-contact transistor was manufactured for ten years starting in 1951 by the Western Electric Division of AT&T. The *a priori* tuning of the point-contact transistor parameters, however, was not simple inasmuch as the device was dependent on the detailed surface structure and, therefore, very sensitive to humidity and temperature as well as exhibiting high noise levels. Accordingly, the devices differed significantly in their characteristics and electrical instabilities leading to "burnout" were not uncommon. With the implementation of crystalline semiconductor materials in the early 1950s, however, p-n junction (bulk) transistors began replacing the point-contact transistor, silicon began replacing germanium and the transfer of transistor technology from the lab to the fab accelerated. We shall review the historical route by which single crystalline materials were developed and the accompanying methodologies of transistor fabrication, leading to the onset of the Integrated Circuit (IC) era. Finally, highlights of the early years of the IC era will be reviewed from the 256 bit through the 4M DRAM. Elements of IC scaling and the role of Moore's Law in setting the parameters by which the IC industry's growth was monitored will be discussed.

INTRODUCTION

Transistor action was experimentally observed by John Bardeen and Walter Brattain in n-type polycrystalline germanium on December 16, 1947 (and subsequently polycrystalline silicon) as a result of the judicious placement of gold-plated probe tips in nearby single crystal grains of the polycrystalline material (i.e., the point-contact semiconductor amplifier, often referred to as the point-contact transistor) [1-3]. The device configuration exploited the inversion layer as the channel through which most of the emitted (minority) carriers were presumed to be transported from the emitter to the collector. The point-contact transistor was manufactured for ten years starting in 1951 by the Western Electric Division of AT&T [4]. The *a priori* tuning of the point-contact transistor parameters, however, was not simple inasmuch as the device was dependent on the detailed surface structure and, therefore, very sensitive to humidity and temperature as well as exhibiting high noise levels. Accordingly, the devices differed significantly in their characteristics and electrical instabilities leading to "burnout" were not uncommon [5]. With the implementation of single crystalline semiconductor materials in the early 1950s [3,6-8], however, p-n junction (bulk) transistors began replacing the point-contact transistor, silicon began replacing germanium [5,7,8] and the transfer of transistor technology from the lab to the fab accelerated.

We shall briefly review the historical route by which single crystalline materials were developed and the accompanying methodologies of bipolar transistor fabrication (i.e., grown junction, alloy and diffused). The oxide masking and photolithographic technique of Carl Frosch and Link Derick [9,10] and its embodiment in the mesa process, the utilization of the silicon oxide for the passivation of the silicon surface by Mohammed (John) Atalla and colleagues [11] and the development of the planar silicon transistor by Jean Hoerni (i.e., the planar process) [12-15] whereby the SiO_2 masking layer, utilized in the fabrication of diffused silicon transistors, was left in place for the passivation of p-n junctions intersecting the wafer surface set the stage for MOSFET fabrication as well as the utilization of the dielectric layer for supporting metallic conductor overlayers in the integrated circuit (IC) era [16].

The Si-SiO_2 diffusion technology, transferred from AT&T's Bell Telephone Laboratories (BTL) to Shockley Semiconductor and, therefore, to Fairchild Semiconductor Corporation led to the phenomenon of "Silicon Valley" and the creation of the IC industry.

This manuscript was published previously in *ULSI Process Integration III*, ECS PV 2003-06, 15-67 (2003).

CP683, *Characterization and Metrology for ULSI Technology: 2003 International Conference*,
edited by D. G. Seiler, A. C. Diebold, T. J. Shaffner, R. McDonald, S. Zollner, R. P. Khosla, and E. M. Secula
© 2003 American Institute of Physics 0-7354-0152-7/03/$20.00

Indeed, Gordon Moore has noted "... and you are once again reminded that this is no longer just an industry, but an economic and cultural phenomenon, a crucial force at the heart of the modern world" [17]. The critical role of John Moll's laboratory at BTL in 1954 and the development of the oxidation, diffusion, lithography, aluminum metallization and thermocompression bonding techniques for the fabrication of the junction transistors and silicon-controlled rectifier [18-21], in conjunction with Nick Holonyak [22], are reviewed.

The oxidation kinetics of silicon by Bruce Deal and Andy Grove [23], the explication of the charge and drift mechanisms in the Si-SiO_2 system by Deal et al. [24-30] and the role of Pieter Balk [31,32] in emphasizing the importance of subsequent hydrogen and nitrogen annealing are briefly discussed. The mesa and planar processes described above paved the way for the invention of the IC by Jack Kilby in 1958 [33-36] (utilizing the mesa methodology in germanium) and Bob Noyce in 1959 [37-39] (utilizing the planar procedure, i.e., in silicon) and the subsequent microprocessor era [40-42]. The critical differences between the two patents (i.e., the interconnection methodology) are clarified by Walter Runyan and Kenneth Bean [43].

The early years of the IC from the 256 bit to the 4 M DRAM are then reviewed [44], building on Bob Dennard's one transistor cell structure [45] and associated scaling methodology [46-49]. Gordon Moore's remarkably prescient assessment that the number of memory bits would double per year (now taken as about 18 months), enshrined as Moore's law, became the productivity criterion by which the IC industry grew at about a 25% compound annual growth rate [50-54] as illustrated in the International Technology Roadmap for Semiconductors (ITRS) [55]. More than just monitoring productivity, whether by staying on the productivity curve or increasing manufacturing effectiveness, however, is required. Rather, modeling productivity—the identification of new productivity measures—is now required [56].

Finally, potential directions for enhanced IC performance, per the ITRS [55], are briefly discussed. These include both carrier transport mechanisms in the channel using variously strained structures to enhance the carrier mobility and new MOSFET device configurations, including various vertical transistor configurations [57].

Single-Crystal Growth

Polycrystalline germanium and silicon were the basic materials used at BTL and elsewhere for transistor research and development in the late 1940's inasmuch as the utilization of single crystals of germanium and silicon for the transistor was a very controversial matter at that time, although the importance of high-purity material to achieve a high rectification characteristic was understood. Gordon Teal has noted [58]:

> "Bill Shockley was opposed to the work on germanium single crystals when I suggested it, because, as he has publicly stated on several important occasions, he thought that transistor science could be elicited from small specimens of polycrystalline masses of material."

Teal, however, was a proponent of the criticality of single-crystal materials for the electronics era, recognizing that Shockley's bipolar junction transistor characteristics [59-62] in single-crystal germanium [63-66] and silicon [67,68] would be substantially better and more reproducible than those of polycrystalline material [6-8,58]. The limited capability for the fabrication of single crystal materials further exacerbated the situation. Indeed, Shockley later realized the shortcomings in his previous assessment of the usefulness and necessity of single crystals [69,70].

Teal believed the fundamental property of a crystalline semiconductor, which would result in its technological importance, was the easily controllable and spatially variable concentration, type and mobility of free carriers, which was indeed found to be the case [71-78]. According to Teal [6-8,58,79-81]:

> "I reasoned that polycrystalline germanium, with its variations in resistivity and its randomly occurring grain boundaries, twins and crystal defects that acted as uncontrolled resistances, electron or hole emitters and traps would affect transistor operation in uncontrolled ways. Additionally, it seemed to me that use of this material to produce many complex units meant to be identical, with close performance tolerances, would be inconsistent with high yields and, therefore, also with low costs. Even in developing complex transistor devices, it seemed to me essential to have a high-perfection, high-purity controlled composition semiconductor in order to achieve a separation of various available electron and hole conduction processes in order to analyze and understand the operation of these devices and thus to finally achieve an optimum functional use of them.
>
> My general aims for the single crystal research were as follows: (1) to produce a conducting medium in which a high degree of lattice perfection, of uniformity of structure and of chemical purity is attained; and (2) to build

into this highly perfect medium in a controlled way the required resistivities and electrical boundaries to give a variety of device possibilities by control of the chemical composition (i.e., donor and acceptor concentration) along the direction of single crystal growth."

The successful initial results obtained in a joint program between Teal and John Little, begun in September 1948, resulted in several germanium rods with some large single crystals in them by pulling from a melt [63,79-81]. The technique employed a single-crystal seed (oriented in a <111> or <100> direction), seed rotation and precise temperature control of the melt-solid interface from which the crystal was pulled [6,82-84] (see Figure 1) [84]. These were vital factors in the attainment of single crystals in which the essential semiconducting properties became highly controlled. The method is variously referred to as pulling, the Teal-Little process [63,82] or, somewhat inappropriately, as the Czochralski (CZ) process. Generally, it is simply called the CZ technique, after Czochralski who in 1918 withdrew thin single-crystal metal filaments from a melt [85]. Czochralski did not use a single-crystal seed, however, and apparently did not recognize the significance of directly controlling the melt temperature to control the crystal diameter. Furthermore, it was noted by Teal in 1952 that in his zeal to develop a single crystal transistor in 1950 [79,80,86]:

"A variety of methods has been employed by various experimentalists to produce single crystals of metals, salts, insulators, and semiconductors. It will be apparent from the scientific literature and the description that follows that the pulling method of growing single crystals of germanium differs materially from the pulling methods of Czochralski, Kyropoulos, Gomberz, Hoyem and Tyndall, and others. There are differences in the materials, designs, and operation of the crystal batch growing equipment. Also, novel techniques were developed to produce germanium single crystals in which the impurity composition and lattice perfection are controlled throughout the crystal. The use of a pulling method for germanium and the employment of new techniques to be described have been vital factors in the attainment of single-crystal in which the essential semiconducting properties are highly controlled."

In retrospect, it should be noted that during the 1940's the concept, let alone the necessity and usefulness of single crystal materials, was not

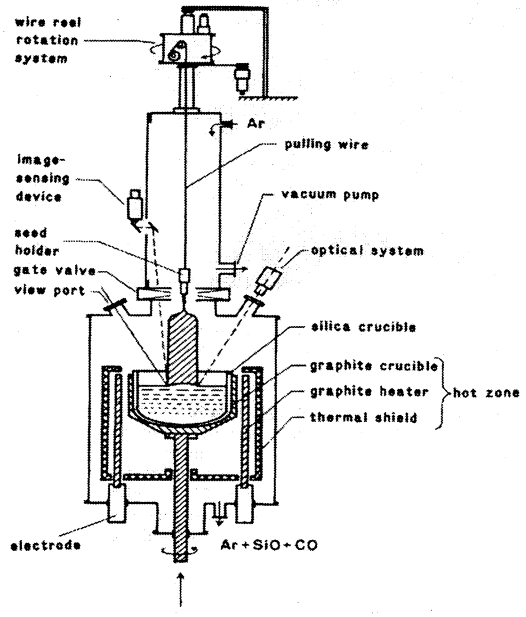

Figure 1. Schematic illustration of a typical Czochralski puller with hot zone, automatic optical and image sensing diameter controls and wire reeling system [84]. Reprinted with permission from Elsevier Science.

appreciated for semiconductor applications. Teal's emphasis on the preparation and characterization of single crystal material, however, facilitated experimental verification of a number of quantum theoretical concepts developed for electrons and holes in crystalline semiconductors such as effective mass, drift and conductivity mobility, carrier lifetime and tunneling [6,58] and clarification of a number of phenomena in p-n junctions [87] and, indeed, exhibited significantly improved characteristics compared to polycrystalline samples. For example, Haynes observed in early 1949 that the lifetime of minority carriers in single crystals of germanium was as much as 140 μs (20 to 300 times greater than observed in polycrystalline germanium) and mobilities three to four times higher, due to the greater perfection and purity (45 ohm-cm) of the single crystals [6,58,88-90]. Teal also reported injected carrier lifetimes greater than 200 μs in single crystal germanium as compared to significantly lower carrier lifetimes of 1-5 μs in polycrystalline germanium [58]. By the early 1950's, all investigators of the semiconducting properties and p-n junction studies of germanium [63-66] and silicon [6,67,68,91,92] preferred to use pulled single crystals. Teal filed for a p-n junction patent in single crystal germanium in 1950 [93] and the first bipolar junction transistor (n-p-n) was achieved in single crystal germanium (grown-junction technique) by Shockley, Morgan Sparks and Teal in 1951 [66], three years after the discovery of transistor action by Bardeen and Brattain [1,2].

The conversion of germanium and silicon ores to metallurgical grade material and their subsequent purification during the 1940's has been reviewed by Frederick Seitz and colleagues [6,94-96]. Seitz also initiated and was the co-editor of the venerable *Solid State Physics – Advances in Research and Applications* series [97], servicing the needs of the physics community from 1955 onwards. Norman G. Einspruch later initiated and was the editor of the invaluable *VLSI Electronics Microstructure Science* series servicing the IC community [98].

Although Teal did not receive the acclaim accorded Bardeen, Brattain and Shockley, his pioneering research and implementation of single-crystal silicon technology as the basis of the IC microelectronics revolution can hardly be over-estimated [7,69,99-101]. The description of dopant distribution during single-crystal growth by normal freezing (see equation 1) was described by William Pfann, via the related zone-refining techniques [102-104], where $C_s(g)$ is the dopant concentration at the fraction of crystal solidified, g, C_l is the initial concentration of dopant in the liquid, d_l and d_s are the density of the solid and liquid phases, respectively, and k_e, the effective distribution coefficient, is the ratio of the solute concentration in the crystal adjacent to the melt-crystal interface relative to the concentration in the bulk liquid [102-104]. The relationship between the effective distribution coefficient, k_e, and the equilibrium distribution coefficient, k_o, during CZ crystal growth with the crystal growth rate, f (later recognized to be the microscopic crystal growth rate by Kenji Morizane in conjunction with Gus Witt and Harry Gatos [105]; the stagnant boundary layer thickness, δ, at the melt-crystal interface; and the bulk melt flow conditions influencing the solute diffusion coefficient, D, was described by the Burton-Prim-Slichter theory as described in equation 2 [106,107]. This equation was instrumental in conjunction with equation 1, in facilitating the availability of single crystals of germanium and silicon with a specific distribution of dopant impurity [7,63,79,80,82,108] to be discussed below. An extensive summary of the equilibrium distribution coefficients and solubilities for a variety of elements in crystalline germanium and silicon were summarized by Forrest Trumbore [109].

$$C_s(g) = k_e C_l (1- g)^{ke(dl/ds) -1} \quad (1)$$

$$k_e = k_o [k_o+(1- k_o) \exp(-f \delta/D)]^{-1} \quad (2)$$

The role of microscopic fluctuations in the dopant distribution, both radially and axially along the crystal, were to have profound implications on device performance to this day [8]. Pfann and colleagues initiated methodologies to rectify this situation [110].

Finally, a rather innocuous observation that silicon crystals grown by the CZ method contain parts per million atomic (ppma) concentrations of oxygen [111-113] (due to the dissolution of the quartz crucible by the molten silicon), and has had extremely important repercussions to the present day [114,115]. The higher melting point of silicon at 1420°C, compared to germanium at 937°C, resulted in contamination using the previous, conventional graphite containers. This contamination was overcome by utilizing fused quartz crucibles to hold the liquid silicon rather than graphite, with the unintended consequence that the melt became saturated with oxygen. This topic and methodologies for the localized removal of the oxygen from the near-surface regions of the fabricated device and its utilization as effective internal gettering bulk sites has been extensively discussed [7,8,115,116] and will not be further discussed in this review.

Bipolar Transistor Fabrication

Grown-Junction Bipolar Transistors

The path by which the role of group III and V impurities were deduced as p- and n-type dopants, respectively, in silicon and the critical role of metallurgy was reviewed by Jack Scaff [117,118]. The n- and p-type impurity dopants such as phosphorus and boron, respectively, were shown by Greiner's x-ray studies of the variation of the lattice constant with dopant concentration to occupy substitutional sites in the group-IVa semiconductors such as germanium and silicon, as reported in [71]. The inference was that all group III and group V dopants behaved in this manner in germanium and silicon [117]. The ground states are typically 40-50 meV from the appropriate band edge for silicon [6,119,120] (donors below the conduction-band edge, acceptors above the valence-band edge) and are readily ionized at 300K where the number of free carriers is essentially equal to the dopant density as determined by neutron activation analysis (NAA) [121].

The deduction of the role of group III and V impurities as p- and n-type dopants, respectively, in germanium, in conjunction with equations (1) and (2), led to the first grown junction n-p-n transistors, based on the "double-doping" technique in 1951 [64]. Since it was easier at the time to make good contact to a p-type base in an n-p-n transistor rather than to an n-type base in a p-n-p transistor, the former became commercially available, subsequently followed by the p-n-p transistor using a more complicated process [122]. Pellets of gallium and antimony alloys of germanium were sequentially (and rapidly) added to the melt during the growth of an n-type germanium crystal [64]. Only one slice of n-p-n germanium junction transistors, however, could be fabricated by this technique, which was subsequently superceded by Robert Hall's "rate-growth" technique, introduced in 1952 [123-125]. This technique is based on the variation of the incorporation of acceptor or donor impurities into the solidifying germanium semiconductor with the crystal growth rate. A series of germanium regions containing numerous p-n junctions (obtained by slicing the crystal) were grown within the same crystal [123] and n-p-n transistors with good yields and performance at intermediate radio frequencies were also achieved [124,125]. Further extension of the utilization of two different impurities in various structures within the same germanium crystal were carried out by Hall [123-125] as well as by Bridgers and Kolb [126,127].

With the subsequent development of the microwatt junction transistor in germanium [66], the benefits of larger current handling capability and less noise in the junction, compared to the point-contact, transistor [5] led to the escalation of the former, especially after improved techniques to control the base width—and thus increase the frequency response, an initial limitation [128]—were subsequently developed [129]. Specialized techniques to improve the point-contact transistor characteristics of germanium, however, such as the gold bonded diode [99] continued. Nevertheless, small-area silicon diodes replaced germanium point-contact diodes by about 1960. The silicon devices, first reported in 1952 by Gerald Pearson in conjunction with Sawyer and Philip Foy utilized the alloying technique (see below) to fabricate silicon diode rectifiers via an aluminum doped (p-type) wire alloyed to an n-type Si material that could operate up to 300°C [130,131]. It was clear, however, that the future was with the silicon bipolar junction transistor [129], which itself became eclipsed with the advent of the silicon MOSFET in the 1970's.

A commercially feasible grown-junction silicon transistor, introduced by Teal in 1954 [132], was subsequently described by Willis Adcock, in conjunction with Mort Jones and colleagues [133], although an experimental silicon transistor was previously announced in 1950 by BTL [134]. The silicon transistor raised the power output and doubled the maximum operating temperature previously attained by germanium transistors. These results clearly demonstrated that silicon was superior for transistor performance compared to germanium and vastly expanded the types of applications for which transistors could be used [6]. Morris Tannenbaum and co-workers subsequently reported that additions of gallium and antimony dopants supported the growth of single crystals of silicon containing up to five n-p-n regions of grown junction silicon transistors [135]. Junction transistors were cut from the slices and the base layer was located and contacted, which was not a trivial process.

Alloy Bipolar Transistors

Concurrently, John Saby developed the alloy transistor in 1952 [136] as did J. Trevor Law and colleagues [137], building on Hall's research [138,139]. The alloy process has been described, in retrospect, as crystal dissolution and regrowth or local liquid phase epitaxy (LPE) on both surfaces of silicon or germanium [22]. For example, arrays of indium dots were positioned on opposite surfaces of an n-type slice of germanium cut from a CZ grown germanium crystal (see Figure 2 for a schematic illustration of a p-n-p transistor) [4]. Alloying was accomplished, on individual die, in an inert atmosphere of about 600°C, during which the recrystallized germanium incorporated some of the indium, thereby converting to p-type. The achievement of a precise and narrow base width,

however, by controlling the alloying temperature was difficult and the achievement of very high cut-off frequency performance was also quite difficult [22].

interface to expose the {111} surface orientation during alloying [18]. This resulted in the decision to utilize that surface plane for device fabrication

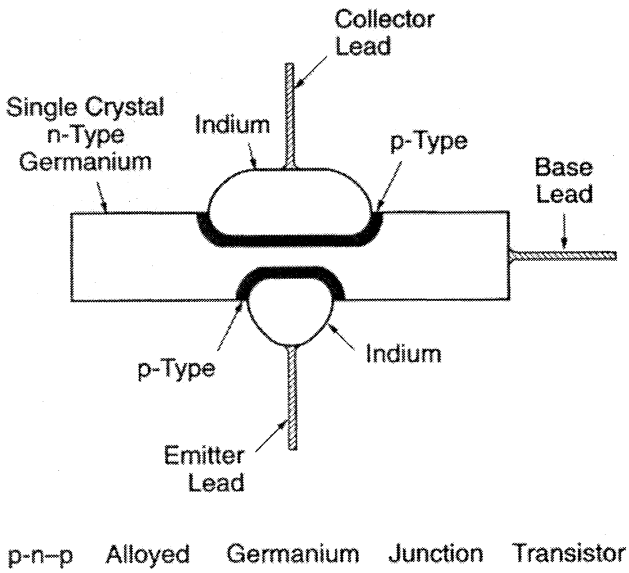

Figure 2. Schematic illustration of the alloy transistor [4]. Reproduced by permission of the IEEE, Inc.

For example, the maximum cut-off frequency of an alloy germanium p-n-p transistor was typically 10 MHz [140] while a germanium point-contact transistor exhibited typical values up to 50 MHz [141]. The fabrication problems were exacerbated inasmuch as the emitter-base junction and the collector-base junction were fabricated from opposite surfaces of the n-type and p-type silicon or germanium slice for p-n-p and n-p-n transistor fabrication, respectively [22]. The viability of achieving a controlled base width for a silicon n-p-n alloy transistor was quite difficult due to the variability of the process [22]. Nevertheless, alloying (because of its presumed simplicity) was readily implemented as a manufacturable process and became widely utilized (compared to the more uncontrollable base widths for grown-junction devices) for transistors in germanium [137,142] and silicon for many years [4], although the silicon alloy transistor was very difficult to fabricate and never commanded a significant market position [43].

The challenge of upgrading to the GHz range was a goal, required to support the extensive range of anticipated electronic applications. Since, as noted earlier, it was easier at the time to make good contact to a p-type base in an n-p-n transistor rather than to an n-type base in a p-n-p transistor [122], the latter did not become prominent in the marketplace. Finally, it should be noted that there was a strong preference of the melt-crystal to enhance the ability to form a uniform alloy; this choice of surface orientation was to continue throughout the bipolar junction and bipolar IC era. The onset of the MOS era in 1968, however, quickly shifted focus to the {100} orientation, which exhibited a reduced concentration of interface states at the $Si-SiO_2$ interface and facilitated better control of the MOS threshold voltage [24,143-145] as well as was advantageous for III-V devices (i.e., lasers).

There is a fundamental difference in the emitter-base and base-collector junctions between the alloy and grown-junction transistors. The alloy junctions are abrupt (of the "step" type) while the grown-junctions are graded. Accordingly, the alloy transistor exhibited a higher alpha cut-off frequency range (5-10 MHz) than the grown-junction transistor (1-10 MHz) due to the emitter-base step junction, although the abrupt base-collector junction for the alloy transistor resulted in a higher capacitance per unit area, tending to limit the high-frequency response. An alternate method of transistor fabrication, referred to as a surface barrier alloy transistor [146], was able to achieve cut-off frequencies up to 50 MHz by utilizing an electrochemical fabrication technique [147]. The approach was pioneered by Philco, using a jet etching technique, in which the germanium is etched by an electronically controlled jet of electrolyte [146]. Subsequent alloy contacts on each side of the thinned

base material resulted in a higher cut-off frequency, due to the factor ten smaller base width, compared to the grown-junction transistor. The mechanical fragility and low production yield, however, precluded the surface barrier transistor from becoming more widely disseminated. Concurrently, although the cut-off frequency of the point-contact germanium transistor had approached 50 MHz [147], it was the grown-junction and alloy junction devices, which continued to be produced on a mass-production basis into the late 1960's [43,147]. Likewise, the micro-alloy diffused transistor [148] received some attention, but was soon eclipsed by the introduction of the diffused mesa and planar transistor due to their lower leakage current and greater mechanical stability as described below.

Diffused Bipolar Transistors

The double-doping and rate-growth techniques were critical to proving out the junction transistor theory in practice. There were, however, inherent limitations in their manufacture as regards their commercial applicability. The junctions, for example, were physically inside the crystal and the p-type base layer was thicker and less uniform than desired. The introduction of solid-state diffusion procedures, with a key patent issued to Scaff and Henry Theuerer in 1951 (filed in 1947) [149] and implemented by Pearson and Calvin Fuller [150] rectified this situation via the in-diffusion of impurities in a controlled ambient over the whole slice of the semiconductor. The technique involved the exposure of the semiconductor slice to a vapor, containing the dopant of sufficient concentration, in a carrier gas to ensure the controlled dopant concentration at the semiconductor surface and at a sufficiently high temperature to create diffusion rates that would provide precise control of the dopant penetration depth in the semiconductor. Diffused layers from a few tenths of a μm to 20 μm were achieved. The initial study of the diffusion of donors and acceptors in germanium was published by Fuller [151] followed by Fuller and Ditzenberger's research of diffusion in silicon [152]. The silicon diffused junction rectifier was described by Prince in 1956 with peak reverse voltages of 400 V and current ratings of 400 mA [153]. By the mid 1950s, improvements in semiconductor processing facilitated the fabrication of both n-p-n and p-n-p transistors structures by solid-state diffusion processes in a mesa structure (see Figure 3) [4]. Lee fabricated a p-n-p germanium mesa transistor in 1954 with a base width of 1.0 μm by a diffused arsenic base and alloyed Al emitter; the current amplification factor and cut-off frequency were 0.98 and 500 MHz, respectively [154]. By 1959, germanium mesa transistors were being fabricated with base widths of 0.2 μm and, in the early sixties, silicon with cut-off frequencies approaching 1000 MHz were fabricated in double-diffused planar epitaxial structures (see below) [155]. Tannenbaum and Thomas fabricated a diffused base and emitter n-p-n mesa Si transistor with a base width of 2 μm in 1956, with a current amplification factor and cut-off frequency of 0.97 and 120 MHz, respectively [156]. Tannenbaum and Thomas's work is of especial importance in that it utilized the simultaneous diffusion of both the acceptor (for the base) and donor (for the emitter) dopants, albeit their concentrations were appropriately different to ensure the desired transistor action. Friedolf Smits reviewed the spectrum of solid-state diffusion techniques now available [157].

Mesa and Planar Processes

The mesa process was described by Aschner et al. in 1959 [158], who noted it was essential that the emitter diffusion proceed more rapidly than the base diffusion to retain control of the base width while the oxide masking and photolithographic technique, pioneered by Frosch and Derick [9,10] to be discussed below, was also utilized to remove a portion of the top of the semiconductor slice, resulting in the characteristic mesa structure. This geometrical modification was essential in order to reduce the p-n junction area so as to reduce the depletion layer capacitance and achieve the desired high frequency characteristic. Utilization of the diffusion (mesa) process, furthermore, opened the pathway for the fabrication of devices slightly below the planar surface of the semiconductor wafer, with far-reaching implications. These included narrow, well-controlled base widths, about a factor ten smaller than for the grown-junction and alloy transistors [159], and, thereby, higher frequency operation. Many transistors could be made at one time on each slice of Si or Ge during the "batch" processing with rather similar characteristics, especially important for adjacent devices with matched device characteristics for high-performance circuit applications, and many slices were available from each CZ grown crystal. Mesa transistors fabricated by the oxide masking and photolithographic technique were less expensive to fabricate compared to grown-junction transistors or alloy transistors, although all these fabrication methods, to some extent, were prone to excessive leakage currents. The initiation of the concept of the "learning curve," based on the reduction in the cost of

producing numerous identical devices through the cumulative process experience, was enunciated by Patrick Haggerty with implications to the present day [160-162].

With the advent of the diffusion process, the device frequency limitation was transferred from the base to the collector region [4]. It appeared that there was a basic, built-in design conflict inasmuch as the diffusion process resulted in the collector having the highest resistivity, compared to the emitter and base. Ross noted that "this led to significant series resistance in the collector, and that, combined with the capacitance of the collector junction, limited the frequency response" [4]. While the collector resistivity could be reduced, this resulted in a higher capacitance and lower breakdown voltage at the base-collector junction inasmuch as the base had to be highly doped to minimize the base resistance and small width to reduce the transit time. Jim Early had previously (before the introduction of the diffused transistor) proposed a device design solution by the introduction of an intrinsic or very high resistivity layer between the base and the collector to create a p-n-i-p structure [163]. This allowed an increased collector doping while retaining a low capacitance and high breakdown voltage of the collector since, under a reverse bias, the space charge region would penetrate the total width of the intrinsic region, suggesting a cut-off frequency as high as 3000 MHz. The state-of-the-art cut-off frequency in 1954 of 95 MHz was realized for a p-n-i-p germanium transistor proposed by Early [163]. Ross noted that Early "had the distinction of being the only person other than Shockley to propose a basically new transistor structure" [4], although one may regard the innovation more of a design modification. The fabrication of such a structure, however, required the diffusion or alloy process to be incorporated from both surfaces of the semiconductor slice, thereby resurrecting the experimental problem of accurate control of the base width. Nevertheless, Early significantly improved the understanding of the static characteristics of conventional bipolar transistors by also utilizing a heavily doped, thin base such that the space-charge widening in the collector (due to the reverse bias at the collector-base junction) enhanced the bipolar transistor's transit time [164,165]. This also had the effect of causing a portion of the depletion region to spread onto the base side of the junction (even though the base was more highly doped than the collector), thereby reducing, somewhat, the effective base width for the transport of minority carriers to the collector. This resulted in an increase of the current gain factor (α) and a decrease in the emitter potential required to ensure a given emitter current [164,165].

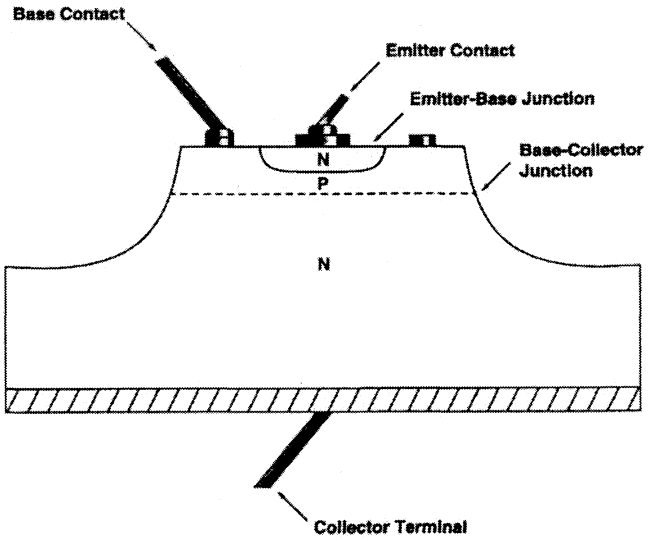

Figure 3. Schematic cross-section of an early mesa transistor made by Fairchild Semiconductor Corporation [122]. Reproduced by permission of the IEEE, Inc.

The solution to achieving the p-n-i-p or n-p-i-n structure was obtained by the fabrication of a lightly doped layer of silicon on a heavily doped silicon substrate, referred to as epitaxial fabrication [166-169]. Henry Theurerer and colleagues [169] expanded the applicability of epitaxial structures by implementing Bernard Murphy's localized, high-concentration sub-collector diffusion in a lightly doped silicon substrate [170-172], before epitaxial deposition, which enhanced bipolar performance by reduced collector resistance. The development of high-frequency, small signal devices via the newly developed planar process (see the *Planar Process* section below) was, however, limited to reverse junction breakdown voltages of only a few hundred volts and, since the junction area was small, to limited power handling capabilities. As a result, the development of high-power, high-voltage rectifiers and transistors with reverse junction breakdown voltages of several thousand volts as well as thyristors continued to proceed via utilization of the mesa process, although there was, naturally, strong interaction between these two complementary approaches (i.e., mesa and planar) so that the separation was not as sharp as might be indicated. Table 1 broadly summarizes the evolving technological trends to control the base width for junction transistors.

Table 1. Technology Evolution For Controlled Base Width Transistors

Technology	Author	Approximate Year	Reference
Double-doping	Teal	1951	64
Rate-grown	Hall	1952	123-125
Alloy	Hall	1950	138,139
Electrochemical thinned base	Tiley and Williams	1953	146
Diffused base	Pearson and Fuller	1954	150
P-N-I-P (N-P-I-N)	Early	1954	163
Planar process	Hoerni	1960	12
Epitaxy	Teal, Sangster, Mark, Theuerer	1954, 1957-1960	166-169

As noted earlier, silicon rapidly replaced germanium for transistor fabrication [6-8,16,43] as a result of silicon's larger energy gap which facilitated higher-temperature device operation and lower reverse current, effective four-layer n-p-n-p or p-n-p-n switching devices [19,173,174] as well as the plentiful availability of single crystals with the requisite purity, perfection and controlled electrical properties [6-8,86,119]. Germanium was relegated as a niche material for specialty devices, such as low-power, requiring performance metrics not readily achievable with silicon. Of especial importance, furthermore, was silicon's oxide, SiO_2, which was insoluble in water, whereas germanium's oxide was water soluble [3,7]. This attribute of silicon facilitated its utilization in the planar process as a diffusion mask for p-n junction fabrication as developed by Frosch and Derrick [9,10], the fabrication of the planar silicon transistor by Hoerni [12-15], passivation of the silicon surface and p-n junctions intersecting the surface by Atalla and colleagues [11] and a dielectric layer for supporting metallic conductor overlayers in the IC era [16]. Indeed, with the advent of the planar process, increased breakdown voltage behavior along with reduced reverse leakage current was achieved [175].

All the elements were now available (oxidation, diffusion, photolithography, aluminum metallization and thermocompression bonding [4,7,20-22] for the fabrication of junction transistors and the silicon controlled rectifier (SCR) (also referred to as the four-layer switch or thrysistor), in Moll's laboratory [18,21], in conjunction with Holonyak [22]. The SCR, developed by Moll in conjunction with Holonyak and colleagues, has a rich history [19,20,22,163,174].

The state of device physics had now reached a sufficiently sophisticated level that Holonyak relates that "Moll asked Holonyak, in conjunction with LeLacheur's familiarity with the systems requirements on switching transistors, to prepare a preliminary design "theory" to permit all of our colleagues to become familiar with what could be expected of the new silicon transistors" [22,176]. Inasmuch as this internal BTL memorandum [176] was not subsequently published, an excerpt is noted below [22].

> "Some of the design variables of diffused impurity transistors are discussed. Design compromises between series collector

resistance and collector capacity are found to be necessary. The relative advantages of linear and circular structures are considered both for base resistance and for collector capacity. Parameters, which are expected to affect the frequency behavior, are considered, including emitter depletion layer capacity, collector depletion layer capacity and diffusion transit time. Finally the parameters which might be obtainable are compared with those needed for a few typical switching applications."

The Planar Process

The development of oxide masking by Frosch and Derick [9,10] on silicon deserves special attention inasmuch as they anticipated planar, oxide-protected device processing. Silicon is the key ingredient and its oxide paved the way for MOSFET integrated electronics [22]. An account of their revolutionary development and utilization of SiO_2 as the vital foundation of today's IC industry has been described by Holonyak [22]:

"In building our various experimental devices, we were in contact with various groups and individuals, but above all with Carl Frosch. Frosch was a consummate process chemist who was familiar with many types of processing procedures and had been working, with his technician Derick, on impurity diffusion into silicon for several years. In spite of his considerable experience, Frosch, with dry gas diffusion procedures utilizing N_2 or H_2, regularly reduced many of our silicon wafers to "cinders, " particularly at higher temperatures ($\geq 1100°C$).

Because we had mastered building a diffused-base alloyed-emitter silicon p-n-p transistor (in spite of our problems with diffusion), one of the p-n-p-n configurations that we could explore was simply a modification of the p-n-p transistor: We could fabricate the diffused-base alloyed-emitter p-n-p on one side of a p-type substrate wafer after it first was prepared with an n-type diffused region (symmetrical) on both sides of the wafer. Either side could be chosen to form the p-n-p. The result was a p-n-p-n switch, in fact, the p-n-p-n switch of example (b) as described in [19]. (The complementary version of this exact structure, an n-p-n-p with Ga diffused into both sides of an n-type silicon wafer and then a Au-Sb emitter alloyed on one side, was later introduced at General Electric as the first commercial silicon controlled rectifier, today's thyristor. This later work was also based on our 1956 research [19].

In the process of diffusing the p-type substrate wafer into an n-p-n configuration for the first stage of p-n-p-n construction, particularly in the redistribution "drive-in" phase of the donor diffusion at higher temperature in a dry gas ambient (typically $\geq 1100°C$ in H_2), Frosch would seriously damage our wafers. The wafer surface would be eroded and pitted, or even totally destroyed. Every time this happened the loss was apparent by the expression on Frosch's face, not to mention, on ours (N.H.). We would make some adjustments, get more silicon wafers ready, and try again.

In the early Spring of 1955, Frosch commented to Holonyak, "Well we did it again," meaning the wafers were again destroyed. But then he smiled and displayed the silicon wafers – nice and green in color (in further instances also pink). He and his technician Derick had switched from a dry-gas (typically N_2 or H_2) impurity diffusion to a wet-ambient (H_2O vapor + carrier gas) diffusion, a consequence of an accident of the exhaust H_2 igniting and flashing-back into the diffusion chamber (because of gas flow fluctuations) and causing H_2O to cover, react with, and protect the silicon samples with oxide. The "wet" ambient, which was then immediately evaluated and adopted, created a protective oxide on silicon. It could be selectively removed for gaseous diffusion into the bare regions, which could then be resealed with oxide for higher temperature anneals or further diffusion. Many processing sequences could be devised for use of the protective oxide, which, of course, prevented crystal pitting and erosion. Frosch and Derick quickly found out which impurities were blocked from diffusion into silicon by the natural protective oxide (SiO_2) created in an H_2O-vapor ambient and which impurities would permeate the oxide (e.g., Ga). It was easy, once the issue of the oxide was known, to devise various schemes to diffuse into or to block impurity diffusion into silicon. The process was so flexible that planar n-type regions of any desired pattern could be prepared on a p-type substrate silicon, or the opposite, p-on-n diffused regions could be prepared on n-type silicon. All other diffusion procedures were suddenly rendered obsolete. We readily converted the Frosch-diffused silicon n-p-n into a working p-n-p-n switch [19]."