The Foundation of the Silicon Age

The transistor was the product of basic research with a clear technological goal, but although the new technology was anticipated, its revolutionary impact was not.

Ian M. Ross

If PHYSICS TODAY had been launched just one year earlier—in 1947 rather than 1948—it might have begun life by soliciting a number of articles to commemorate the 50th anniversary of the discovery of the electron by J. J. Thomson. That 1897 event could surely have qualified as the start of the electronics discipline and the industry that followed. It was the new understanding of the properties of the electron that created the field of electronics and that, combined with our developing capability in the electrical, magnetic and mechanical arts, made possible a rich array of new products and services.

In 1947, the tone of such articles would have been upbeat. Vacuum tube technology had fully matured, with a wide range of tubes—diodes, pentodes, cathode-ray tubes, klystrons and traveling-wave tubes—being in high-volume manufacture. Vacuum tubes were the key component in an array of electronic equipment that seemed to meet all conceivable information needs.

The then-director of research of Bell Telephone Laboratories might well have been invited to write an article. Mervin Kelly, who later became president of Bell Labs, would also have been upbeat. Electromechanical relay technology had provided fully automatic telephone dialing and switching. Microwave radio provided high-quality telephone transmission across the continent. Again, available technology appeared capable of meeting the needs.

Yet Kelly would also have raised a word of caution. Although relays and vacuum tubes were apparently making all things possible in telephony, he had predicted for some years that the low speed of relays and the short life and high power consumption of tubes would eventually limit further progress in telephony and other electronic endeavors. He not only predicted the problem, but had already taken action to find a solution. In the summer of 1945, he had established a research group at Bell Labs to focus on understanding semiconductors. The group also had a long-term goal of creating a solid-state device that might eventually replace the tube and the relay.

Kelly's vision triggered one of the most remarkable technical odysseys in the history of mankind, a journey that has continued through 50 years. The semiconductor odyssey produced a revolution in our society at least as profound as the introduction of steam engines and steel, as well as the total industrial revolution. Electronics today pervades our lives and has an impact on everything we do at work and at home.

In this article, I discuss the events that led to the invention of the transistor, the hurdles that had to be overcome and the breakthroughs that were needed to make the semiconductor revolution a reality.

IAN ROSS *was president of AT&T Bell Laboratories from 1979 to 1991. He joined Bell Labs in 1952 and worked in semiconductor R&D until 1964. He and George Dacey fabricated the first working field-effect transistor in 1952.*

The scientific phase

By January 1946, Kelly's semiconductor group was in place at Bell Laboratories under the leadership of William Shockley and Stanley Morgan. Two key members of the team were John Bardeen and Walter Brattain. Other members included Gerald Pearson, Bert Moore and Bob Gibney. The team was embedded in the unusually creative environment that existed at Bell Labs, in Murray Hill, New Jersey, after World War II. As such, the team members were able to seek the advice of resident experts in almost any relevant discipline.

They had a number of other assets to call on in their pursuit of Kelly's goal. There existed a large body of empirical knowledge of semiconductor devices based on experience with diodes for the detection of radio signals. There was also considerable experience with power rectifiers such as copper oxide diodes. Those devices were made from a variety of semiconductor materials, but most were highly impure and none was single crystal. There was much art and much tinkering, but little engineering understanding and almost no science.

There was already some basis for understanding the physics of semiconductor materials. The concept of bandgaps existed. Two types of conduction, already named n-type and p-type, had been identified, and attributed to the presence of certain impurities in very small concentrations. What were called p-n junctions had been found within ingots formed by melting and refreezing the purest silicon then commercially available. Their electrical and electro-optical characteristics had been explored. And considerable progress had already been made at Purdue University, Bell Labs and elsewhere in producing semiconductor materials of increasing purity and in understanding their properties.

However, there was also much uncertainty, much still unknown. The highest purity semiconductor available was orders of magnitude short of that eventually needed. Semiconductor materials were polycrystalline at best and frequently used in powder form. The key properties of

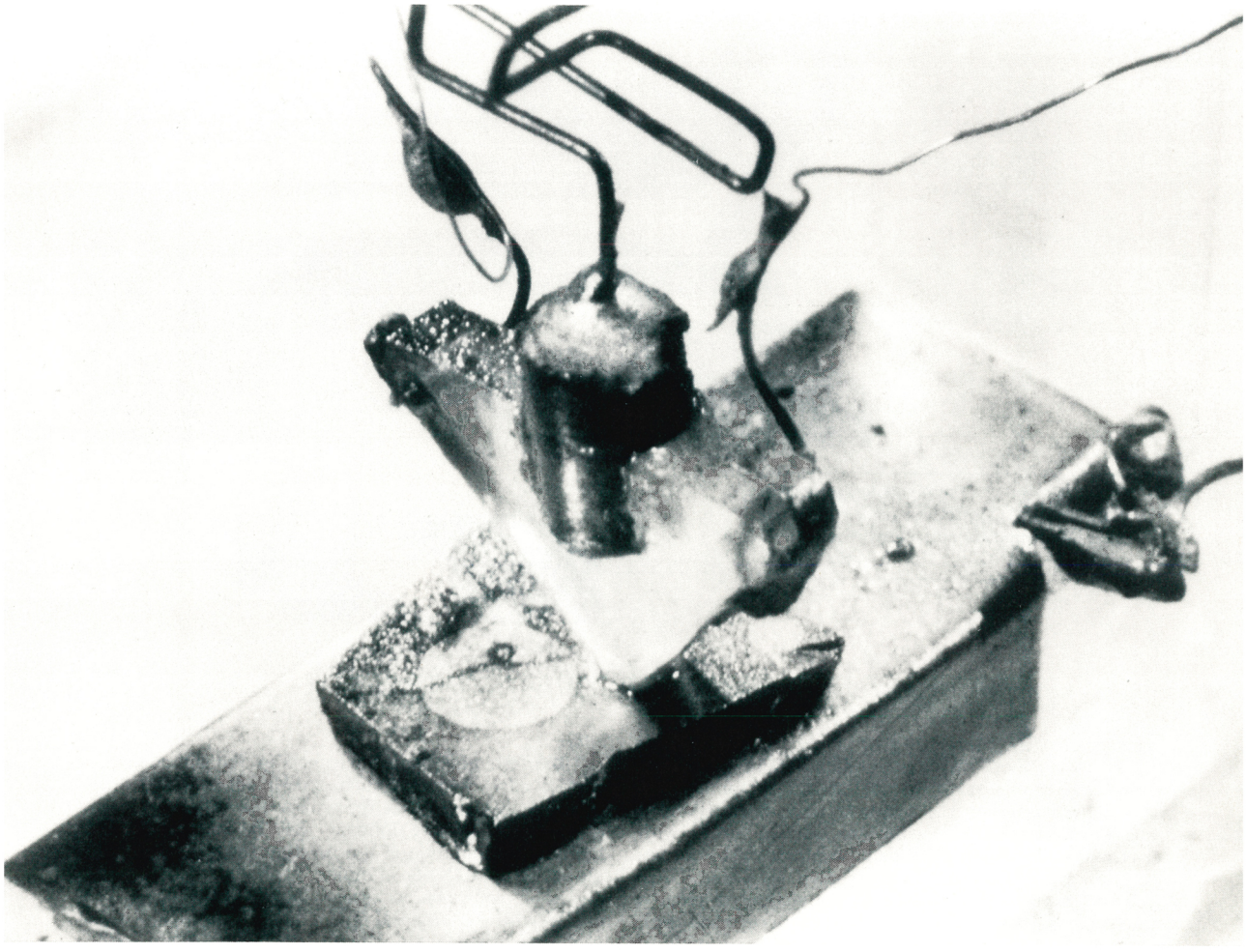


FIGURE 1. THE ORIGINAL TRANSISTOR structure. Gold foil was cemented over the point of a wedge of insulator and sliced at the point with a razor blade. The point was then pressed onto the germanium surface.

the materials relevant for device applications had yet to be fully understood and evaluated.

Finally, there was a long and persistent history of proposals for a solid-state amplifier. Most were based on the so-called field-effect mechanism. The concept was that an electric field applied through the surface of a semiconductor could modify the density of mobile charge in the body of the material and thereby change its conductivity. The first documented proposal of this kind had been made by Julius Lilienfield as early as 1925. All attempts to make such a device, however, had failed.

Both before and after the war, Shockley had studied and analyzed possible field-effect structures and had concluded that the effect could lead to amplification in achievable structures. Shockley's existence proof provided major encouragement that the challenge undertaken by the Bell Labs group could indeed be accomplished.

By January 1946, two critical decisions had been made. The first was to focus the group's attention on crystals of silicon and germanium and ignore other, more complex materials. It was recognized that silicon and germanium were stable elements that readily assumed the crystalline state, and therefore showed the best promise of being made into high-purity, high-perfection single crystals. Such materials would permit the investigation to move forward on a sound scientific base. The second decision was to pursue the field-effect principle as the one most likely to lead to a useful device.

Given this renewed focus, a number of new experiments were carried out at Bell Labs by J. Richard Haynes, Henry J. McSkimin, William A. Yager and Russell S. Ohl in attempts to observe the field effect. All gave negative results. Bardeen proposed that the experiments failed because the electric field was not penetrating the body of the semiconductor material but was terminated by immobile charges trapped in states at the semiconductor surface. He calculated that a quite small number of such surface states, low compared to the density of surface atoms, would be adequate to shield the body from any measurable field effect.

Bardeen and Brattain attempted to confirm this theory by experimenting with metal probes on the surface of

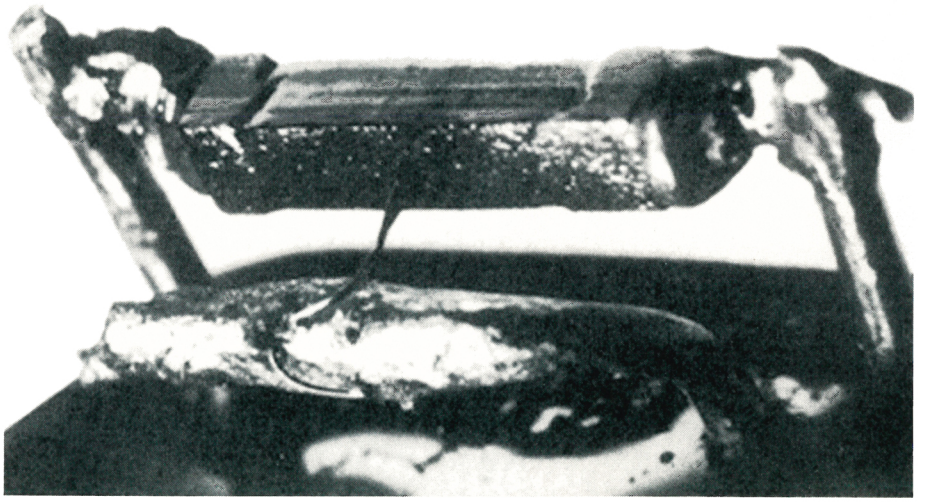
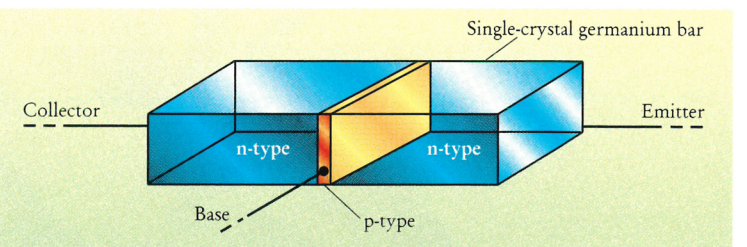


FIGURE 2. FIRST GROWN-JUNCTION TRANSISTOR. Contact was made to the base region by manipulation of a gold wire under a microscope.



germanium. The theory seemed to be correct. Thus for the first time there was some understanding of the persistent failure to observe the field effect, and an opportunity to intervene. In the course of their work, Bardeen and Brattain tried to modify the surface states with electrolytes surrounding the metal contacts to the germanium surface. Following a suggestion by Gibney, they found that applying a voltage to the electrolyte created major changes in the current flow through a reverse-biased contact. Brattain later replaced the electrolyte with an evaporated gold spot adjacent to the point contact. Finally, he replaced both contacts with an ingenious arrangement of two strips of gold foil separated by a gap of some tens of micrometers and pressed onto the germanium surface. With one gold contact forward-biased and the other reverse-biased, he observed power gain. The transistor effect had been discovered. (See figure 1.) That was on 16 December 1947, a mere two and a half years after the formation of Shockley and Morgan's group!

On Christmas Eve of 1947, Brattain and Moore demonstrated the transistor action for the top management of Bell Labs. This time the device was operated as an oscillator, an acid test of the existence of power gain. The announcement of the transistor discovery was delayed, however, until June 1948. This six-month period was used to gain more understanding of the device and its possible applications, and to obtain an adequate patent position.

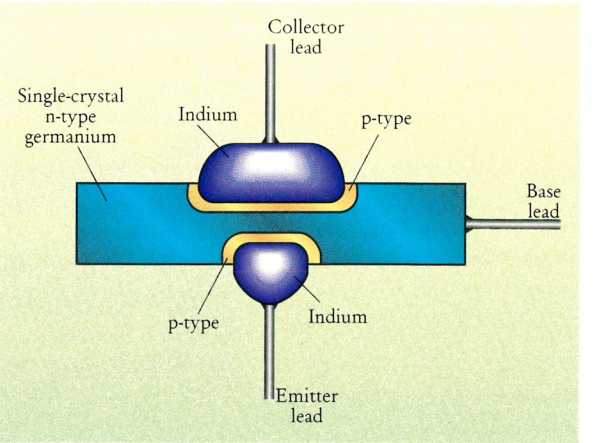
The invention of the point-contact transistor—the gold foil having been replaced by two closely spaced point contacts—opened the door to a whole new era of electronics. But the process of inventing the transistor still had a long way to go! Transistor action had been observed, but no one understood just what was the mechanism. Was it a surface effect or was the action occurring in the semiconductor body? Ironically, the mechanism certainly was not the field effect that had helped guide the whole effort.

Bardeen and Brattain leaned in the direction of a surface effect and continued experiments on that basis. Shockley, however, had recognized the role of minority carriers, and by late January 1948 he had completed a thorough formulation of p-n junction theory and the role played by the injection of minority carriers in forward bias and their collection in reverse bias. His analysis concluded with the proposal of a junction transistor, a sandwich of lightly doped n-type material between two regions of p-type—or the other way around. With one p-n junction forward-biased and the other reverse-biased, minority carriers would be injected from the forward-biased junction into the n-type material. They could then diffuse across the n-type region and, if it were thin enough, a large fraction would be collected at the reverse junction. Thus, current generated in a low-impedance circuit, the emitter, would create a similar current flow in a high-impedance circuit, the collector, and power gain would result. But so far this was just theory.

One month later, in February 1948, John Shive carried out a critical experiment. He applied two phosphor–bronze contacts to the opposite sides of a 0.1 mm thick slice of germanium. With this arrangement, he observed transistor action from one contact to the other with substantial power gain. The length of the surface path around the semiconductor slice effectively ruled out a surface effect. The action had to take place through the semiconductor body. The behavior he observed was nicely explained by Shockley's recently developed theory of the junction transistor. Thus, while the point-contact transistor may have exhibited some surface effects, bulk propagation was also surely taking place and was probably the dominant effect.

The next major advance was made later in 1948. Gordon K. Teal and John B. Little succeeded in growing

FIGURE 3. ALLOY TRANSISTOR scheme. This device was produced through a batch process. However, the base thickness was difficult to control, being the small difference between the thickness of the original wafer and the sum of the alloy depths.



a single crystal of germanium by slowly pulling a seed crystal from a melt of high-purity germanium. Using such material, it was at last possible to detect and characterize minority carriers injected by metal contacts into filaments of germanium. Various elegant experiments by Haynes, Pearson, Harry Suhl and Shockley confirmed the behavior of both types of minority carriers and yielded measurements on injection efficiency, mobility, diffusion coefficients and lifetime. Their results showed that useful devices could be made according to Shockley's junction transistor theory. All that remained was to make one.

That required further refinement of the techniques of crystal growth and particularly of the controlled doping of the crystals during growth. In April 1950 a team consisting of Shockley, Morgan Sparks and Teal succeeded in growing a crystal containing a thin region of p-type material embedded in n-type material. The crystal was cut into n-p-n rods and contacts were applied. The electrical properties of the resulting devices (figure 2) were largely consistent with Shockley's theory. Transistor electronics now had a solid foundation.

So, in a period of only five years following the establishment of the semiconductor group at Bell Labs, the invention of the transistor was essentially complete, understood and documented. The scientific phase was coming to an end. The next phase would focus on solving development and engineering issues so that a brilliant invention could be converted into an important innovation.

Once the transistor was invented, the challenge was then to find ways to design a product that could be manufactured, and that could sustain a market. This phase took the industry approximately eight years, during which many challenging problems were addressed and solved. Whereas Bell Labs had dominated the scientific phase, there were now other companies in the business, and they also made major innovations.

Below I describe some of the major hurdles that had to be overcome and the major breakthroughs that were made. Many events made a difference; I focus here on those that made *the* difference.

Early manufacturing problems

In early 1951, there were two transistor structures that were proven to work, but neither was suitable for large-scale manufacture. The point-contact transistor was difficult to make and its electrical characteristics were far

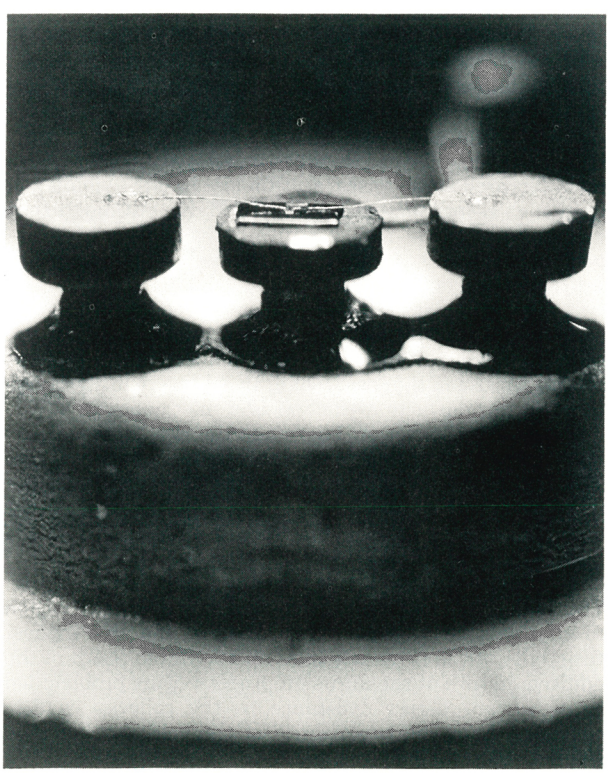
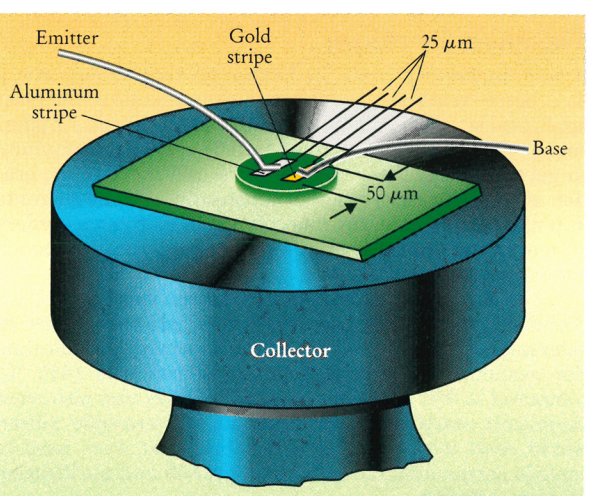


FIGURE 4. MESA TRANSISTOR. The collector junction was contained in a mesa-like area created by etching away the outer material.



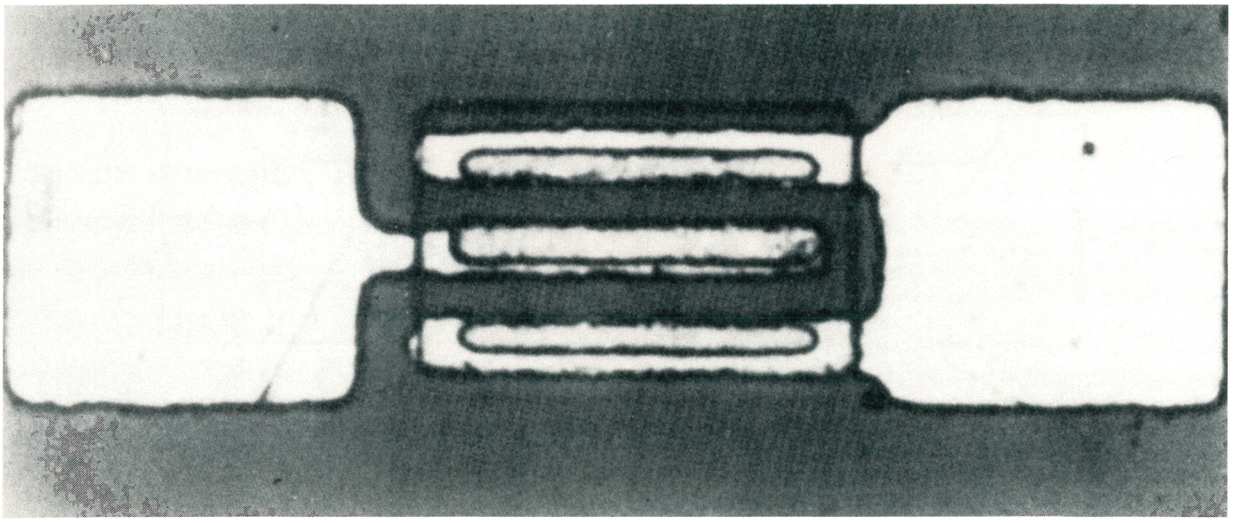


FIGURE 5. PLANAR TRANSISTOR. The light areas are aluminum contacts, made over the protective oxide, to the emitter region (the single central stripe) and to the base region (the two outer stripes).

from ideal—very variable, hard to control and inherently unstable. The junction transistor, on the other hand, had predictable and more desirable electrical characteristics. It was, however, prodigal in its use of precious semiconductor material and required tricky contacting techniques not conducive to automation.

In 1952, John E. Saby at the General Electric Co announced the development of the alloy junction transistor. The original version was made by alloying dots of indium (an acceptor material that makes germanium a p-type semiconductor) on opposite sides of thin slices of n-type germanium. The starting point was the growth of uniformly doped crystals that were relatively easy to produce. Slices were cut from the crystal, most of which could be used. Arrays of indium dots could be positioned in jigs on either side of the slices and, after alloying, the slice could be diced to yield a great many individual transistors. Contacts were easy to apply. The alloy transistor had well-behaved performance characteristics, made efficient use of semiconductor material and could be manufactured with some degree of batch processing and automation. The alloy device was the first transistor to be readily manufactured, and for some years was the mainstay of the industry. (See figure 3.)

The quest for silicon

It was understood from the beginning that silicon would be a better transistor material than germanium for most applications. This mainly resulted from silicon's higher energy gap—1.1 eV compared to 0.67 eV for germanium. In germanium at room temperature, the thermal generation of minority carriers led to substantial reverse currents in p-n junctions. The reverse current in silicon was orders of magnitude smaller and made the material a much superior rectifier.

The most serious problem with silicon was that critical chemical and metallurgical processes all took place at substantially higher temperatures than with germanium. For example, the melting point of silicon was 1415 °C, compared to 937 °C for germanium. Silicon was also more chemically reactive than germanium. For example, silicon would react with the quartz crucibles that were used to contain germanium during crystal growth and purification by zone refining.

The critical breakthrough came in 1953 with the development by Henry C. Theuerer of the floating-zone method. He was able, in a vertical rod of silicon, to create a zone of molten material contained only by surface tension. Thus the zone refining technique could be used for silicon and resulted in crystals of purity comparable to the best obtained in germanium.

In 1954, Teal, then at Texas Instruments, made the first silicon transistor using the grown-junction method. All the pieces were then in place for silicon devices to assume a major role. It was soon realized that that role was not simply to replace the vacuum tube, but to do things the vacuum tube could never do, such as producing high-density logic circuits.

The speed problem

The fundamental determinant of the frequency response of a junction transistor was the transit time of minority carriers across the base region—and therefore the thickness of the base layer. In practice, alloy transistors were manufactured with bases as thin as 10 μm, yielding a frequency response approaching 10 MHz. Although that was quite a feat of manufacturing engineering, performance up to a few gigahertz was needed to support a full range of electronic applications.

The base width problem was solved by using the process of diffusion of donors and acceptors into the semiconductor surface. This solution eventually yielded precise control of the depths of diffused layers in the range from 20 μm down to a fraction of a micrometer.

In 1954, Charles A. Lee made the first diffused germanium transistor, with a base thickness of about 1.0 μm. This transistor, also known as the mesa transistor, had a cutoff frequency of 500 MHz. (See figure 4.) A year later the first diffused silicon transistor was made and had a frequency cutoff at 120 MHz.

The speed problem was almost solved—but not quite. The frequency limitation had moved from the base region to the collector region. The collector had the highest resistivity of the three regions—an inevitable result of the additive nature of the diffusion process. This problem led to significant series resistance in the collector, which, combined with the capacitance of the collector junction, limited the frequency response.

The eventual solution was to add a totally different process: the epitaxial growth of a lightly doped layer of single-crystal semiconductor on a substrate of a heavily doped single crystal. A transistor base and emitter layer was then diffused into the epitaxial layer. The results were published in 1960 by Theuerer, Joseph J. Kleimack, Howard H. Loar and Harold Christensen.

Oxide masking and photolithography

In 1955, Carl J. Frosch and Link Derick at Bell Labs made a very important observation. They discovered that a layer of silicon dioxide a few hundred nanometers thick and grown on the surface of silicon prior to diffusion could mask the diffusion of certain donor and acceptor atoms into the silicon. They also demonstrated that diffusion would occur unimpeded through windows etched in the oxide layer. Somewhat later, Jules Andrus and Walter L. Bond at Bell Labs showed that certain photoresists deposited on the oxide surface would prevent etching of the oxide. Hence, optical exposure of the resist by projection or through contact masks could be used to create precise window patterns in the oxide and in turn provide precise control of the areas in which diffusion would occur.

This combined process of oxide masking and photolithography has since been developed to the point that junction areas can be controlled to a fraction of a micrometer. This development complements the precision of the depth control of junctions diffused into the silicon surface, providing the means to control the fabrication of silicon devices in three dimensions to the precision of a fraction of a micrometer.

These advances have also ended the role of germanium as a major player. No material was found that would provide diffusion masking for germanium.

The reliability problem

It was found in the early days that the transistor was very sensitive to its environment, particularly to humidity. This lack of reliability was a huge setback and embarrassment to the semiconductor community. The transistor had been lauded as a device with no failure mechanisms. Instead, we had a severe reliability problem, and one that took almost 20 years to solve completely.

The immediate remedy was to hermetically seal the devices in packages using the metal-to-glass seals from vacuum tube technology. This was a further blow to the pride of the semiconductor engineer. The packaging art evolved using a variety of empirical procedures, including vacuum baking, dry gas baking and gettering. It is remarkable that with these unscientific approaches, germanium transistors were eventually manufactured with failure rates of less than one per hundred million socket-hours.

There were also ongoing systematic studies to try to understand the problem and find a more fundamental solution. At Bell Labs, M. M. "John" Atalla led a group that studied the surface properties of silicon in the presence of a silicon dioxide layer. They speculated that growing an oxide layer under very clean and controlled circumstances on the surface of carefully cleaned silicon could lead to a reduced density of surface states and might serve to protect the surface against further change. In 1959, they did confirm that the presence of an oxide layer could reduce the density of surface states to such a level that the field effect could be observed. However, they had difficulties gaining enough control of the process to get reproducible results. Nevertheless, the concept that an oxide layer might provide a solution to the reliability problem was a major step forward.

The final breakthrough in reliability came with an invention made by Jean A. Hoerni at Fairchild in late 1957 or early 1958 and published in 1960. Hoerni proposed that, in the course of fabricating diffused silicon transistors, the silicon dioxide layer that was used as a diffusion mask be left in place. The junctions at the silicon surface were thus under the oxide layer, and Hoerni speculated that the oxide could protect the junction areas from contamination. He indeed found that such junctions had acceptable characteristics without further treatment. It was a startling result, particularly for those who believed that a passivating oxide would need to be grown under meticulously clean conditions.

Hoerni's result was not the end of the story, but put us on the right track. Not until about 1966, though, were techniques developed to produce satisfactory oxide layers and to "overcoat" them to retain their properties. Silicon devices then needed only to be further encapsulated in plastic for protection against gross environmental effects. Transistors, after all of 20 years, no longer looked like small vacuum tubes.

The planar transistor

In his 1960 paper, Hoerni also described the planar transistor. In that concept, both the base and emitter regions were diffused through windows in silicon dioxide masks so that both collector and emitter junctions terminated at the surface. The masking oxides were left in place and provided protection and eventually passivation of the silicon surface. Ohmic contact was made to both the base and emitter regions through windows in the oxide layer. The metal used for all contacts was aluminum, which Moore and Robert N. Noyce had previously shown would make good contact to either n- or p-type silicon. Moore had also shown that the aluminum could be extended over the oxide to form larger pads to ease connections to the chip. Somewhat later the epitaxial process was added to the planar transistor to minimize collector resistance.

This proposed structure brought it all together. (See figure 5.) All the key development and engineering problems were either solved or on course for an elegant solution. There was a sound foundation for the long-term manufacture of semiconductor devices. Silicon, the semiconductor of choice, could be produced with a crystalline perfection and purity more than adequate to the task. Critical dimensions in all three directions could be controlled if necessary to a fraction of a micrometer. Electrical contacts could be made with a single metal and without the need for microscopic precision. The resulting devices would eventually be solidly reliable. And all of these properties could be achieved by using batch processing, with the promise of high yield and low unit cost.

By the late 1950s, scarcely more than a dozen years after its discovery, the transistor had a sound engineering foundation, which provided the base for the next giant step: The integrated circuit was invented in 1958 by Jack S. Kilby at Texas Instruments with a major added contribution from Noyce at Fairchild.

This is a modified and shortened version of an article I have written for the Proceedings of the IEEE, transistor anniversary special issue, January 1998. I have relied heavily on accounts in Engineering and Science in the Bell System, volume 4, published by AT&T Bell Laboratories in 1985, and particularly on the section entitled "The Transistor," written by John Hornbeck and edited by Friedolf Smits. Friedolf was also helpful in clarifying some of the events covered in the text. I am grateful to Gordon Moore for input on the planar transistor. Bill Troutman of Bell Laboratories, Lucent Technologies (formerly AT&T Bell Laboratories) provided valuable reference material and even more valuable encouragement. I thank AT&T and Lucent Technologies for their support and particularly for providing material from their archives. ■