The Invention of the Transistor

IAN M. ROSS, LIFE FELLOW, IEEE

Invited Paper

The invention of the transistor almost 50 years ago was one of the most important technical developments of this century. It has had profound impact on the way we live and the way we work.

The first part of this paper covers the events that led to the discovery of the transistor effect and the invention of the point contact transistor in December of 1947. It continues with the development of the theory of the junction transistor in early 1948 and the fabrication of the first grown junction transistor in 1950. It is fair to say that this event completed the invention of the transistor and developed a fundamental understanding of how it worked.

The second part of the paper describes the major hurdles that had to be overcome and the major breakthroughs that had to be made to turn an exciting invention into a far reaching technical innovation. This phase took approximately another 10 years. By that time, high performance, high reliability transistors could be manufactured in large quantity and at low cost. Importantly the foundation had been laid for the invention of the integrated circuit and the dramatic development of the microelectronics industry.

The final part of the paper suggests some of the reasons why such an important technological innovation could occur in a relatively short period of time and be such an unqualified success. Finally, there are some comments on how much further this technology can go and when its rapid progress may come to an end.

***Keywords*—** *Bipolar transistors, history, materials processing, MOSFET, semiconductor device fabrication, technological innovation, transistors.*

I. SETTING THE COURSE

In the summer of 1947, the forerunner of the IEEE could have scheduled a special fiftieth anniversary edition. It was indeed 50 years since the discovery of the electron by J. J. Thomson. That event in 1897 could surely qualify as the start of the electronics discipline and the industry that followed. It is true that there already existed products and services that today would be considered electronic. Telephone service was developing rapidly. The first telephone switching machine had been invented in 1889. New York City and Chicago, IL, had been linked by telephone in the same year, and the telephone dialer was developed in 1896. The first magnetic sound recorder was developed in 1893. In 1874, F. Braun observed rectification at metal contacts to galena (lead sulfide). These semiconductor devices were the forerunners of the "cat's whisker" diodes used to detect radio signals. In 1896, Marconi received patents on his radio systems, which were then capable of ranges of 2–9 miles. Most significantly, the cathode-ray tube (CRT), the first electron tube, was developed before 1897. Indeed, Thomson used a form of cathode-ray tube to make his discovery. Despite the existence of these products and services, 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, enabled a rich array of new products and services.

The special edition would have been an upbeat event. Vacuum-tube technology had fully matured with a wide range of tubes—diodes, pentodes, CRT's, klystrons, and traveling-wave tubes—in high-volume manufacture. Vacuum tubes were the key component in an array of electronic equipment that seemed to meet all conceivable information needs. Radio equipment—AM, FM, and microwave—was in wide use. There were automobile radios and even radio sets that were considered portable. Radio transmission had provided worldwide telephone connections. Black-and-white television was in commercial service, and color was already planned. There were on the market any number of electronic devices, including fax machines, calculating machines, and both audio and video tape recorders. Digital technology was beginning to emerge. Several electronic computers were in operation, and software programs were developing. Electronics did indeed seem capable of satisfying all conceivable needs for information services.

The then director of research of Bell Telephone Laboratories might well have been invited to submit a paper to the special edition. M. 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. Radio also was spanning the oceans with somewhat flaky service, but a quality solution was soon to be available. Plans were under way to design and deploy undersea cables with vacuum-tube repeaters. The system was designed to operate for at least 20 years. Again, available technology appeared capable of meeting the needs.

Manuscript received September 30, 1996.

I. M. Ross, retired, was with Bell Laboratories, Murray Hill, NJ 07974 USA. He is now at 5 Blackpoint Horseshoe, Rumson, NJ 07660 USA.

Publisher Item Identifier S 0018-9219(98)00747-6.

Yet Kelly also would 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, he had already taken action to find a solution. In the summer of 1945, Kelly had established a research group at Bell Labs to focus on the understanding of semiconductors. It 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 steel, of steam engines, and the total industrial revolution. Electronics today pervades our lives and affects everything we do, whether at work or at home.

Before the invention of the transistor, we had two devices to perform the logic and amplification functions we needed. The first was the relay, in which an electromagnet was used to move one piece of metal on a spring into contact with another piece of metal. A small amount of power in the magnet circuit could control a larger amount of power in the contact circuit. The device was simple, rugged, and quite reliable—an excellent switch. The limitation was that relays moved at "mechanical" speeds—it took about a thousandth of a second to open or close the contact. It was a million times slower than its competitor, the vacuum tube.

The vacuum tube relied on the flight of electrons in a vacuum. Electrons were emitted from a hot wire filament within a glass bulb and were collected at a positively charged plate. The flow of electrons out of the filament was controlled by a small voltage on an electrode, called the grid. The vacuum tube could switch and amplify, and since electrons can travel at high speed in a vacuum, it was very much faster than the relay. But the vacuum tube used considerable standby power, and its lifetime was limited. The filament would burn out or the bulb would leak and the tube would fail. Thus, the vacuum tube was fine in applications where only a few were needed, such as in radios. Where thousands were needed, as in computers, the vacuum tube could not hack it.

A solid-state device promised the best of all worlds: electrons traveling short distances in a solid, no moving parts, no hot filaments, and no vacuums. It should thus be fast, cheap, and reliable. The invention of the transistor gave us such a device. Eventually, it met this promise and much more. But it took many years and a lot of hard work to convert an exciting invention into the revolutionary innovation that has changed our lives in profound ways.

My purpose in this paper is to discuss the events that led to the invention of the transistor, plus the hurdles that had to be overcome and the breakthroughs that were needed to make the semiconductor revolution a reality. In doing this, I have tried to select those events that made "the" difference rather than cover the multitude of contributions that made "a" difference. I admit that there is some judgment in making this selection.

II. The Scientific Phase

By January 1946, Kelly's semiconductor group was in place at Bell Labs under the leadership of W. Shockley and S. Morgan. Shockley was a very capable physicist, an analyst, and a man with a fascination for finding practical applications of science. Two key members of the team were J. Bardeen and W. Brattain. Bardeen was a remarkably talented theoretical physicist, as evidenced by the fact that he was awarded two Nobel Prizes in physics, each in a field of major significance. Brattain also was an accomplished physicist with a flair for ingenious experiments. Other members included G. Pearson, B. Moore, and R. Gibney. The team was embedded in the unusually creative environment that existed in Bell Labs Murray Hill (NJ) after World War II. As such, team members were able to seek the advice of resident experts in almost any relevant discipline.

The group 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 detection of radio signals. These diodes ranged from the "cat's whisker" crystal diodes at the heart of early radio receivers to the microwave diodes used in great quantities during the war for radio and radar detection. There also was considerable experience with power rectifiers such as copper oxide diodes. These devices were made from a variety of materials, including selenium, lead sulfide (galena), copper oxide, germanium, and silicon. All were semiconductor materials, most were highly impure, and none was single crystal. There was much art, much tinkering, but little engineering understanding and almost no science.

There already was a basis for understanding the physics of semiconductor materials. The concepts of band gaps and two types of conduction, already named n-type and p-type, had been identified in semiconductors and attributed to the presence of certain impurities in very small concentrations. Some of this work was done at Bell Labs during the war by J. Scaff, H. Theuerer, and R. S. Ohl. In the case of germanium, it was Theuerer who first identified the presence of phosphorus as an n-type agent as a result of smelling traces of phosphine during ingot preparation. p-n junctions had been found within ingots formed by melting and refreezing the purest silicon then commercially available. Their electrical and electrooptical characteristics had been explored. Considerable progress had already been made at Purdue University (West Lafayette, IN), Bell Labs, and elsewhere on producing semiconductor materials of increasing purity and on understanding their properties.

There also was much uncertainty, however, much still unknown. The highest purity silicon available—99.8%—was characteristic of a soap advertisement and orders of magnitude short of that eventually needed. Semiconductor materials were polycrystalline at best and frequently used in powder form. Single crystals of adequate perfection had yet

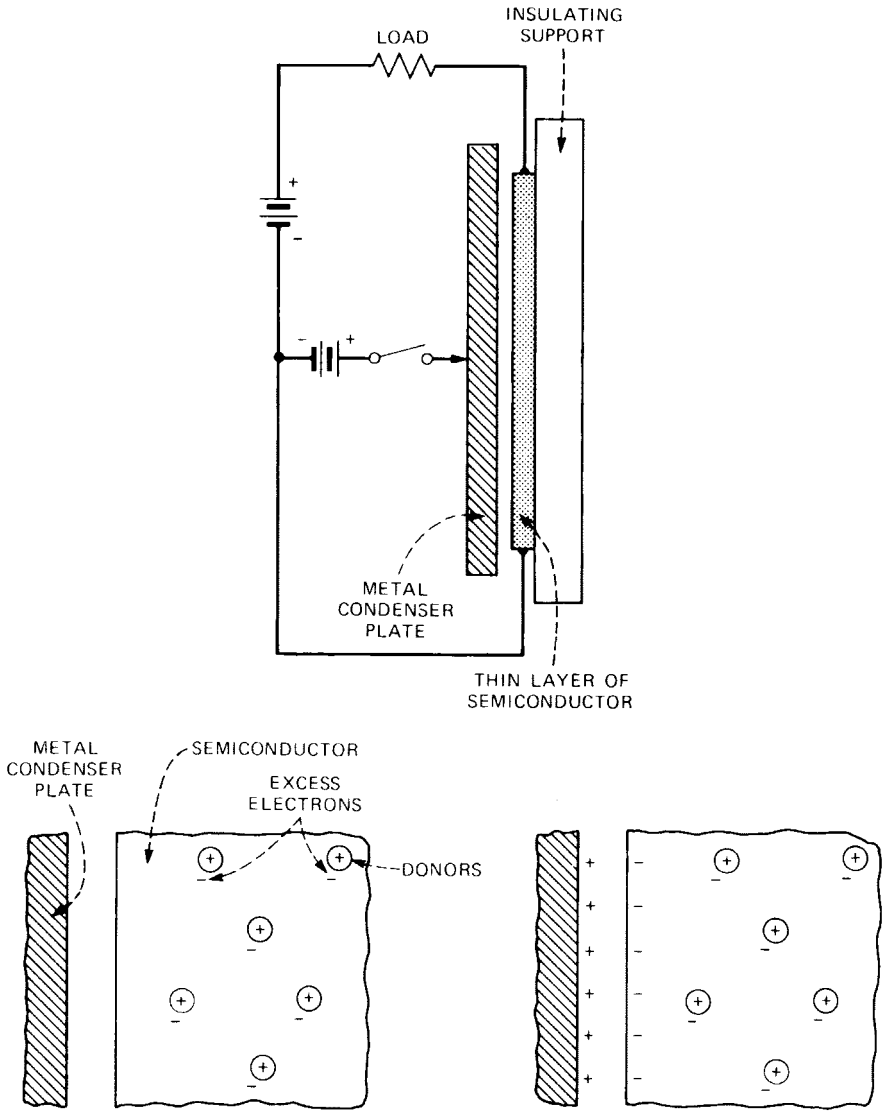


Fig. 1. Schematic of Shockley's field-effect idea. [12].

to be grown. The key properties of these materials relevant for device applications had yet to be fully understood and evaluated.

Last, 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 charge in the body of the material and thereby change its conductivity. Typically, the field was to be created by applying voltage to a metal plate close to but insulated from the base material. Modulating the voltage on the plate would modulate a current flow through the base material with the possibility of power gain. The first documented invention of this kind was made by 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 should lead to amplification in achievable structures (Fig. 1). Shockley's existence proof that amplification was theoretically possible in practical semiconductor materials 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 frequently used in prior investigations. 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 having the most assurance of leading to a useful device.

Numerous attempts to demonstrate the field effect in semiconductors had been made over the years and all had failed. Before the war, Shockley had participated in one such failure using a structure with a grid of metal filaments

buried in the body of a semiconductor. Given the renewed focus, a number of new experiments were carried out by J. R. Haynes, H. J. McSkimin, W. A. Yager, and Ohl in attempts to observe the field effect. All gave negative results. Bardeen proposed that these 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 [1]. 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 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, they tried to modify the surface states with electrolytes surrounding the metal contacts to the germanium surface. Following a suggestion by Gibney [2, p. 97], they found that applying 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. Last, he replaced both contacts by an ingenious arrangement of two strips of gold foil separated by just a few mils 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 [3] (Fig. 2). This was on December 16, 1947, a mere two-and-a-half years after the formation of the Shockley group.

On Christmas Eve, 1947, the transistor action was demonstrated by Brattain and Moore 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, however, was delayed 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 above is an abbreviated account of the events that led to the invention of the transistor. I believe it to be essentially correct. It is consistent with a memorandum written in December 1949 by W. S. Gorton, an assistant to the director of research of Bell Labs [2, p. 97]. Gorton had been asked by his management, "while the memories were reasonably clear, to write an account of the thinking, work, and events which resulted in the transistor." Gorton's memorandum is probably the most authentic summary in existence. In preparing his account, Gorton addressed the question of giving full credit to all who had contributed. Gorton's memorandum includes the names of 12 people who had taken a substantial part in the work. Those names all appear in the foregoing account.

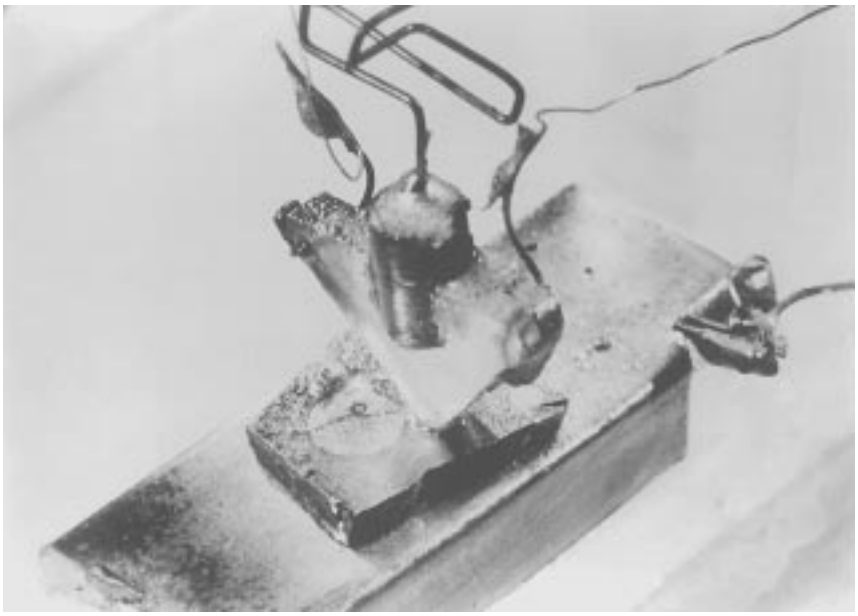
With the invention of the point-contact transistor—the gold foil having been replaced by two closely spaced point contacts—and the demonstration of transistor action, the door had been opened to a whole new era of electronics. But the process of inventing the transistor still had a long way to go. Transistor 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 invention 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 [4] (Fig. 3). But this so far was just theory.

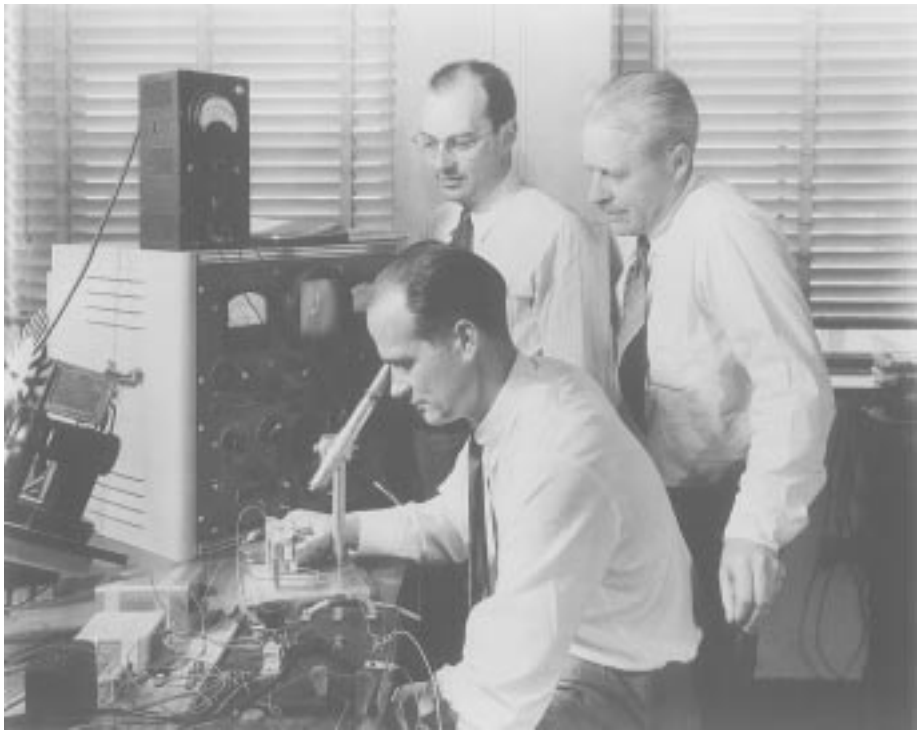
One month later, in February 1948, J. Shive carried out a critical experiment [5] He applied two phosphor-bronze contacts to the opposite sides of a 0.01-cm-thick slice of germanium. With this arrangement, he observed transistor action from one contact to the other with substantial power gain. The length of the surface path around the semiconductor slice effectively ruled out a surface effect. The action had to take place through the semiconductor body.

Shive reported the results of his experiment at a group meeting in February. These meetings were held regularly among a group of about a dozen people who were studying the transistor effect. The group included members of Shockley's organization but was extended to a select few people outside the Shockley group, including Shive. After Shive's presentation, Shockley went to the blackboard and described his recently developed theory of the junction transistor. The Shockley theory nicely explained Shive's observations. Thus, it was clear that while the point-contact transistor may have exhibited some surface effects, bulk propagation was also surely taking place and was probably the dominant effect.

There was one additional feature in Shockley's analysis of the junction transistor that should be mentioned at this point. He proposed a structure that had a fourth region by the addition of an extra junction at the collector. This was sometimes called a "hook collector" [6]. Shockley noted that this would lead to a current gain of greater than unity. This had indeed been observed in some point-contact transistors and was probably due to the creation of a hook collector in the process of "forming" the phosphor bronze collector contact. It was later discovered that silicon p-n-p-n diode had a bistable characteristic, behaving like a reverse-biased junction in one state and a forward-biased junction



(a)



(b)

Fig. 2. (a) The original transistor structure. (b) W. Shockley (seated), J. Bardeen (left), and W. Brattain (right) photographed in 1948.

in the second state [7]. This was due to silicon transistors' having current gains that increase with current [8]. The device could be switched from the first to the second state by exceeding the breakdown voltage, or by a pulse of light, or by injecting current through a contact to one of the base regions.

The p-n-p-n diode later was extensively studied in my development group in the mid-1950's in the hope of producing a cross-point switch for telephone switching machines. The device turned out to be difficult to control and at that time could not compete economically with a ferreed relay. Much later, it did find use in the switching matrix of some customer premises switches. The properties of the p-n-p-n triode were also explored, and it soon became known as the "thyristor." This device eventually was widely used in power conditioning applications. Since p-n-p-n devices did not play a significant role in the evolution of the transistor itself, I will not discuss them further.

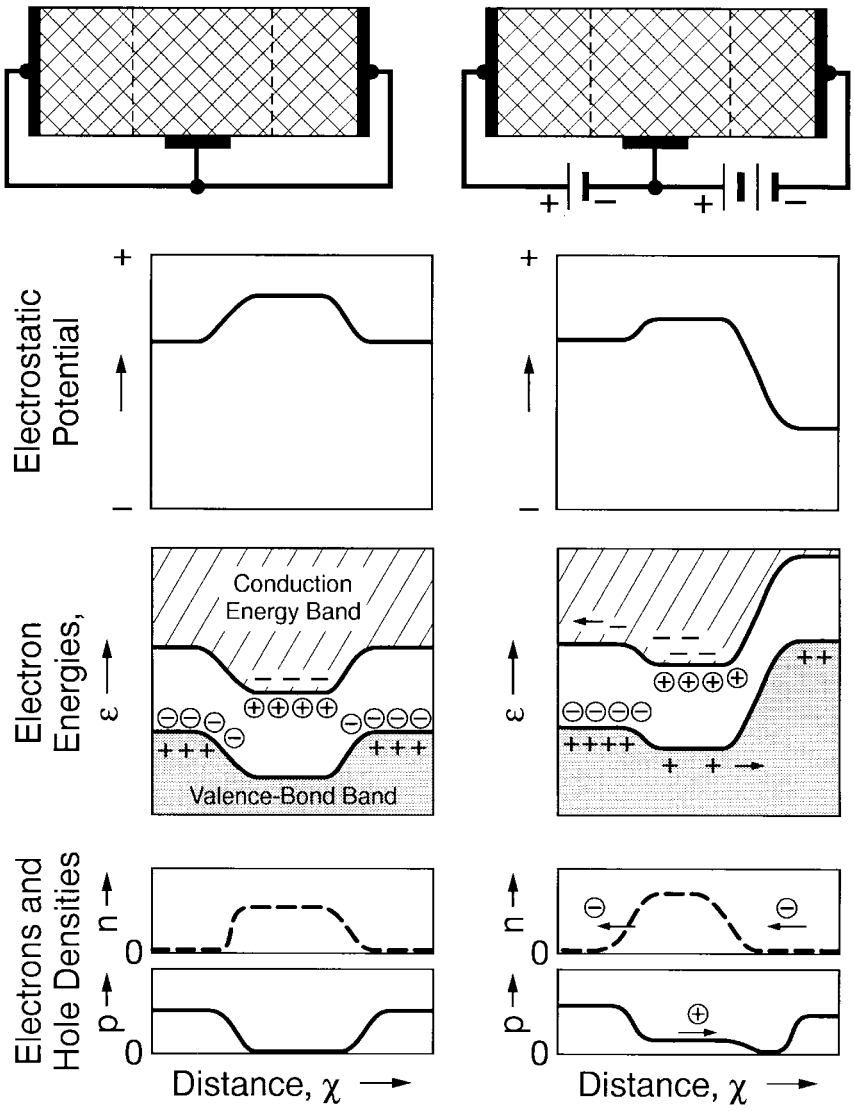


Fig. 3. Schematic of the junction transistor [12].

Returning to the transistor story, the next major advance was made in 1948. G. K. Teal and J. B. Little succeeded in growing a single crystal of germanium by slowly pulling a seed crystal from a melt of high-purity germanium [9]. 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, Suhl, and Shockley confirmed the behavior of both types of minority carriers and yielded measurements on injection efficiency, mobility, diffusion coefficients, and lifetime [10]. These results showed that useful devices could be made according to Shockley's junction transistor theory. All that remained was to make one. (Fig. 4).

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 of Shockley, M. Sparks, and Teal succeeded in growing a crystal containing a thin region of p-type embedded in n-type material. The crystal was cut into n-p-n rods and contacts applied. The electrical properties of the resulting devices were largely consistent with the Shockley theory [11]. Transistor electronics now had a solid foundation (Fig. 5).

There was one other event that completed this phase of the transistor saga. That was the publication in 1950 of Shockley's book *Electrons and Holes in Semiconductors* [12]. This was an exquisite account of the current understanding of semiconductors and transistors. It makes enlightening reading today after almost 50 years. In the 1950's, it provided an excellent means, and almost the only means, for scientists and engineers to get up to speed on a rapidly developing technology. It was required reading for those entering the business in its early days, and particularly so if you found yourself reporting to its author, as I did in March 1952.

So, in a period of only five years from 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.

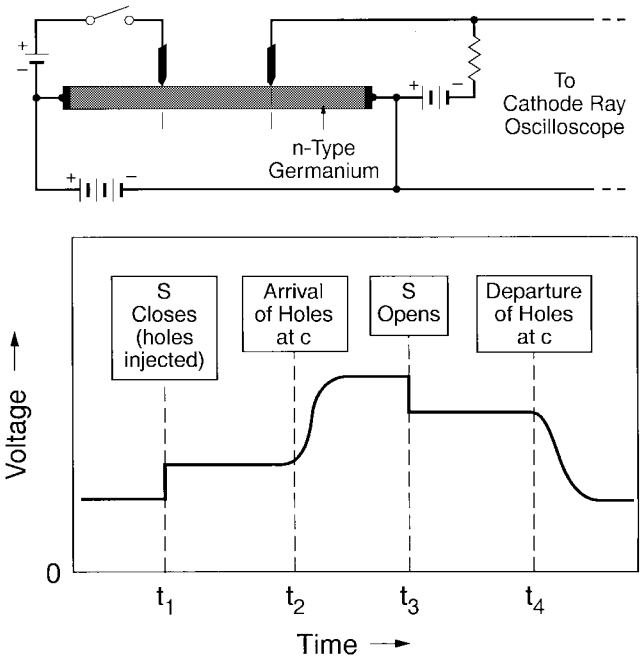


Fig. 4. Schematic of experiment to measure the properties of minority carriers [12].

The next phase would focus on solving development and engineering issues so that a brilliant invention could be converted into an important innovation. But before leaving the scientific phase, it is instructive to inquire as to why it was so successful and why it could occur at such a startling pace.

The vision of Kelly was no doubt a major contributing factor. R. W. Galvin, chairman of the executive committee of Motorola and chairman of SEMATECH, once said that the successful leaders of industry are those who can "anticipate and commit"[1] [37]. Kelly anticipated the need for a replacement for the vacuum tube and the relay and the possibility that a solid-state amplifying device could be made to do the job. He was also willing to commit the necessary quantity and quality of resources to the task. Kelly provides a good example of the kind of contribution that middle management can make. In today's enthusiasm for the reengineering of organizations, there are those who suggest that the only role of middle management is in communications up and down the line. They forget that management somewhere must develop the vision—must anticipate and commit.

Kelly's goal provided a compelling mission for the group. One major factor that distinguishes first-rate industrial research from academic research is that the first is mission oriented and the second is discipline oriented. The mission of creating a semiconductor amplifier focused the effort, created enthusiasm, and stimulated cooperation.

[1] "Total customer satisfaction starts with anticipating every 'what' that the customer wants and/or a competitor might be preparing to offer and committing effectively to fulfilling the timely needs of the customer. Effective, timely commitment is all too often the missing link to success because of our unduly rationalizing insufficient funds, people, etc. The courageous leader finds the means to commit to the achievement of the anticipated" [37].

The science was right. Enough was known to understand qualitatively what might work and what might not. Enough was known to encourage the group to insist on picking materials that were capable of being produced with extremely high purity and exquisite crystalline perfection. They were able to focus on mechanisms that qualitatively had a chance to work. Last, when discoveries were made, enough science existed for the theory to be taken the next step. This focus on basic understanding created a legacy that remains in the semiconductor industry today.

The trio that formed the core of the semiconductor group was an outstanding assembly of complementary talents. Each one made a unique and necessary contribution to the end result. When additional expertise was required on the project, it was enthusiastically available from people working just a few laboratories away. And that support was supplied promptly according to the needs of the mission and was not barred or delayed by any bureaucratic requirements or protocols. This was a benefit of the highly creative environment that had been built within Bell Labs. Sparks, who helped fabricate the first grown junction transistor, describes the atmosphere at Bell Labs in the following words:

> Bell Labs was a marvelous institution to give birth to and nurture a discovery such as the transistor. The labs existed as a separate company, wholly owned and supported by AT&T, the holding company of the Bell System, and Western Electric Company, its manufacturing arm. The Bell System was a nationwide end-to-end telecommunications conglomerate, regulated by various government commissions. The Research Department was about 10% of Bell Labs, but it was the part that most distinguished it among industrial laboratories of that time. Most research in 1947 was located in a new building complex in Murray Hill, NJ. There was a tradition of openness and cooperation throughout the laboratory. The culture was an unusual blend of business and academic attitudes. Much of the work was basic research in fields relevant to telecommunications, which included most of the physical sciences. A fundamental approach to the technical work was encouraged and supported by management. In the preface to his classic book, *Electrons and Holes in Semiconductors,* Shockley wrote:
>
> > The endeavor to probe deeply into the logical consequences of the fundamental theory, to reduce these consequences to pictorial terms and to find experimental counterparts to the theoretical concepts is in keeping with the philosophy of research at Bell Telephone Laboratories. The invention of the transistor occurred in connection with a research program based on this philosophy.

R. Gibney was a physical chemist who had worked with Bardeen and Brattain on some of their studies of electronic surface states on germanium.

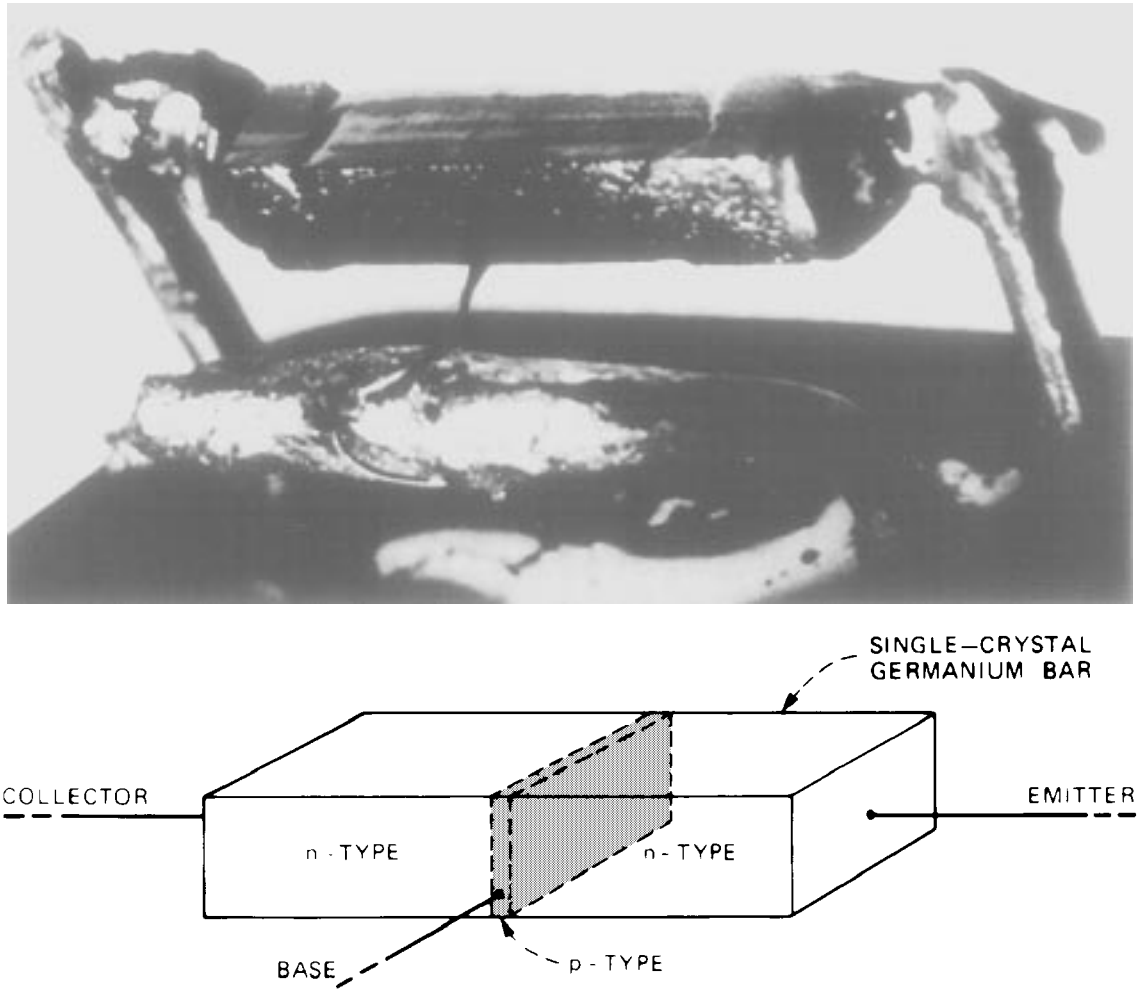


Fig. 5. The first grown junction transistor.

Shortly after Bardeen and Brattain made their historic discovery of the transistor, Gibney left Bell Labs for Los Alamos, NM, hoping to improve his wife's health. I knew Shockley, and my technical background was similar to Gibney's. Shockley invited me to join his semiconductor group. The next few years were an exciting and rewarding time for me. I worked directly with Shockley. His physical insight and ability to cut quickly to the essential factors of any technical consideration were phenomenal. He had what seemed to be an intuitive understanding of the physics of semiconductors. However, he used conventional concepts and analysis to develop his ideas. He hoped his unusual talents might be teachable and later spent considerable effort in collaboration with the Bell Labs Education Department to this end—alas, to no positive result.

There were frequent informal meetings, usually in small groups of half a dozen or so, to discuss results and plans. Shockley presided, inevitably at a chalkboard. Everyone was encouraged to participate. I particularly enjoyed Brattain's comments, delivered with his somewhat gravelly voice and salty vocabulary. Bardeen was quiet and polite; his thoughts were always well organized. He was readily available to those of us who sought his help. He was enormously respected. Dick Haynes was our expert on minority carrier lifetimes. Lifetime measurement techniques and interpretation of the roles of various crystalline imperfections in determining lifetimes were critical topics of the day.

The group remained small for a couple of years, but it was well represented in appropriate disciplines and talents. The mission orientation of the group encouraged cooperation. I learned more crystalography from Walter Bond and circuit theory from Moore in a few months than I had known from all previous exposures to the subjects. Gerald Pearson was a versatile and clever experimentalist with good communication skills. In addition to his own formidable contributions, he was up to date on problems and new results in the group and a valuable source of suggestions on procedures.

Eminent scientists from all over the world converged on Murray Hill to visit the group. They usually were hosted at larger and more formal

gatherings. Peter J. W. Debye was a consultant who visited on a regular basis.

Transistor-related activity in other parts of the laboratory sprang up and grew rapidly. Gordon Teal successfully adapted to germanium the Czochralski method for growing single crystals from a molten reservoir. The nearly perfect crystals were also greatly purified by the growth process. They quickly became widely used throughout the transistor project.

Shockley developed, from first principles, an elegant theory of the electronic properties of a p-n junction in a continuous crystal of germanium or silicon. His extension of p-n junction theory predicted that two parallel junctions very close together in a crystal and with alternating conductivity types such as n-p-n would constitute a new transistor structure. The transistor action would occur entirely within the interior of the crystal.

I was working with germanium p-n junctions, mostly cut from cast ingots, where junctions occasionally formed during solidification. My interest focused on the current–voltage characteristics of the junctions, which were good rectifiers by the standards of the day. There were different results from sample to sample, however, and agreement with theory was not very good. A crystal growing apparatus, based on Teal's work, was built for my use. A design feature provided for the controlled addition of donors or acceptors to the melt during crystal growth, allowing the formation of a p-n junction within the continuous crystal. Junctions made in this way had rectification properties in excellent agreement with Shockley's theory. We were extremely excited with this achievement and proceeded quickly to attempt to make a junction transistor. A critical requirement for operation of a junction transistor is the injection of minority carriers across a forward-biased p-n junction. After some refinements in the controls, crystals with an n-p-n section were produced with the necessary small dimensions. Specimens cut from the section were indeed transistors with operating characteristics as predicted.

My work during those few years gave me an immense feeling of accomplishment. Grown junctions in crystals were soon superceded as a technique for the practical manufacture of transistors. However, it settled the basic physics of transistor action and formed a solid foundation for a phenomenal industry to evolve based on Shockley's theoretical insight. [38]

Last, there was an element of luck, as there usually must be in an exploration of this kind. Pure crystals could indeed be grown, and the minority carrier properties within them were more than adequate to support transistor action. The very first crystal to be grown had a minority lifetime of 100 μs, more than adequate to support devices that needed to operate at frequencies above 10 megacycles. This was fortunate for the technical community, and indeed for society at large. Nature need not have been so generous.

III. The Development and Engineering Phase

Having invented the transistor, the challenge was then to find ways to design a product that could be manufactured and that could sustain a market—a traditional development and engineering task. This is not to suggest that the physicists, the chemists, and the metallurgists ceased to be involved—far from it. But clearly, more developers and engineers became involved, and the goals were predominantly engineering goals. This phase took the industry approximately eight years, during which many challenging problems were addressed and solved. Whereas the scientific phase had been dominated by Bell Labs, there were now other companies in the business, and they also made major innovations. These contributions were rapidly shared within the industry despite the pressures of increasing competition.

What follows is not an attempt to create a chronological history of this era but rather to select and describe some of the major hurdles that had to be overcome and the major breakthroughs that were made. There are many events that made "a" difference. I will focus here on those that made "the" difference.

A. The Early Manufacturing Problems

In early 1951, there were two transistor structures that were proven to work, but neither of them was suitable for large-scale manufacture. The point-contact transistor had all the frailties of its cat's whisker heritage. It was difficult to make and its electrical characteristics were far from ideal, very variable, hard to control, and inherently unstable. Point-contact transistors were, nevertheless, manufactured for ten years, starting in 1951, by the Western Electric division of AT&T. They found application in telephone oscillators, hearing aids, an automatic telephone routing device, and the first airborne digital computer. They were never popular with the manufacturing engineer, however, nor with the circuit designer.

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. Crystals were grown using a precise doping procedure to create a single thin layer of base material embedded in emitter and collector material of the opposite type. Only one "slice" of base material could be grown in one crystal.[2] Rods containing the base material were cut from the crystal, and the base layer in each rod was located and contacted. This was a labor-intensive process and not conducive to automation.

[2] In 1953 at General Electric (GE), R. N. Hall described an ingenious technique known as "rate-growing with meltback" by which multiple p-n junctions could be grown into a germanium crystal.

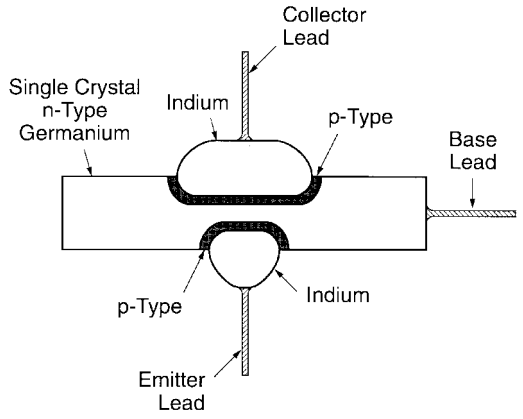
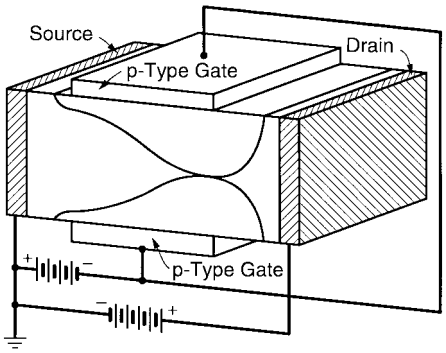


Fig. 6. Schematic of the alloy transistor.

The grown junction transistor was manufactured starting in 1952. In the same year, J. E. Saby at GE announced the development of the alloy junction transistor [13]. The original version was made by alloying dots of indium, an acceptor material, 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 (Fig. 6). One drawback was that precise control of dimensions and alloying temperatures were required to create thin enough base layers for high-frequency performance.

The emergence of the alloy junction process had an interesting byproduct. It was used to fabricate, at last, a functioning field-effect transistor. In 1951, Shockley had reinvented the field-effect transistor, but this time as a junction device. He proposed to use the space-charge region of a reverse-biased junction to constrict the area of a semiconductor material in which majority carriers could flow. His analysis showed that such a device could produce power gain. Since the action of the device involved only majority carriers, he called it a "unipolar" transistor to distinguish it from the "bipolar" junction transistor, which involved both minority and majority carriers. Incidentally, one of Shockley's drawings showed p-n junctions formed in a grid-like array perpendicular to the direction of the flow of majority carriers. The desire to emulate the vacuum tube died slowly.

In 1952, Shockley asked G. C. Dacey and me to try to build a unipolar field-effect transistor. We chose to avoid the unneeded complication of building a grid structure and formed the "gate" region by alloying indium over the



(a)

(b)

Fig. 7. (a) Schematic of the junction field-effect transistor. (b) G. C. Dacey and I. M. Ross (seated) testing a field-effect transistor.

surfaces of an n-type filament [14]. This structure behaved in all respects according to the Shockley theory. At this time, however, it had no significant theoretical performance advantage over the bipolar transistor and was no easier to fabricate. Thus, the field-effect theory was validated but the field-effect transistor went back into obscurity—at least for a while (Fig. 7).

B. 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 the higher energy gap of silicon—1.1 eV compared to 0.67 eV for germanium. In germanium at room temperature, a substantial number of electrons could penetrate the energy gap and enter the conduction band. The reverse current in germanium p-n junctions was therefore substantial and increased rapidly with temperature. The reverse current in silicon was orders of magnitude smaller. Hence, a reverse-biased silicon junction better approximated an open circuit and made it much more suitable for relay-type operations as found in logic and switching applications. Silicon junctions also retained their

properties to much higher temperatures than germanium, implying a capacity to handle higher power.

There were some minor disadvantages anticipated for silicon. The higher energy gap also meant that the forward bias needed to create a significant flow of current was higher by the ratio of the energy gaps. This would lead to a minimum operating voltage that would be higher in silicon circuits and require more operating power. Silicon also has a lower minority carrier mobility, by about a factor of three. This would reduce the velocity of carriers and result in a lower limit on speed of operation. In practice, this drawback could be eased by a reduction in device dimensions, which was not a serious problem given the miniaturization technology that was eventually developed for silicon.

The most serious problem with silicon was that critical chemical and metallurgical processes all took place at substantially higher temperatures. For example, the melting point of silicon was 1415°C compared to 937°C for germanium. Silicon was also more chemically reactive than germanium. The uniform increase in processing temperatures severely aggravated the problem of obtaining the required material purity and crystal perfection. Materials that could be used to contain germanium during crystal growth would either fail completely at silicon growth temperatures or were highly contaminating. These were, of course, the reasons that germanium had been the favored material for the early transistor explorations.

Given the relative advantages of silicon, it was considered important to develop a silicon transistor capability if only to complement the performance of germanium. In 1952, Teal and E. Buehler managed to produce silicon crystals by pulling from a melt contained in a silica crucible [15]. The crystals were heavily contaminated, mainly by oxygen from the silica. Theuerer solved the contamination problem by a modification of an invention made by B. Pfann in 1951 [16]. Pfann had developed a zone refining technique that could purify germanium crystals to a level of one part in 10^{10}. The Pfann process involved multiple passes of molten zones from end to end of germanium ingots held in a graphite boat. With each pass, impurities were moved along the ingot in the molten zone, leaving ever purer germanium behind. But no suitable boat material could be found for molten silicon. The critical breakthrough came in 1953 with the development by Theuerer of the floating zone method [2, p. 582]. He was able, in a vertical rod of silicon, to create a zone of molten material contained only by surface tension. By moving the rod relative to the heating element, the molten zone could be moved from end to end in the crystal. Thus, the zone refining technique could be used for silicon and resulted in crystals of purity comparable to the best obtained in germanium (Fig. 8).

In 1954, Teal, who by that time had moved to Texas Instruments (TI), made the first manufacturable silicon transistor using the grown junction method [17]. All the pieces were then in place for silicon devices to assume a major role. The initial belief was that silicon would tend to dominate the logic and switching applications while germanium would retain the high-frequency linear applications. As technology developed, the relative ease of fabrication of silicon devices made them ever more dominant, while germanium tended to occupy special niches—a situation that still prevails.

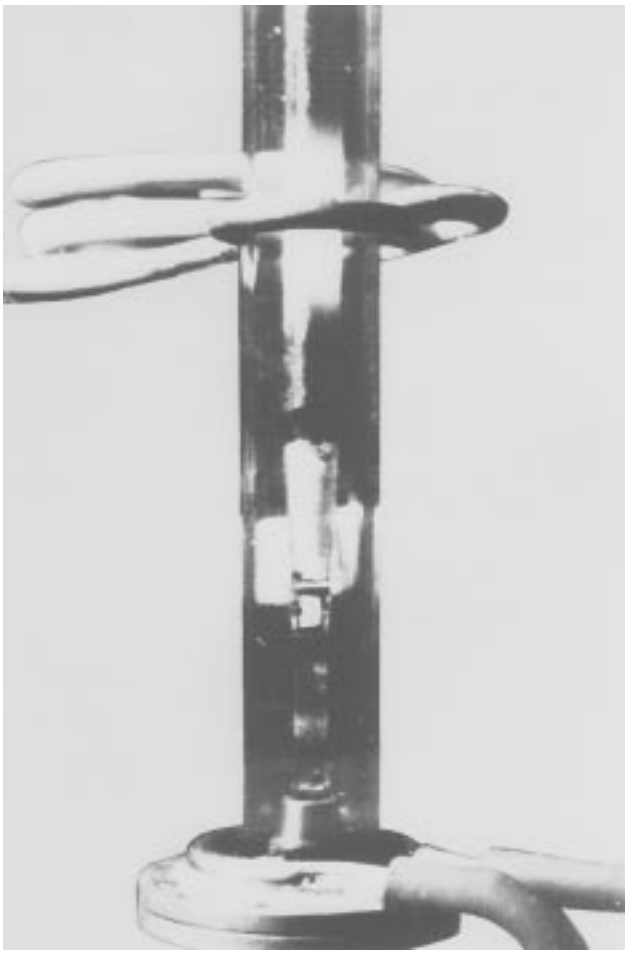


Fig. 8. Early floating zone apparatus.

C. The Bob Wallace Revelation

Having overcome the hurdle of being able to make transistors with some degree of reproducibility, a major goal was to replace the vacuum tube in as many applications as possible. This was not as simple as it first appeared. Transistors were easiest to make in small sizes, which inherently led to limited power-handling capability. Early transistors had lower frequency performance than most tubes. However, increased frequency response called for smaller, not larger, devices. In seeking higher power at higher frequencies, we seemed to be bucking nature. But that is the challenge of engineering, a challenge we would strive to meet.

One day, I was in a small meeting at Bell Labs with a colleague, R. Wallace. Wallace had been involved in the very early days of the transistor. He had in fact participated in the electrical characterization of the first grown junction transistor. He was a thoughtful and very creative person. As one illustration, later in his career, he invented the

phased-array microphone, which greatly reduced the sense of hollowness in the reception from a speaker phone.

In the meeting on that day, we were, as was frequently the case, discussing our problems in emulating the vacuum tube. Wallace suddenly said:

> Gentlemen, you've got it all wrong! The advantage of the transistor is that it is inherently a small size and low power device. This means that you can pack a large number of them in a small space without excessive heat generation and achieve low propagation delays. And that's what we need for logic applications. The significance of the transistor is not that it can replace the tube but that it can do things that the vacuum tube could never do!

And this was a revelation to us all. We realized that in chasing the vacuum tube, we had the wrong emphasis.

I tell this story because that is the way this important insight came to the people in that room. I know that we talked to others and the word spread rapidly. I am also sure that the same idea occurred independently to other people in other organizations at about that time. The net result was that the semiconductor community began to relax about replacing the tube and focused on developing the transistor in its own right. The transistor did eventually replace the tube in all but a few special applications, the magnetron being one outstanding example. But it took decades. In the meantime, semiconductor technology opened up important new fields that the tube could never have supported.

There is a lesson in this story. Having the clear goal of an application for an invention is a powerful stimulus for innovation. But frequently, the original application turns out not to be the most important. This was true in the case of the fax machine, the computer, and, most recently, the Internet. All of these found their major applications beyond the purpose for which they were originally devised.

The important point is that when a major breakthrough occurs, its further development should not be restricted to the originally intended applications. We must encourage the Wallaces of this world.

D. The Speed Problem—Controlling the Depth Dimension

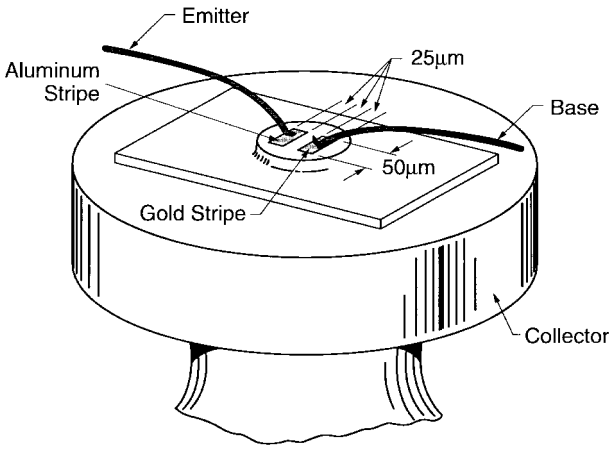
The fundamental determinant of the frequency response of a junction transistor was the transit time of minority carriers across the base region and therefore the thickness of the base layer. A base width of about 10 μ was needed to yield a frequency response approaching 10 MHz. This resulted in a major limitation in the performance of the alloy transistor. In this device, the base thickness equaled the thickness of the semiconductor wafer minus the penetration of the alloyed material. While the wafer had to be thick enough to be mechanically rugged, the base needed to be very thin, no more than 10 μ and preferably much less. We thus had the classical problem of controlling the small difference of large numbers. Frequency response was further limited by the minimum practical area of the alloy dots. In practice, alloy transistors were manufactured with bases as thin as 10 μ. Although this was quite a feat of manufacturing engineering, the devices were still only able to function up to a few megahertz. Realistically, performance up to a few gigahertz was needed to support a full range of electronic applications.

It was recognized that dopants could be introduced to very shallow depths by diffusion from the semiconductor surface. In 1952, C. S. Fuller published studies of diffusion of donors and acceptors in germanium [18]. The method involved surrounding the semiconductor with a vapor containing the desired dopants and raising the temperature to drive them into the surface. The temperature could be chosen to create diffusion rates that would provide precise control of the depth of penetration. The combination of temperature and concentration of dopant in the vapor controlled the density of dopant at the semiconductor surface. Eventually, the diffusion process yielded precise control of surface concentration over a range of 10 000 to 1, and control of the number of atoms introduced over a range of 100 000 to 1. Most important, the depths of diffused layers were nicely controllable in the range from 20 μ to a fraction of a micrometer—just what was needed.

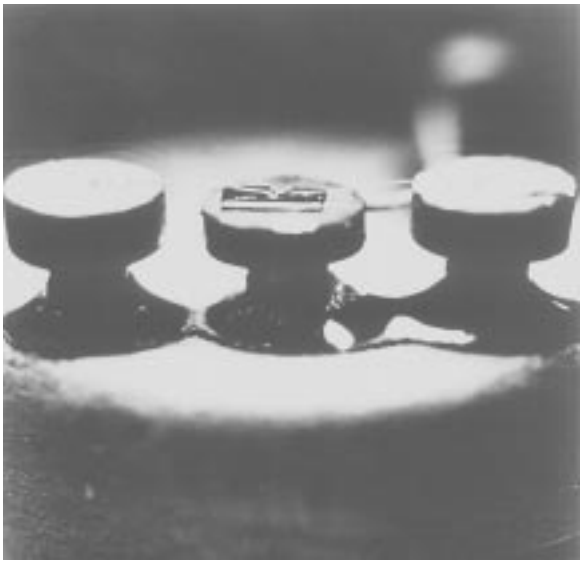
In 1954, C. A. Lee made first the germanium diffused transistor [19]. He diffused a base layer of arsenic to a depth of 1.5 μm and then created an emitter region, 25 by 50 μm, by alloying aluminum to a depth of 0.5 μm. The area of the collector junction was determined by etching away the base material except in a small region containing the emitter. The active device was thus contained in a "mesa" region. The resulting base thickness was about 1.0 μm, a factor of ten less than the base regions of the fastest alloy transistors, and promised a factor of 100 improvement in frequency performance. Indeed, this first diffused p-n-p transistor had a cutoff frequency of 500 MHz (Fig. 9). A year later, the first diffused silicon transistor was made and had a frequency cutoff at 120 MHz [20].

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 led to significant series resistance in the collector, and that, combined with the capacitance of the collector junction, limited the frequency response. One could conceive of cunning fabrication techniques to minimize this effect, but the real problem arose from a basic defect in the junction transistor concept. There was a built-in design conflict. The base region had to be thin to reduce transit time. It also needed to be highly doped to minimize the base resistance. Minimizing series resistance in the collector region required that it too should be highly doped. But with the collector junction formed between two highly doped regions, its capacitance would be high and its breakdown voltage low. One could juggle the resistivities and widths of regions to trade off among these factors but could not minimize them all. There just were not enough design degrees of freedom.

The theoretical solution to this problem had been proposed by J. Early in 1952 [21]. His solution involved the addition of a layer of intrinsic, or very high resistivity,



(a)



(b)

Fig. 9. (a) Schematic of the mesa transistor. (b) A mesa transistor.

material at the collector junction to create a p-n-i-p structure. In reverse bias, space charge would penetrate the total width of the "i" region. The width of the "i" region would then determine both the capacitance and the breakdown voltage of the collector independent of the resistivity of the adjoining "n" and "p" regions, which could thus be heavily doped. Early's device gave that extra degree of design freedom. His analysis led to the conclusion that "oscillations as high as 3000 MHz may be possible." Not only did Early invent the theoretical solution to a vexing design problem, he had the distinction of being the only person other than Shockley to propose a basically new transistor structure.

The theory was fine but the device was difficult to make. Using the best techniques available in 1954, p-n-i-p transistors were made with a cutoff frequency of about 95 MHz [22]. This performance, although far from the desired optimum, agreed well with the theoretically predicted performance for the specific device. The essence of the problem was that it was just too difficult to create a thin, controlled layer of intrinsic material in the body of a semiconductor wafer.

The Early structure represented the optimum design to achieve optimum performance. Fabrication difficulties could, of course, be eased somewhat by backing away from the optimum structure. High-performance devices could still be made if the high-resistivity region was not strictly intrinsic. It could be moderately doped p-type and still provide lower capacitance and higher voltage. Good results could be obtained even if the space charge did not totally penetrate the lightly doped p-type region, provided the region was thin enough to avoid substantial series resistance. The basic problem was that all high-performance devices called for a relatively thin layer of lightest doped material in the structure to be in the middle of the structure. The diffusion process and the alloy process naturally led to decreasing doping with distance from the surface. This made it necessary to diffuse or alloy from both sides of the wafer, leading again to the problem of controlling the small difference of large numbers.

The semiconductor community struggled with this problem for more than five years with much ingenuity and little success. We needed an extra degree of freedom in the fabrication process. The eventual solution was to add a totally different process, that of growing a lightly doped layer of single-crystal silicon on a substrate of a heavily doped single crystal while propagating the crystal structure of the substrate—a process called epitaxial growth. It was recognized in early 1960 that this indeed would be a solution to the problem and that such a layer could be grown. At that time, Theuerer was growing single-crystal silicon on silicon rods by the thermal deposition of $SiCl_4$. Some members of my development group and I visited Theuerer and asked if he could grow a thin, lightly doped n-layer on the surface of a heavily doped n-type crystal. He said he could, and in a day or so, we had our crystal. The team then produced a base and emitter layer in the lightly doped material. The resulting transistor behaved as we had hoped.

The next step was to use Theuerer's chemistry in our own lab to grow a lightly doped epitaxial layer on a heavily doped wafer and to fabricate a transistor in the epitaxial material. That process was also made to work, and the transistor behaved according to theory. The results were published in June 1960 by Theuerer, J. J. Kleimack, H. H. Loar, and H. Christensen [23]. The challenge of producing the desired doping profile across a wafer had finally been solved with the addition of the epitaxial process to those of diffusion and alloying. We now had the fabrication degrees of freedom that we needed.

E. Oxide Masking and Photolithography—Controlling the Surface Dimensions

In 1955, C. J. Frosch and L. Derick made a very important observation. They had been studying the pitting of the surface of silicon wafers during the diffusion process and its dependence on the presence of oxygen. They found that the pitting problem could be eliminated by assuring that

oxidation of silicon during the diffusion process favored the formation of silicon dioxide rather than silicon monoxide. They achieved this by the addition of water vapor. They further discovered that a few-thousand-angstrom layer of silicon dioxide grown on the surface 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 [24]. Somewhat later, J. Andrus and W. L. Bond showed that certain photoresists deposited on the oxide surface would prevent etching of the oxide [25]. Hence, optical exposure of the resist by projection or contact masks could be used to create precise window patterns in the oxide and in turn provide precise control of areas in which diffusion would occur.

Thus, four people in the course of a few weeks had invented the complete process of oxide masking of diffusion and the application of photolithography to the precise control of the geometry of diffused regions. This was a natural batch process that promised control of patterns of diffusion to the precision of the wavelength of light. Over the years, these same principles have been laboriously and ingeniously developed to the point that junction areas can be controlled to a fraction of a micrometer. This complements—indeed, roughly equals—the precision of the depth control of junctions diffused into the silicon surface. We thus had the means eventually to control the fabrication of silicon devices in three dimensions to the precision of a fraction of a micrometer.

The oxide masking development is another illustration of the innovation process being aided by the generosity of nature. It is quite remarkable that this powerful capability to determine the "horizontal" dimensions of silicon devices would be conveniently provided by the oxide of silicon itself. This piece of fortune assured the future of silicon. It also sealed the fate of germanium. No material was found that would provide diffusion masking on germanium. The oxide of germanium is a fragile material. It is even soluble in water. With such a large fabrication advantage, silicon steadily dominated the semiconductor applications. Germanium became the niche material for specialty devices that rely critically on some special property.

One further observation before leaving this subject: many people these days complain that they no longer see major breakthroughs emerging from industrial labs. The development of masking and photolithography could be cited as one such prior contribution. But that development took only a few people and a few months to go from discovery and invention to first useful application[3] [26]. Photolithography is today running into some fundamental limitations and eventually may have to be replaced by techniques using direct exposure by electron beams or by x-rays. These may well use ideas already generated in the labs of IBM and AT&T. But now, the step from invention to proven process needs investments of more than $200 million over a period of five to ten years. This is not the kind of investment that a single company can justify today. Nor could it afford its equivalent 40 years ago. Great ideas are still coming from industrial labs but, at least in the semiconductor industry, the ante has changed.

[3]The first application of these techniques was in the fabrication of a stepping transistor, which was developed in my group. Andrus made the oxide windows. The stepping transistor was designed to perform a counting function much like that of a shift register. Although the device worked, it turned out to be a wrong solution to an important problem.

F. The Reliability Problem

I joined Shockley's organization at Bell Labs Murray Hill in March 1952. When the transistor invention was announced in mid-1948, I was an undergraduate at Cambridge University in England aspiring to be an engineer. I was excited by the news, but had no idea that this was the beginning of a major technical revolution, nor did I have any concept of the influence it would have on my own career.

As an undergraduate at Cambridge in those days, there was no opportunity to specialize in a particular branch of engineering such as electrical, mechanical, or civil. Instead, students took courses in all major engineering topics. Indeed, Cambridge was so out of date, or perhaps so farsighted, that the department was not even called "Engineering" but used the title "Mechanical Sciences." Had I been permitted to specialize as an undergraduate, I would have chosen civil engineering and missed the electronics revolution altogether. As I got more deeply involved in engineering studies, I realized that my interests lay more in the fundamentals, more toward physics than structures, and I chose to stay at Cambridge to pursue the Ph.D. degree in electrical engineering but in an area that soon came to be known as electronics.

My thesis work involved measurements of low-frequency noise, mainly in radio tubes. I also did similar studies on semiconductors. As I was finishing my Ph.D. work, I was made an offer I could hardly resist. Shockley invited me to join his Bell Labs group and to work on semiconductor devices. So, in March 1952, I arrived in New Jersey with the idea of working at Bell Labs for a year. That was the beginning of a 40-year career.

When I arrived at Murray Hill, most of the world's knowledge of transistor technology resided in that building. Most of the original cast of contributors were still there. They were a fascinating group of people, wrapped up in the excitement of a very rapidly developing field. They were very kind and helpful to someone like me with a shiny new Ph.D. degree and very little knowledge of semiconductors.

In April 1952, a transistor-technology symposium was held for Western Electric licensees. The objective was to provide enough information to the 40 attendees to "enable qualified engineers to set up equipment, procedures, and methods for the manufacture of these products." The symposium took eight days, six at Murray Hill and two at the Western Electric facility in Allentown, PA. I was asked to organize a laboratory session in which the attendees, in small groups, could measure the characteristics of transistors. For most of them, this would be the first time they had actually seen a transistor.

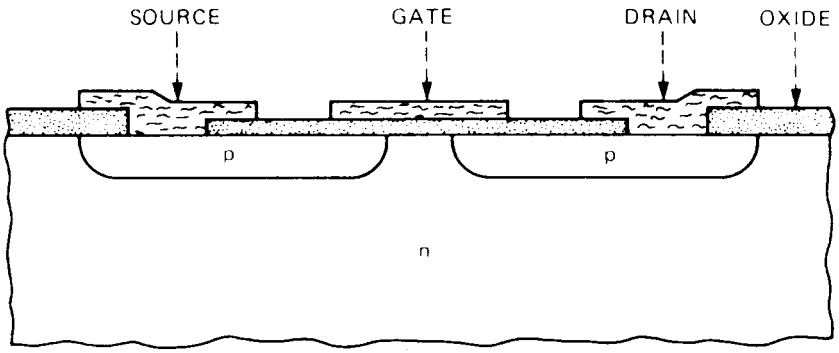


Fig. 10. Schematic of the MOS transistor.

I had arrived from England just a month earlier. I was used to a climate that made a gradual transition from winter to summer during a period called spring. To my surprise, New Jersey that year went from winter to summer in one day, and I suddenly found myself exposed to temperature and humidity levels that were totally new to me. That transition took place two days before the start of the symposium. Murray Hill at that time was not air conditioned. When I got to the lab that morning, I found that my transistors for the session had lost all their electrical characteristics—the CRT traces were flat.

I had discovered for myself what many people already knew. The transistor was sensitive to its environment and particularly to humidity. This was no time for investigation, nor for pride. The session took place, but with each transistor in a test tube containing desiccant and a wad of cotton wool. The tubes were sealed with rubber bungs. The students did not seem to care at all. But I did.

The lack of reliability of the transistor was a huge setback and embarrassment to the semiconductor community. The transistor had been lauded as a device with no failure mechanisms, with nothing to wear out. Instead, we had a severe reliability problem, and one that took almost 20 years to solve completely.

The immediate remedy was to seal the devices hermetically in packages using the metal-to-glass seals from vacuum-tube technology. This was a further blow to the pride of the semiconductor engineer. Nevertheless, using these techniques, point-contact transistors attained a reliability level of a million socket hours per failure, which compared favorably to tubes. 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 mesa transistors were eventually manufactured with failure rates less than ten per billion operating hours.

There also were ongoing systematic studies to try and understand the problem and find a more fundamental solution. At Bell Labs, Brattain continued his experimental work on surface states, as did Shive. M. M. Atalla had 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 states at the silicon surface 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 [27].

A byproduct of their investigations was the fabrication of another design of a field-effect device. Their method of exploring surface states involved the application of voltage to a metal contact overlaying an oxide layer grown on the silicon surface. The density of states at the silicon surface was sufficiently low that they were able to create a substantial inversion layer in the underlying silicon. This was the first operation of a metal–oxide–semiconductor (MOS) transistor, and it behaved according to theory (Fig. 10). Their results were published by Atalla and D. Kahng at the same conference at which the epitaxial transistor was described [28]. At this point, the performance of the MOS device could not, however, compete with bipolar devices. A major problem was a high-threshold voltage, which turned out to be caused by the high contact potential of the aluminum that they used for the gate. (This problem was eventually solved by the use of a silicon gate.) So, the field effect having been verified again, the concept had its moment of glory and returned to obscurity.

The final breakthrough in the solution of the reliability problem came with an invention made by J. A. Hoerni at Fairchild in late 1957 or early 1958. His idea was later reduced to practice and published in 1960 [29]. 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 thus intersected the silicon surface under the oxide layer, and Hoerni speculated that the oxide could protect the junction areas from contamination. It was indeed found that such junctions had acceptable characteristics without further treatment. This was a startling result, particularly for those who believed that a passivating oxide would need to be grown under meticulously clean conditions.

This was not the end of the story, but the Hoerni result put us on the right track. It was later found that not all

"diffusion" oxides gave adequate initial performance and that all were subject to degradation with time. The source of this degradation was identified, also at Fairchild, as alkali ions, mainly sodium, moving through the oxide in the electric fields near the biased junctions.

There followed extensive efforts throughout the industry to eliminate sodium from the oxide and prevent it from reentering. Although many important contributions were made and some continue to be made, the key advance was that made by J. V. Dalton in 1966 [30]. He demonstrated that an overcoating of silicon nitride would provide an effective seal against sodium ions. (Actually, silicon nitride proved such an effective seal that it even trapped hydrogen in the device structures, which leads to threshold-voltage instabilities. Silicon oxynitrides were then shown to still be a barrier for sodium but not for hydrogen. Most circuits today therefore use a plasma deposited overcoat of silicon oxynitrides as a seal over the entire circuit structure, except, of course, the contact pads.) Given this batch process to eliminate the ions from the oxide, the reliability problem was basically solved. Silicon devices 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.

It is indeed remarkably fortunate that passivation could be achieved with a simple surface coating and that the effective material would be silicon dioxide and not even in a very pure form. Again, nature need not have been so generous.

G. The Planar Transistor

In his 1960 paper, Hoerni also described the planar transistor. In this 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 emitter and base regions through windows in the oxide layer. It was noted that connection to the collector region could also be made on the top surface if that were desirable. The metal used for all contacts was aluminum, which Moore and R. N. Noyce had previously shown would make good contact to either n- or p-type silicon [31]. Moore had also shown that the aluminum could be extended over the oxide to form larger pads to ease connections to the chip (Fig. 11). Somewhat later, the epitaxial process was added to the planar transistor to minimize collector resistance.

This structure brought it all together. 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 crystalline perfection and purity more than adequate to the task. Critical dimensions in all three directions could, if necessary, be controlled 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.

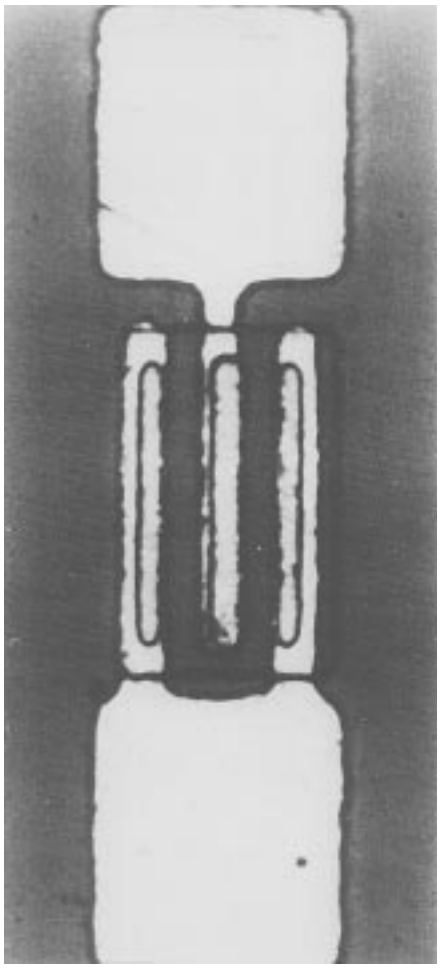


Fig. 11. The planar transistor.

And all this could be done with batch processing with the promise of high yield and low unit cost.

Some 13 years after its invention, the transistor now had a sound engineering foundation. This provided the base for the next giant step. The integrated circuit (IC) was invented in 1958 by J. S. Kilby at TI [32], with a major added contribution from Noyce at Fairchild [33].

H. The Integrated Circuit

The problem the integrated circuit was designed to solve had been vexing the semiconductor community for a number of years. Given the transistor's inherent small size, low power dissipation, and potential for high reliability, it had long been appreciated that the transistor should make it possible to build systems with thousands of active devices working together and operating at high speed. There were pressing applications for such systems in computers, telephone switches, and in several military projects.

Transistors in principle could be packed closely together without excessive problems of heat dissipation. In large systems, they needed to be closely spaced so that the signal propagation delays across the whole system did not become the factor limiting system speed. This all called for miniaturization of systems, not just components. There was, however, a major concern that assembly yield and

Fig. 12. The first integrated circuit. (Reprinted by permission of Texas Instruments.)

reliability in systems with thousands of components would be unacceptably low. It was further recognized that even if the components were perfect, a multitude of connections of the kind then in use could still be a weak link. This "tyranny of numbers" drove the desire to try and reduce the number of components and to simplify their interconnection.

It was Kilby, in 1958, who first demonstrated that it was possible to produce transistors, diodes, capacitors, and resistors in one piece of semiconductor and interconnect them to create functioning circuits. His early circuits had about ten components. Kilby used wire bonding to interconnect the components within the chip (Fig. 12). This would have made manufacturing very difficult and surely would have limited the number of components per chip. Kilby recognized this problem. In his patent, he suggested that suitable contacts could be made by deposition of a metal over a previously deposited layer of silicon dioxide. Later that year, and independent of the Kilby suggestion, Noyce proposed replacing the wires with the batch deposition of aluminum on a planar structure. As a consequence, Kilby and Noyce are jointly recognized as the inventors of the IC.

There was, at the time, a surprising degree of reluctance in the industry to pursue the IC approach. There had been considerable prior speculation about the benefits of fabricating several components on a single chip. As early as 1953, there had been proposals to fabricate more than one transistor in a single chip, but very few multiple transistor structures had been made. It was known, of course, that diodes, capacitors, and resistors could be made in semiconductors. But no one had had the vision to put it all together as Kilby did.

There were several objections to pursuing the idea of multiple components on a chip. The most serious concern was the expectation of very low yields and reliability. It was argued that if a single transistor could be made at only 20% yield, which was an acceptable yield for many years, or even at 90% yield, which was considered to be excellent for many years, the yield of a chip containing 100–1000 transistors would be minuscule. It was similarly argued that the reliability of a chip would approximate the reliability of a discrete transistor degraded by the power of the number of transistors. These arguments carried great weight. They assumed, of course, that yields and failure rates were governed by random events, which turned out not to be the case.

At the time of the invention of the IC, there were basically two other proposals for addressing the tyranny of numbers. The first was to assemble and interconnect tested, discrete components according to a uniform, size-reduced configuration. One ramification was the micromodule, which can be thought of as a stack of small circuit boards containing connected components. The boards, in turn, were interconnected by metal risers along the sides of the stack. This kind of disciplined assembly approach was the most actively pursued and was funded by the U.S. Army.

The second approach sought to reduce system complexity by seeking physical phenomena that could be used to perform functions in place of circuits. It was postulated that single, hopefully simple, structures could replace circuits with many components. The piezoelectric crystal resonator was cited as an example of a simple device that replaced a multicomponent circuit. The acoustic delay line was another example. The concept went under many names but was perhaps best known as "molecular electronics" in a project funded by the U.S. Air Force. The term "functional devices" was used at Bell Labs for a similar concept. This approach had its attractions but depended for success on some basic inventions, which did not occur fast enough to compete with other events. Perhaps the best example of a successful invention of this kind was the charge-coupled device. It found major applications but in a specialized field.

These other approaches were pursued for several years. They consumed considerable effort and created much heated debate. They persisted for at least five years beyond the IC invention. This, in retrospect, is quite surprising. After all, Kilby had demonstrated the fabrication of about ten components in one structure with every reason to go to higher numbers. This about matched the gain in simplicity that was speculated to result from molecular electronics. And Noyce had replaced the wiring complexity with a simple batch process. That should have eliminated much of the incentive for micromodules. But the yield and reliability problem for higher levels of integration was still an issue. Good engineers had been trained to be conservative.

Kilby and Noyce, however, by inventing and pursuing the IC concept, effectively made the choice on which microelectronics approach to take. It took hard work and several years to demonstrate that the feared problems hardly existed. As it turned out, neither yield nor reliability was dominated by random events. This was really the consequence of the batch nature of silicon processing. In the case of yield, there tended to be large areas of a silicon slice in which the yield was effectively 100% while the yield in the remaining areas approached zero. Thus, if the IC chip was small compared to the areas of "good" material, the yield would be substantially independent of the size of the chip and the number of components on the chip.

The yield and reliability bugaboo was the only critical hurdle, the only make-or-break issue, that had to be over-

come to permit the IC to proceed on its incredible journey. All the needed transistor principles and processes were in place. From then on, it took much ingenuity, much effort, and much investment to apply them and further refine them, but no major transistor breakthroughs were needed. These thoughts are well expressed in words written by Kilby in 1976 [34]. In discussing the buildup of IC production he said:

> It should be noted that one of the great strengths of the integrated circuit concept has always been that it could draw on the mainstream efforts of the semiconductor industry. It was not necessary to develop crystal growing or diffusion processes to build the first circuits, and new techniques such as epitaxy would be readily adapted to integrated circuit fabrication. Similarly, new devices such as MOS transistors and Shottky barrier diodes would be phased in as they became available. Even today, it is difficult to identify a process that is used only for integrated circuits.
>
> Another strength of the concept was that it could draw on existing circuit technology to produce a broad range of useful devices. . . .
>
> Because of the commonality with existing processes, integrated circuits moved rapidly into a production status. [34]

There was, however, one more critical and interesting choice to be made. That was the choice of transistor. The planar bipolar transistor was the favored device for general application. But, given the planar technology, the MOS field-effect transistor was finally a practical device. However, the theoretical performance limits of the MOS still did not match those of the bipolar transistor, which therefore remained the preferred device. At some point, it was recognized that the simplicity of the MOS structure in fact gave it the advantage for most IC uses. Being simpler, the MOS was naturally smaller, and in addition, at any given time was easier to fabricate at the leading edge of design rules. This meant that its performance deficiencies were minimized and, most important, that more MOS devices could be formed on a given chip. Thus, MOS had a distinct advantage in terms of function per chip as compared to function per device and gradually became the transistor of choice for most IC applications. In 1964, RCA was the first to introduce MOS technology into IC manufacture.

This decision was interesting because it went contrary to our engineering heritage to embrace the better performer. It was even more interesting because of its historical significance. The field-effect concept had existed since 1925. Throughout semiconductor history, it had appeared several times as a highly desirable concept. But it had always been a bridesmaid. Now, after just less than 40 years, it was to be given a very major role in the most important of all electronic technologies.

I. The Major Impact of Moore's Law

There is, I believe, one other notable IC phenomenon that derives from the inherent properties of the transistor's technology. That is captured in Moore's law. In 1965, G. Moore published in *Electronics Magazine* a semilog plot of the number of components on a silicon chip versus the date of first availability. The result was a straight line representing almost a doubling per year. This became known as Moore's law and has been updated ever since. In later years, the rate relaxed somewhat to a doubling every 18 months. That is the rate today, and it is expected to prevail for some time.

Moore did not merely use his plot to record the past . He also used it to anticipate the future. Given the nature of silicon fabrication processes and his understanding of the limits of the physics involved, he saw no reason that this progress should not continue. His colleagues in the semiconductor industry saw it the same way and were willing to commit resources to make it happen. This is another example of the "Galvin principle" of the need to anticipate and commit.

Moore's law, however, had a major influence beyond the semiconductor industry itself. It demanded a change in mindset of any engineer working in electronics or designing systems that incorporated electronic equipment. Good engineers have been trained to design within the limits of their understanding. Indeed, in many cases, they intentionally back away from the leading edge of their knowledge and apply a factor of safety, sometimes properly called a factor of ignorance. Moore's law required that the successful engineer would design a product today anticipating a factor-of-two improvement by the time the product could reach the market. This was not an easy cultural change to make.

A factor-of-two improvement in 18 months, sustained for several decades, is almost unique in human experience. Epidemics may be the only other example with which the population at large is familiar. There is, however, one other technology that obeys Moore's law, and that is fiber-optic communication. The measure of performance here is the product of the bit rate that can be transmitted over a fiber times the spacing of repeaters. This factor has been doubling each year for almost 20 years, and we expect that that rate can continue for at least a decade. It is an interesting coincidence that fiber-optics technology is based on the use of glass, our old friend silicon dioxide. It was also fortunate because the chemical processes used for silicon and silicon dioxide in semiconductor technology were substantially helpful in developing extremely pure glass and in controlling the introduction of desired additives.

The existence of the two technologies, microelectronics and photonics, each repeatedly doubling their performance in short intervals, underlies the explosion in the so-called information technologies. This is frequently, and in my view incorrectly, referred to as the information age. It is the two underlying technologies that are driving the progress, not information. If the analogy is with the Stone Age, the Bronze Age, or the Iron Age, then a better term would be the "silicon age." That is what it really is.

Returning to Moore's law for IC's, how long can this kind of performance improvement continue? What are

the estimates for the ultimate capability of conventional silicon technology? Recent analysis by the Semiconductor Industry Association (SIA) suggests that we will reach some fundamental physical limitations toward the year 2005, with about 0.1-μm minimum dimensions and around 5–10 billion components per chip. There are numerous factors that contribute to this estimate. Perhaps the most fundamental factor is that we are beginning to run out of silicon atoms. At 0.1-μm dimensions, components such as transistors will contain just a few hundred atoms along their linear dimensions and contain only about 1 million atoms altogether. Beyond this point, it is no longer realistic to think of these as being bulk materials in the normal sense.

How stable, how firm is this estimate of a 10-billion-component limit? I believe it is quite firm. In November 1983, I was invited to deliver the Mountbatten Lecture in London. I was asked to address the question of an ultimate limit to progress in silicon technology, and I predicted a fundamental limit at about 1 billion components on a 1-cm^2 chip with 0.1 minimum dimensions. My prediction was greeted with some skepticism and even a little ridicule. The important point is that in a period of 13 years, the estimated limit has increased by only a factor of ten. The limit of design rules is little changed. What has changed is the willingness to anticipate chips as large as 10 cm^2 and to commit to make them. However, I expect the 10-cm^2 limit to also prove to be robust.

Some people argue that Moore's law will still continue to apply beyond these limits—that there will be some new invention that will save the day. I have little doubt that there will be new inventions. But, if they are to provide a capability beyond the limits of silicon microelectronics, they will not involve mere extensions of the use of the properties of bulk silicon. It is unlikely they will use silicon at all, and probably they will not depend on the bulk properties of any material. Some totally new approach will be needed with a totally new technology base. Thus, just as in the case of the transistor, the road from a brilliant invention to an effective innovation will be a long one, with many opportunities to fall by the wayside. To replace today's microelectronics will require surmounting many hurdles, creating many breakthroughs. All that will likely also take 15 or more years to accomplish.

Given that these projections are correct, the accustomed rapid advance in silicon chip technology will slow down around 2005 without the prospect of an immediate replacement. At that point, it will be necessary to change the culture of a large number of engineers who for 50 years have relied on the impact of Moore's law to solve many of their problems.

IV. What Made It Possible? What Made "the" Difference?

A. The Search for Understanding

From its beginning, the exploration of the transistor was accompanied by a search for sound scientific understanding. Kelly set this direction by establishing a research group, albeit a group with a mission that contemplated important practical applications. This concept was reflected in the composition of the group he formed and particularly the three principal members. They strongly believed in the importance of basic understanding and avoiding the empirical approach. Luck, of course, played some role. Their experiments, although designed to illuminate theoretical understanding, did on occasion have unexpected results that led in fortunate directions. But on these occasions, they devised further experiments to gain understanding of the new results. The major step from the discovery of the point-contact transistor to the development of the theory of the junction transistor was the direct result of the search for understanding.

In 1976, Shockley described events surrounding the invention of the transistor and discussed the emphasis given to the "respect for the scientific aspects of practical problems as an important creative principle in industrial research" [35]. He further wrote:

> In 1946, when the semiconductor research group focused on the basic science, leaders of some other research department groups urged me to emphasize practical semiconductor difficulties in the telephone plant. Our group was of one mind and we followed the wise course of working, not upon such practical but messy semiconductors as selenium, copper oxide, and nickel oxide, but instead on the best understood semiconductors of all—silicon and germanium—a "try the simplest cases" approach.
>
> For these best understood semiconductors, not all of the theoretical concepts developed . . . largely during World War II, had been experimentally verified. We elected to concentrate on the remaining gaps, among these being the recently proposed surface states. We felt that it was better to understand these two simplest elemental semiconductors in depth rather than to attempt to add piecemeal contributions to a variety of other materials.
>
> In assigning our highest priority to the primarily scientific aspects, we chose those related to the problems that blocked our approach to the long-range practical goal—the creation of a semiconductor amplifier, later to be called the "transistor."

This attribute has remained with the industry. Even during the many years when empirical solutions were applied to the reliability problem, the search for a basic solution continued and eventually won out.

B. A Willingness to Share Information

The semiconductor industry has operated with an unusual willingness to share information. This, of course, derived from the special nature of AT&T as the manager of the Bell System. AT&T was a private company with a government monopoly franchise. Its purpose was to benefit its stockholders by providing excellent telephone service to customers in the United States. In meeting customer needs,

the company developed and exploited technology, much of it built on the foundation of basic science. Many of the resulting products were manufactured in AT&T's subsidiary, Western Electric. However, it was also recognized that, given access to technology developed within AT&T, other companies would also produce products that could benefit AT&T's customers. This led to AT&T's practice of releasing research results at an early stage and licensing its patents to all comers in industry.

This practice was followed in the case of the invention and development of the transistor, as evidenced by the offering, in 1951 and 1952, of two symposia and one summer school. These were intended to transmit the latest transistor technology to academia, government, and industry. This established a semiconductor community of people who understood the benefits of sharing information. Of course, at first it was a one-way deal.

The willingness to share information prevailed in the industry. Many of the breakthrough results described in this paper were first published at the Solid State Research Conference. This conference was sponsored by the IEEE but attendance was by invitation only. The invitees, about 40 people, were chosen from those who were contributors and were willing to release results of their work at an early stage. Mere observers were not invited. Indeed, if attendees were found not to be contributing or were perceived to be holding back results, they were not invited again. These unusual conferences started in 1952 and continued into the early 1960's. The spirit of communication probably was further sustained by the institutional climate of Silicon Valley. Movement of key people among companies was so easy and occurred so often that open communication was inevitable and was indeed difficult to avoid.

The continuation of this spirit is evidenced today by the existence of two successful consortia in the semiconductor industry. The first, the Semiconductor Research Corporation, was formed in 1982 to manage a pool of resources to sponsor semiconductor research in academia. The second, SEMATECH, formed in 1988 in response to competition from overseas, initially shared effort on advanced processing but later played a vital role in strengthening the semiconductor materials and equipment industry in the United States. Last, the SIA has orchestrated a cooperative effort to maintain a technical road map for the industry. That SIA is one of the most effective industry associations is a tribute to the willingness of the industry to share information.

C. Leadership

The leaders of the semiconductor industry largely came up through the technology side of the business. That is not to deny that some of the successful leaders had no technical experience and fared well. But their colleagues and competitors mostly had "silicon under their fingernails." It is also true that some of the technical leaders did not make good businessmen or managers and fell by the wayside.

I believe that the prevalence of technical knowledge in the leadership contributed to the remarkable success of the industry. It took a deep understanding of a complex technology to appreciate what present limitations were and to anticipate what improvement was likely to occur next. It also took deep understanding and confidence to make the large commitments to the creation of the next generation of the technology. The proportion of revenues that was devoted to R&D was unusually high, surpassed perhaps only by the pharmaceutical industry. The capital investment was huge, increasing rapidly for each new factory that in turn had to be replaced in only a few years. The ability to anticipate and the willingness to commit were critical factors in the success of the industry.

D. The Role of Government

The U.S. government made some major contributions to the transistor story. Unfortunately, some of these contributions have on occasion been overstated. But that should not result in their being overlooked.

The transistor would not have been possible were it not for the scientific understanding that had come from basic research. In the United States, much of that work was funded by the government. Government funding of research continues to this day and has enabled the universities to continue to contribute to the technology of the industry and to be a source of excellent scientists and engineers for the industry.

During the war, effort on behalf of the military advanced the state of semiconductor technology, especially in microwave diodes. Much of that technology was, however, empirically based. Toward the end of the war, there was a recognition in the government that the field needed more research work, and some effort was funded in several universities. One good example was the work at Purdue University under Lark-Horovitz. He and his team did much to improve the purity of germanium and to understand its properties. Soon after the formation of their group, Shockley and Morgan visited the Purdue team to learn of their results.

In addition to the study of germanium material properties, Purdue had two projects investigating the performance of metal contacts to germanium. These were separate investigations by separate investigators, one working with the contacts forward biased and the other reverse biased. It has been speculated that had the two gotten together and, more important, had the contacts to the germanium been placed close together, they might have discovered the transistor effect. But the fact is that they did not.

There was no government funding of the Bell Labs work that led to the transistor. I believe that there was no significant funding of the work that led to any of the breakthroughs that made "the" difference. These were largely industry efforts funded by industry. This did not arise from a total unwillingness of the government to fund some of the work but resulted in part from a reluctance in industry to accept government funding for key development work.

There was one contribution that the Department of Defense (DOD) wanted to make but fortunately did not.

Shortly after the invention of the transistor, there was an effort to have the activity classified. Had the DOD not been persuaded otherwise, the whole story could have been different.

The government, through the military, made a major contribution to industry progress by being a customer for transistors. They were willing to buy the product as produced and rarely tried to control the product. (In the few cases where they did try, the track record was not very good.) During the 1950's, the government purchased as much as 50% of the output of the fledgling transistor industry. They were anxious to test and find applications for these devices, including the leading-edge product. There is hardly any better support that can be given to a new industry.

The above assessment, I believe, is fair and represents the government's contributing in ways it does best. This position is supported by the following statement of Noyce [36]:

> With very few exceptions, the major motivation behind technology development cannot come from the Military . . . the major motivation, I feel, is the commercial one. . . I would say that the research that was motivated by getting to a given end result was far more productive than the research that was carried on for the sake of carrying on research. A lot of the Military directly funded research was the latter. I would say that the Military created more motivation for doing good research by creating a market for advanced products. . . . The main reason we stayed clear of military involvement was because I thought it was an affront to any research people to say that you are not worth supporting out of real money. . . . In a sense, the military funding made whores out of all the research people. You were dealing with a critic of the research you were doing who was not capable of critiquing the work. . . . There are very few research directors anywhere in the world who are really adequate to the job . . . and they are not often career officers in the Army. [36]

E. An Element of Luck

There are a number of events that made "the" difference in the transistor story that must be recognized as being very fortunate, fortunate for society as a whole. It was surely fortunate that both of the elemental semiconductors, germanium and silicon, were relatively easy to purify and produce as single crystals and had electronic properties that were suitable for transistor action. The energy gaps provided a good balance between maximum operating temperature and minimum operating power. Dielectric strength permitted very usable maximum operating voltage. Minority carrier lifetimes supported fabrication of transistors with achievable dimensions. Carrier mobility ideally could have been higher, but it has been high enough to allow the achievement of multigigahertz performance before we "run out of atoms." Had any one of these factors been one order of magnitude less favorable, the hurdles may well have been insurmountable. The fact that all were in reasonable ranges is quite remarkable. Nature indeed need not have been so generous.

The capabilities of silicon dioxide are also most fortunate. Silicon dioxide is an excellent insulator and in a sandwich of metal or silicon makes a fine capacitor. Its diffusion masking properties combined with photolithography have led to a batch process with the ultimate ability to create 10 billion components on one chip of silicon. And as grown during the diffusion process, and with the addition of an overcoating of silicon nitride, it provides environmental protection to yield highly reliable devices that need only simple further packaging for mechanical protection, ease of handling, and access.

We surely were incredibly lucky to find one material and its oxide with which we could fabricate and encapsulate high-performance transistors and IC's. However, as L. Trevino says about his golf, "The harder I practice, the luckier I get." It took a lot of hard practice on the part of the scientists and engineers who created this technology to be smart enough to recognize and build on the luck that nature bestowed.

ACKNOWLEDGMENT

The author has relied heavily on accounts in [2], published by Bell Labs, and particularly on the section on "The Transistor" written by J. Hornbeck and edited by F. Smits. Smits was also helpful in clarifying some of the events covered in the text. The author also is grateful to G. Moore and J. Kilby for inputs on their contributions and the contributions of their companies. W. W. Troutman of Bell Labs provided valuable reference material and even more valuable encouragement. The author also wishes to thank AT&T for its support and particularly for providing material from its archives.

The author recognizes that in choosing the events that made "the" difference, he has exercised judgment that is far from perfect. There were a multitude of contributions that he classified as making "a" difference. Many of these were at least as meritorious and deserving of recognition as the ones described.

REFERENCES

[1] J. Bardeen, "Surface states and rectification at a metal semiconductor contact," *Phys. Rev.*, vol. 71, pp. 383–388, May 15, 1947.
[2] *Engineering & Science in the Bell System.* Indianapolis, IN: AT&T Bell Laboratories, 1985, vol. 4.
[3] J. Bardeen and W. H. Brattain, "The transistor, a semi-conductor triode," *Phys. Rev.*, vol. 74, pp. 230–231, July 15, 1947.
[4] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," *Bell Syst. Tech. J.*, vol. 28, pp. 435–489, July 1949.
[5] J. N. Shive, "Double-surface transistor," *Phys. Rev.*, vol. 75, pp. 689–690, Feb. 15, 1948.
[6] W. Shockley and M. Sparks, "Semiconductor translating device having controlled gain," U.S. Patent 2 623 105, Dec. 23, 1952.
[7] W. Shockley, "Bistable circuits including transistors," U.S. Patent 2 655 609, Oct. 13, 1953.

[8] J. L. Moll, M. Tannenbaum, J. M. Goldey, and N. Holonyak, "p-n-p-n transistor switches," *Proc. IRE*, vol. 44, pp. 1174–1182, Sept. 1956.
[9] G. K. Teal and J. B. Little, "Growth of germanium single crystals," *Phys. Rev.*, vol. 78, p. 647, June 1950.
[10] W. Shockley, G. L. Pearson, and J. R. Haynes, "Hole injection in germanium—Quantitative studies and filamentary transistors," *Bell Syst. Tech. J.*, vol. 28, pp. 344–366, July 1949.
[11] W. Shockley, M. Sparks, and G. K. Teal, "p-n junction transistors" *Phys. Rev.*, vol. 83, July 1951.
[12] W. Shockley, *Electrons and Holes in Semiconductors*. New York: Van Nostrand, 1950.
[13] J. E. Saby, "Fused impurity p-n-p transistors," *Proc. IRE*, vol. 40, pp. 1358–1360, Nov. 1952.
[14] G. C. Dacey and I. M. Ross, "Unipolar field-effect transistor" *Proc. IRE*, vol. 41, pp. 970–979, Aug. 1953.
[15] G. K. Teal and E. Buehler, "Growth of silicon single crystals and of single crystal silicon p-n junctions," *Phys. Rev.*, vol. 87, p. 190, July 1952.
[16] W. G. Pfann, "Principles of zone melting," *Trans. AIME*, vol. 94, pp. 747–753, July 1952.
[17] G. K. Teal, "Single crystals of germanium and silicon—Basic to the transistor and integrated circuit," *IEEE Trans. Electron Devices*, vol. ED-23, pp. 136–137, July 1976.
[18] C. S. Fuller, "Diffusion of donor and acceptor elements into germanium," *Phys. Rev.*, pp. 23–34, vol. 86, Apr. 1952.
[19] C. A. Lee, "A high-frequency diffused base germanium transistor," *Bell Syst. Tech. J.*, vol. 35, pp. 1–22, Jan. 1956.
[20] M. Tannenbaum and D. E. Thomas, "Diffused emitter and base silicon transistors," *Bell Syst. Tech. J.*, vol. 35, pp. 1401–1406, Jan. 1956.
[21] J. M. Early, "Effects of space-charge layer widening in junction transistors," *Proc. IRE*, vol. 40, p. 517, Nov. 1952.
[22] J. M. Early, "p-n-i-p and n-p-i-n junction transistor triodes," *Bell Syst. Tech. J.*, vol. 33, pp. 1642–1643, May 1954.
[23] H. C. Theuerer, J. J. Kleimack, H. H. Loar, and H. Christensen, "Epitaxial diffused transistors," *Proc. IRE*, vol. 48, pp. 547–552, Sept. 1960.
[24] C. J. Frosch and L. Derick, "Surface protection and selective masking during diffusion in silicon," *J. Electrochem. Soc.*, vol. 104, pp. 151–162, Sept. 1957.
[25] J. Andrus and W. L. Bond, "Photograving in transistor fabrication," in *Transistor Technology*, vol. III, F. J. Biondi, Ed. New York: Van Nostrand, 1958.
[26] I. M. Ross, L. A. D'Asaro, and H. H. Loar, presented at the Solid State Research Conf., Lafayette, IN, June 1956.
[27] M. M. Atalla, E. Tannenbaum, and E. J. Scheibner, "Stabilization of silicon surfaces by thermally grown oxides," *Bell Syst. Tech. J.*, vol. 38, pp. 749–783, May 1959.
[28] D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field induced surface devices," presented at the Solid State Research Conf., Pittsburgh, PA, June 1960.
[29] J. A. Hoerni, "Planar silicon diodes and transistors," *IRE Trans. Electron Devices*, vol. ED-8, p. 178, Mar. 1961.
[30] J. V. Dalton and J. Dorbek, "Structure and sodium migration in silicon nitride films," *J. Electrochem Soc.*, vol. 115, pp. 865–868, Aug. 1968.
[31] G. E. Moore and R. N. Noyce, U.S. Patent 3 108 359, Oct. 29, 1963.
[32] J. S. Kilby, "Invention of the integrated circuit," *IEEE Trans. Electron Devices*, vol. ED-23, pp. 648–653, July 1976.
[33] R. N. Noyce, U.S. Patent 2 981 887, Apr. 25, 1961.
[34] J. S. Kilby, "Invention of the integrated circuit," *IEEE Trans. Electron Devices*, vol. ED-23, p. 653, July 1976.
[35] W. Shockley, "The path to the conception of the junction transistor," *IEEE Trans. Electron Devices*, vol. ED-23, p. 619, July 1976.
[36] E. Braun and S. Macdonald, *Revolution in Miniature*. Cambridge: Cambridge Univ. Press, 1978, p. 142.
[37] R. W. Galvin, personal communication, Mar. 1996.
[38] M. Sparks, personal communication, Aug. 1996.

Ian M. Ross (Life Fellow, IEEE) was born in Southport, England. He received the bachelor's degree in engineering from Gonville and Caius College, Cambridge University, England, in 1948 and the M.A. and Ph.D. degrees in electrical engineering from Cambridge University.

In 1952, he joined AT&T's Bell Laboratories, and for the next decade was engaged in the development of semiconductor devices. He served as Director of the Semiconductor Laboratory in Murray Hill, NJ, from 1959 to 1962 and as a Director of the Semiconductor Device and Electron Tube Laboratory in Allentown, PA, for the next two years. In 1964, he became Managing Director of BellComm, the Bell System unit that provided systems engineering support for the Apollo manned space flight program. He became President of BellComm in 1965. Dr. Ross returned to Bell Laboratories in 1971 as Executive Director of the Network Planning Division. He held several key positions before becoming the laboratories' President in 1979. In July 1991, he was named President Emeritus of the laboratories.

Dr. Ross is a fellow of the American Academy of Arts and Sciences and a foreign associate member of the Engineering Academy of Japan. He was elected to the National Academy of Engineering in 1973, the National Academy of Sciences in 1982, and the Royal Academy of Engineering of Great Britain in 1990. He received the IRE Morris N. Liebman Award in 1963, the NASA Public Service Group Achievement Award in 1969 and 1975, the Industrial Research Institute Medal in 1987, the IEEE Founder's Medal in 1988, the American Electronics Association Medal of Achievement in 1991, and the Semiconductor Industry Association's Robert Noyce Award in 1992. He chaired the National Advisory Commission on Semiconductors established by Congress and the president from 1988 until June 1992. The commissions and councils on which he has served include the President's Commission on Industrial Competitiveness, Cochair of the Committee on R&D Manufacturing and Council on Competitiveness (the private council). He is Chairman of the Science and Technology Advisory Board, Taiwan, R.O.C., and a member of the National Science Board; Board of Directors, B. F. Goodrich Company; Board of Directors, Thomas & Betts Corporation; Board of Trustees, University of Medicine and Dentistry of New Jersey Foundation; Board of Directors (1979–1992), Sandia National Laboratories; and Board of Directors, NACCO Industries, Inc.