Part One
Fundamentals on Nanoelectronics

1
A Brief History of the Semiconductor Industry

Paolo A. Gargini

Stanford University, Department of Electrical Engineering, 475 Via Ortega Stanford, CA 94305, USA

1.1
From Microelectronics to Nanoelectronics and Beyond

The nineteenth century was the time when science discoveries began to morph into commercial applications. Electric lighting became a reality and soon after electron tubes paved the way for the rise of the electronics industry. By the mid-twentieth century, the transistor effect was demonstrated at Bell Labs, but it was the move of W.B. Shockley back to Palo Alto that laid the foundation of the semiconductor industry. The "Traitorous Eight" left Shockley Semiconductors in 1957 and went on to found Fairchild Semiconductors and later on were the seed to the formation of Intel. By 1972, more than 40 companies had been created in the surrounding area, which came to be known as "Silicon Valley."

1.1.1
You Got to Have Science, Genius!

Mapping and analyzing the relation between science, technology, and manufacturing has always yielded the most instructive lessons one can ever imagine. In essence, none of them can really survive without the others, so studying their relations and timing is fundamental to getting a better understanding of how revolutionary inventions are made.

> "Nothing is new but never is the same."

Scientists worked with electricity long before they understood that current was made of electrons. Thomas A. Edison brought electrical illumination to the world, but his major problem was not the science behind the creation of light but the filament lifetime. He kept on trying any materials known at the time and any possible technique to bring the lifetime of an illumination bulb in the 40 h

Nanoelectronics: Materials, Devices, Applications, First Edition. Edited by Robert Puers, Livio Baldi, Sebastiaan E. van Nooten, and Marcel Van de Voorde.

range with no success. In 1883, among his many failed attempts, he tried to place a secondary filament adjacent to the one that was powered up in the hope that this "cold" filament would somehow divert some of the heat away from the primary heated filament. During the experiments, he observed a current flowing in the "cooling" filament, took note of it, wrote a patent, but moved on since it had not produced any lifetime improvement. He eventually found the right filament material.

Still it was not clear "what was flowing" and it took until 1897 to find the answer. Joseph John Thomson was the British physicist who discovered the electron in a series of experiments designed to study the nature of the "rays" created in a cathode tube. Thomson interpreted the deflection of the rays by electrically charged plates and magnets as evidence of "bodies much smaller than atoms" that he calculated as having a very large value for the charge-to-mass ratio. Later he estimated the value of the charge itself.

J.J. Thomson received the Nobel Prize in 1906 "in recognition of the great merits of his theoretical and experimental investigations on the conduction of electricity by gases."

J.A. Fleming and L. DeForest invented the electronic diode and triode, respectively, by using T.A. Edison's observation of current flowing from one filament to the adjacent one. The main addition made by De Forest to the Edison's concept consisted in the insertion of a grid surrounding the cathode that controlled and modulated the flow of electrons with minimal power consumption. As a consequence of this action, the cathode-to-anode current carried the modulation information created with minimal power consumption by means of the grid voltage. The current flowing to the anode to a much higher power level transported the information carried by means of this modulation. With this experiment, the concept of signal amplification had been reduced to practice for the first time.

For the next 40 years, this technology revolutionized the world and created the field of electronics.

In the first 30 years of the twentieth century, new discoveries in the field of "pure science" completely changed our understanding of the world of physics. Quantum mechanics changed forever the purely deterministic perception of the world brilliantly formulated by Newton with the publication of his *Principia Mathematica* in 1687 and turned fundamental physics into a probabilistic world that would forever challenge our perception of what reality really is! But with this new understanding of physics, many new theories on how solid-state physics fundamentally worked began to come together.

Quantum physics explained how electrons were confined in specific energy bands in a solid and how these bands were in general separated from each other. The distance, as measured in energy terms, between bands determined whether these materials were conductors or insulators. If the upper bands were too far from each other, quantum mechanics showed that little or no flow of current was possible (insulator); but if these upper bands overlapped each other (metal), a large flow of charge was possible even with very little voltage applied. Of course, insulators could not become also good conductors and good conductors

could not also become good insulators on demand. So, in the end, semiconductors, characterized by the fact that the upper conduction band and the valence band (right below it energy-wise), were very close to each other demonstrated that this specific band combination could make the material work as a reasonable conductor and as reasonable insulator under the proper conditions; because of this property the materials were named semiconductors. Armed with this new knowledge, Julius E. Lilienfeld asked a very simple question:

> "If electrons are already in any solid and they can be moved around in a controlled way, why are we extracting them (via a heated filament), manipulating them via a grid and finally collecting them again at the anode?"

Couldn't we do all of these operations within a solid material, he thought? With this in mind, he published multiple patents between 1928 and 1935 in which he outlined the functionality principles of at least seven solid-state devices, including the basic MOS device!

1.1.2 What Would Science Be Without Technology?

Even though Lilienfeld understood how an MOS device could ideally function, he still had to deal with the limited level of solid-state technology existing at the time. In one of his patents he described how to make a gate for an MOS device. It consisted in creating a structure whereby a foil of aluminum, or any other conductor, was sandwiched between two layers of glass and then placed perpendicularly on the surface of a semiconductor. Very simple, but hardly functional!

So time went by with good ideas coming forward, but still without a real demonstration of a solid-state device showing some gain. It was not until 1945 that a concerted effort toward the demonstration of the "transistor effect" got on the way at Bell Labs under the direction of W.B. Shockley.

Shockley was born in 1910 in London, UK to American parents, and was raised in his family's hometown of Palo Alto, CA, from age 3.

After college, Shockley worked at Bell Labs where he filed a patent in 1945 showing a device composed of a source, a gate, and a drain region. In this patent he outlined the concept of how the flow of charges from a first region (source) to a receiving region (drain) could be controlled by the voltage applied to an electrode (gate) placed parallel and in proximity to the semiconductor surface without touching it.

It is however interesting to notice that many of the patents submitted by Bell Labs on the concept of transistors were rejected because they infringed on Lilienfeld's patents.

However, the group of researchers at Bell Labs discovered that it was almost impossible to make a real MOS device (Figure 1.1) in germanium because the "dangling bonds" left by the nonterminated bonds of the atoms on the surface of the semiconductor trapped charges and by so doing prevented the electric field

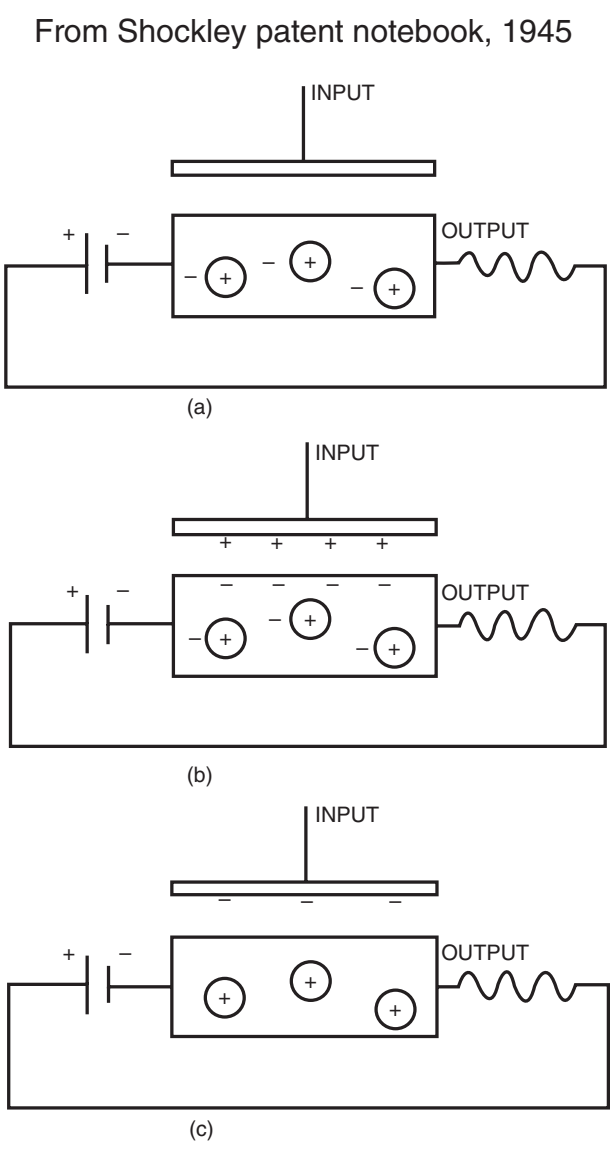


Figure 1.1 Concept of field-effect transistor at Bell Labs in 1945.

generated by the gate from controlling the flow of charge from source to drain. Finally, John Bardeen and W.H. Brattain, after trying just about anything they could think of, created an apparatus (Figure 1.2) not too different from what Lilienfeld had proposed 20 years before and thus placed emitter and collector connectors (narrowly separated by cutting a small gap in a wrap-around gold wire with a razor) in direct contact with the surface of the semiconductor. The base was contacted from the back of the semiconductor slice. Much to their surprise,

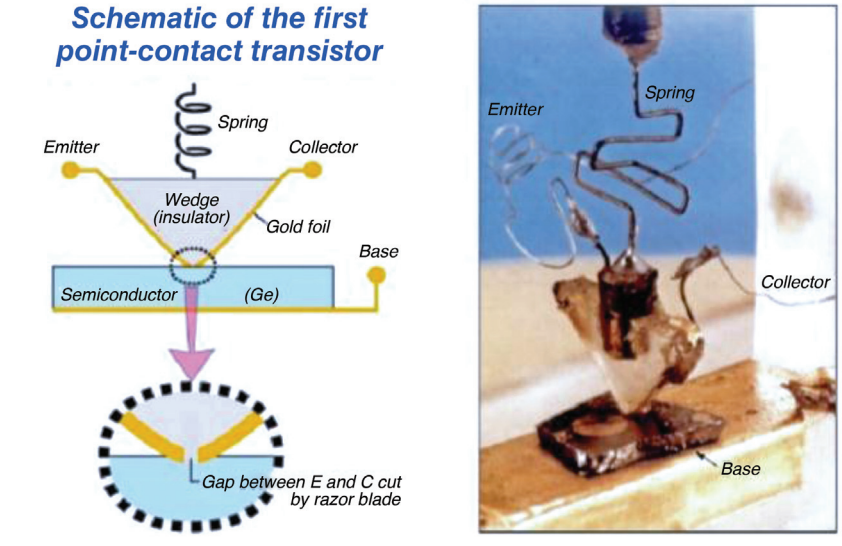


Figure 1.2 Apparatus used by John Bardeen and W.H. Brattain to demonstrate the first transistor.

on December 16, 1947, the device showed a gain of 15 when comparing the input voltage signal applied to the base of the device with the voltage signal measured on the collector! Amplification by a solid-state device had been finally demonstrated!

Shockley was very unhappy since he was not involved in the patent and therefore began relentlessly working on his own approach that actually was much better thought out. If the surface of the semiconductor was the source of the problem, he thought, why not trying to flow the current just below the surface? In fact, he diffused source and drain into the semiconductor and by so doing he demonstrated the "diffused transistor" on January 23, 1948. These transistors were termed bipolar transistors because the operation was a very complicated interaction of electrons (negative charge) and holes (positive charge). It would take another 20 years before commercialization of MOS transistors could take off. The news of the discovery of the transistor did not reach the front page of any famous newspaper then, but it did make the cover of *Electronics*, the trade magazine of the time.

In 1956, Shockley moved from New Jersey to Mountain View, CA, to start Shockley Semiconductors Laboratory to live closer to his ailing mother in Palo Alto, CA. He hired several new bright but quite inexperienced engineers to start his operation at a very reasonable cost.

This event marked the destiny of what was to become the so-called Silicon Valley!

On December 10, 1956 Shockley, Bardeen, and Brattain received the Nobel Prize for their inventions of the transistor. However, as it happens at times scientific genius and friendly human rapport do not seem to easily coexist in a single person and on September 18, 1957, eight of Shockley's brightest engineers decided to leave the company because of the very difficult relation with their leader who in accordance with his personality simply wrote in his daily journal: "Wed 18 Sept, Group resigns."

At that time stable employment and loyalty to a company were considered a lifetime commitment, and this group resignation broke all the rules of Corporate America. So the "traitorous eight," as they were labeled, left for a very uncertain future. One by one all the companies in their list refused to give them any financial support, but after failing 35 times they were finally saved by an investment from Fairchild Camera & Instruments.

On October 4, 1957, less than 1 month after their departure from Shockley Semiconductors, the world changed forever as the then Soviet Union launched Sputnik 1 and a month later Sputnik 2 (do you remember Laika?). All of a sudden the race to space had initiated and each pound of weight carried into space came at a high premium. Replacing vacuum tubes with much lighter transistors was worth a fortune.

Single transistors were then selling for over $100!

The first U.S. satellite was launched on January 31, 1958 and it weighted 14 kg (30.8 lbs).

IBM was also quick to realize the importance of the transistor and in 1958 Fairchild business in silicon transistors was already $500 000 and it reached $21 million in 1960.

Lilienfeld had conceived the MOS transistor in the 1925–1928 period, the Bell Labs effort had demonstrated the transistor effect in the 1947–1948 time and 20 years had gone by. It was not until a viable, even though very simple, technology was finally developed in the late 1950s that transistors became a viable business but much more technology was still needed to claim success.

In the end, it is all Manufacturing, stupid!

Until the nineteenth century, skillful and specialized individuals who spent a lifetime to learn their trade individually made most products. A single expert craftsman or team of craftsmen would create each part of a product. They would then use their skills and tools to put it together. The first organized production lines came into being with the invention of the steam engine, but it was the automobile industry at the beginning of the twentieth century that really created and implemented the modern assembly line method of manufacturing complex items (i.e., Ransom Olds and Henry Ford). It is reported that a model T could be produced in 93 min. No longer were the rich people the only customers targeted for the sales of one-of-a-kind articles, but the new category of anonymous "consumers" had been created.

Robert Noyce, the leader of the founders of Fairchild Semiconductors, realized that the financial success stemming from sales of transistors to the government was eventually going to end and challenged the whole team to produce transistors that could be selling, at a profit, for $1. He realized that the easy government

money was eventually going to substantially decline and that the semiconductor industry still needed to generate a cost-effective manufacturing process.

In the 1950s, it became clear that germanium was not easy to process and silicon became quickly the material of choice. Semiconductor wafers were cut from ingots, wafers were then divided into dice and each individual die was individually processed. Initially, all the three regions of a bipolar transistor were made by placing the right dopants on each transistor, one-by-one, and then alloying them. Subsequently, the technology evolved and the base region was made by diffusion and eventually base and emitted were made by simultaneous diffusion by taking advantage of the faster rate of diffusion of aluminum (p-type) versus antimony (n-type). This meant introducing the right chemicals into selected regions, assembling the "to be transistor" in a package, and making connections between the various regions of the transistor and the package with golden wires. The whole process was highly artisanal in nature and very labor-intensive. It is not by accident that assembly was early on transferred to Asia where cost of labor was significantly less that in the United States.

Classical manufacturing "assembly lines" at the time (e.g., automobiles) relied on some form of mechanical automation to move the product from one workstation to the other. Most of the operations occurring at each station consisted in assembling prefabricated (mostly mechanical) parts onto a body and so on. None of these techniques seemed applicable to the production of semiconductors. The size of the dice was so small that it was practically impossible to pick them up and move them around with the automation technology available at the time. But the advent of the upcoming semiconductor industry had brought into the scene a completely new breed of scientists. Physicists and most of all chemical scientists had become the key players in transistor manufacturing, so a very different approach to a new type of "assembly line" had to come from something hidden deep into the making of the transistor and not from mechanization.

In early 1955, Bell Labs researchers encountered a major problem with pitting on the surface of silicon wafers during high-temperature diffusion. This problem was overcome by chemist Carl Frosch during a serendipitous accident in which the hydrogen gas carrying impurities through the diffusion furnace briefly caught fire, introducing water vapor into the chamber. The resulting "wet-ambient" diffusion method had covered the silicon surface with a layer of glassy silicon dioxide (SiO_2).

Developed further by Frosch and his technician Lincoln Derick in the ensuing months, this technique allowed semiconductor workers to seal and protect silicon wafers during the diffusion process by means of covering the surface with an oxide layer. The two men established what impurities, such as gallium, could penetrate the oxide layer and which others (boron and phosphorus, for example) could not. They also demonstrated how to etch small openings in the layer in order to diffuse these impurities into selected portions of the silicon surface and pattern it precisely in tiny n- and p-type regions. In 1957, they patented and published this extremely important technique.

The first breakthrough toward realizing an "assembly line"-type process came from Jean Hoerni (one of the traitorous eight) in December 1957. He suggested

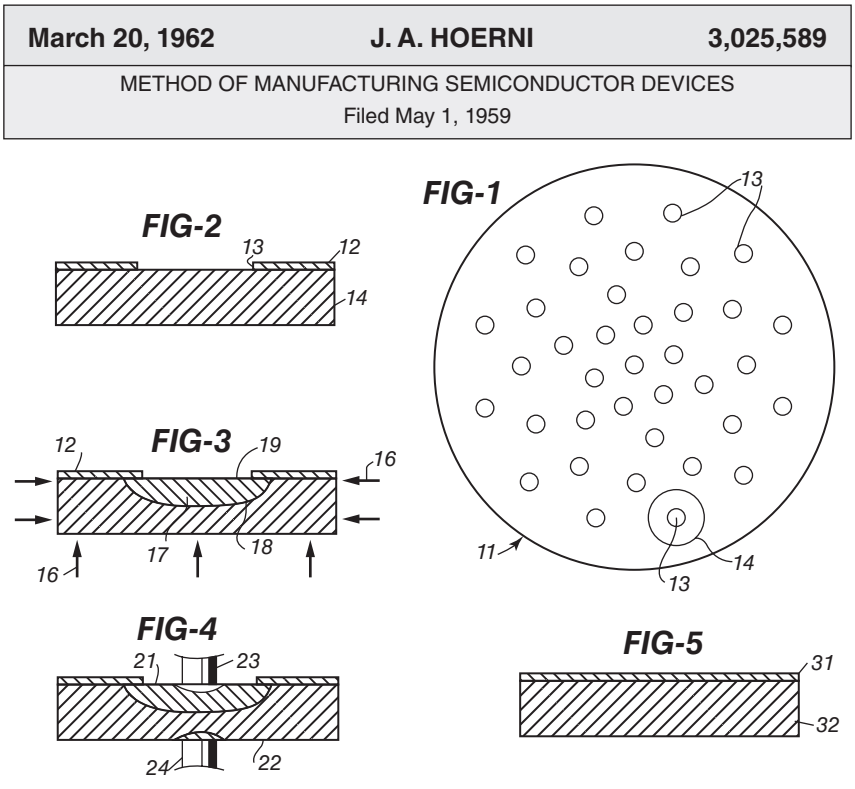


Figure 1.3 Jean Hoerni invented the planar process in May 1959.

using a layer of oxide grown on the silicon wafers as a way of protecting the underlying silicon from any form of contamination. Subsequently, selected windows were opened into the oxide exposing the silicon where the base and subsequent emitter forming operations were performed. All the transistors on a single wafer could be simultaneously processed in a batch process and dicing was done only before assembling the transistors into the package (Figure 1.3) (US patent 3,108,914, filed May 1959).

The final contribution to the manufacturing process came from Robert Noyce in 1958. Noyce conceived the idea that instead of dicing the individual transistors, packaging, and then assembling them to make a practical circuit, it was perhaps possible to connect them to realize a circuit when the transistors were still in the same wafer. He proposed that once all the regions of all the transistors (still in the same wafer) were completed by means of the (planar) process described by Jean Hoerni, windows could be opened in the final oxide layer exposing emitter and base contacts and then all these regions could be covered by a uniform metal deposition. The metal patterns were finally defined via a photolithographic and etch process that connected all the transistors according to the specific circuit design, he suggested.

The dream of an "assembly line" manufacturing process for semiconductors was finally realized.

However, Noyce tried to use all different types of metal to interconnect the planar transistors without success and so he finally asked Gordon Moore, which metal was still left to try. Gordon replied that they had not tried aluminum but he was not sure if it is going to work. Aluminum forms a diode when in contact with bare silicon. However, it turned out that due to the high concentration of dopants introduced in silicon to form the different emitter and base regions, aluminum formed a very leaky diode (i.e., a good contact). The missing element needed to realize the first integrated circuit (IC) was finally in place (Figure 1.4) (US patent 2,981,877, filed July 30, 1959).

In the fall of 1958, Texas Instrument's Jack Kilby succeeded in demonstrating that the monolithic circuit concept was also a practical possibility. He produced an integrated circuit – a linear oscillator involving a transistor, a resistor, and a capacitor formed from a single slice of germanium crystal. Gold wires interconnected the various regions. These methods made the connection of components within the circuit still very labor-intensive.

TI announced this breakthrough on March 6, 1959 and Noyce and his team found themselves somewhat on the defensive. As you can imagine, discussions between the two companies were quite intense for the next few months.

In April 1960, Fairchild sold its first planar transistor, the 2N1613 – a metal cylinder about half a centimeter in diameter and almost as high, with three little metal legs sticking out beneath it. A few months later, Noyce and Moore decreed that henceforth all the company's transistors would be planar.

Jay Last (another Fairchild founder) formed a group in the fall of 1959, aiming to manufacture integrated circuits based on Hoerni's planar process. It took another 18 months before the first commercial microchips, Fairchild's Micrologic series, reached the market.

But Fairchild still came out with its microchip more than 6 months ahead of TI, which succeeded only after it began using the *planar technology it had licensed from Fairchild.*

From 1962 to 1964, multiple integrated circuits were commercialized, but the real popularity of ICs was reached in 1965 when the mA709, the first general-purpose operational amplifier, reached the market and in no time it was used throughout the electronic industry!

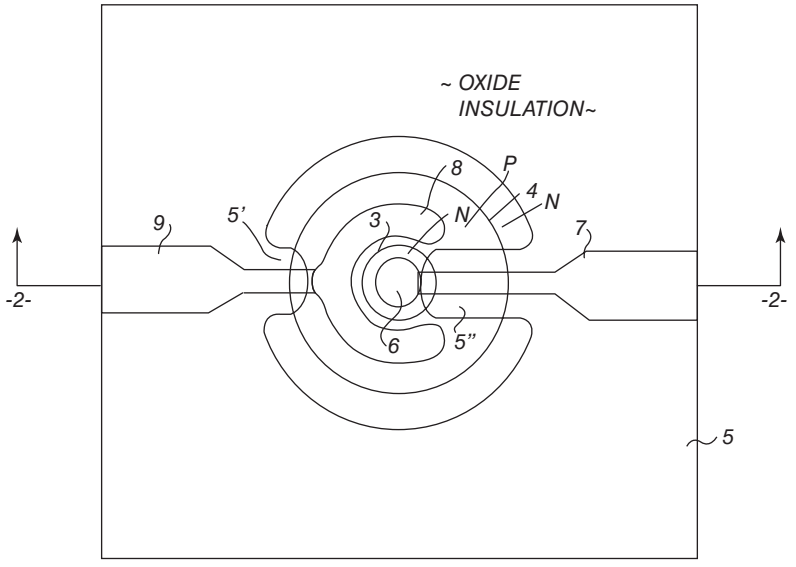
It should be noticed that all these integrated circuits were constituted of bipolar devices of the type demonstrated by Bell Labs. The MOS transistor still remained an evasive goal.

1.1.3
The Magic of Economics

In the mid-1960s, most electrical engineers were still quite skeptical about the use of integrated circuits. After all they were used to design their own circuits, buy the transistors, and build their systems.

April 25, 1961	P. N. NOYCE	2,981,877
SEMICONDUCTOR DEVICE-AND-LEAD STRUCTURE		
Filed July 30, 1959		3 Sheets-Sheet 1

FIG-1



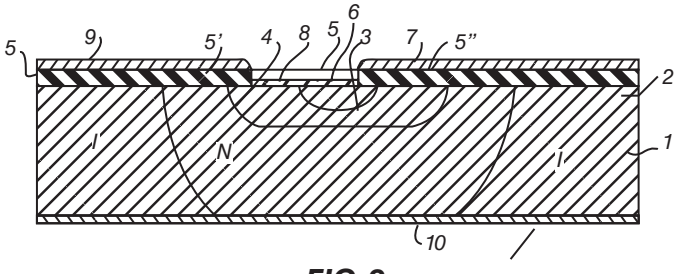


FIG-2

Figure 1.4 Robert N. Noyce invented the integrated circuit in July 1959.

Prices for integrated circuits were still above what many systems producers were accustomed to paying for electronic components. The technology was new, the volumes were relatively low compared to discrete devices, and the integrated circuits were most frequently built to stringent military specifications for performance and reliability. Moreover, integrated circuits required potential customers to adopt a new mode for evaluating the cost of electronic components, one that encompassed the costs associated with a set of the equivalent discrete components and the labor to interconnect them to form a complete circuit.

By 1964 the supporters of IC joined forces to elucidate the economical advantages of ICs.

The back of the envelope calculations began to emerge. In an idealized case, assuming that the cost of processing 1 in.^{2} of silicon to produce integrated circuits was \$10, as said C. Lester Hogan, VP of Motorola semiconductor division, with a yield of 100% such wafer could produce 400 ICs at a cost of \$10/400 = 0.025 cents. This cost was placing the entire IC in direct competition with the cost of individual discrete transistors. Noyce also advanced similar calculations. The concept began to emerge that if ongoing manufacturing costs were growing at a small rate as IC complexity was increasing at a faster rate, then the cost of individual transistors in ICs would continue to decrease. Discussions were heated and quite animated.

But it was Gordon Moore who finally brought it all together in 1965. His observation and forecast was that for any new technology generation in the evolution of integrated circuit manufacturing, there was an *optimal manufacturing point, as measured by the number of components, on an integrated circuit associated the minimum manufacturing cost per component.* He observed that the more the transistors packed in a fixed space on the wafer, the lower the cost/transistor, but at some point transistors will get "too close to each other" and electrical defects will decrease yields and increase the cost (Figure 1.5). However, with the introduction of any new technology generation, Moore argued, this

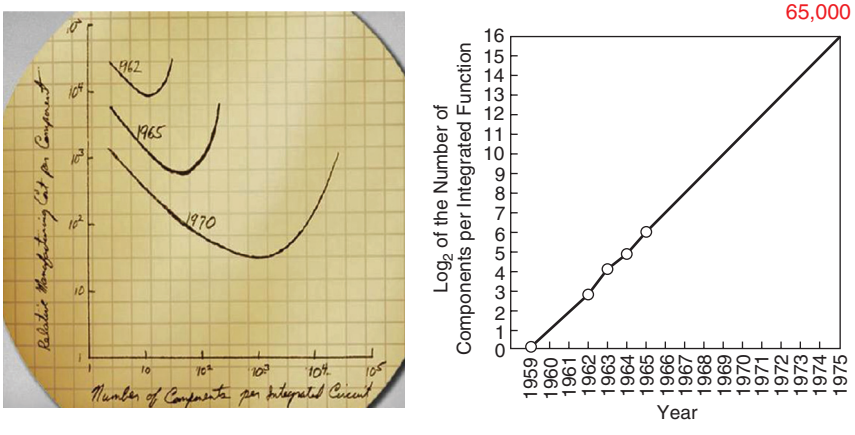


Figure 1.5 Gordon E. Moore's first enunciation of his famous law.

optimal point would shift to both greater complexity and lower minimum manufacturing cost. In few words, in the long run, the cost of transistors in integrated circuit was going to be reduced to a negligible amount.

Moore was then requested to estimate the annual rate at which the density of transistors was going to grow in the foreseeable future. Using all the available historical data available to him at the time, he noted that the complexity of integrated circuits – as measured by the number of components per integrated circuit – had doubled every year between 1958 and 1965, and he stated that this trend was going to last for at least the next 10 years! (Figure 1.5). So Gordon then asked the question:

> "What could you do with 65,000 transistor in a single IC?"

1.1.4 Back to the MOS

It is important to point out that all the integrated circuits up to this point were constituted of bipolar transistors. The problem associated with the surface states that had plagued Shockley's first attempt to making an MOS transistor was still baffling the scientists.

The first long-awaited breakthrough came from John Atalla and Dawon Kahng at Bell Labs. They envisioned a surface passivation method that substantially reduced the influence of the dangling bond that had prevented the Shockley research team from realizing the first MOS device. Atalla and Kahng announced their successful MOSFET at a 1960 conference.

The final push toward the realization of commercial MOS transistors came when Bruce E. Deal and Andrew S. Grove finally reported a method on how to minimize the effect of surface states; a practical MOS transistor was finally in sight. The MOS technology of the time started with the definition and doping of source and drain followed by the growth of a thin oxide (~1000–1500 A). Contacts were then etched in the oxide covering the whole wafer and an aluminum layer was then deposed. By means of sequential photolithographic and etch steps, the gate and interconnections were then simultaneously defined. Due to the inevitable misalignment of the gate mask with respect to the source and drain mask, it was necessary to have a fairly large overlap area between the gate region and the source and drain to ensure that the aluminum gate would cover the edge of the source and drain even under worse-case misalignment. This requirement resulted in rather large gate-to-source and gate-to-drain parasitic capacitances varying from wafer to wafer, depending on mask-to-mask misalignment. This effect highly reduced the speed at which the transistor could operate and also made the speed distribution across the wafer quite broad.

In 1967, John C. Sarace and collaborators at Bell Labs replaced the aluminum gate with an electrode made of vacuum-evaporated amorphous silicon and

succeeded in building working *self-aligned gate MOS transistors*. However, the process, as described, was only a proof of principle, suitable only for the fabrication of discrete transistors and not for integrated circuits; and was not pursued any further by its investigators.

In late 1967, Tom Klein, under Les Vadasz guidance, while working at Fairchild R&D Labs, realized that the work function difference between a p-type doped silicon gate and the n-type substrate could actually be lower than the work function difference between an aluminum gate and the same silicon substrate by as much as 1.1 V.

In February 1968, Federico (Freddy) Faggin joined Les Vadasz's research group and was put in charge of the development of a low-voltage, self-aligned silicon gate MOS process.

Federico Faggin quickly developed a precise etching solution to define the polysilicon gate. It should be noticed that when I joined Fairchild Semiconductors in 1973, the self-aligned polysilicon gate etch solution (whose composition was still known to only a few) was still called "Freddy's etch!" He also developed the whole architecture for the fabrication of the MOS self-aligned silicon gate. With this last step, the development of the manufacturing technology had been completed and the MOS device of Shockley's dreams had finally become a reality!

Federico Faggin designed the first integrated circuit using polysilicon gate, the Fairchild 3708, an 8-bit analog multiplexer with decoding logic that went on sale in July 1968. By the end of 1968, the silicon gate technology had achieved impressive results.

Finally, the integrated circuit was beginning to surge to a higher level of engineering inventions. In the mean time, back at Bell Labs two researches, Kahng and S.M. Sze, were experimenting with a novel device capable of retaining information in a nonvolatile mode. By May 1967, they had fabricated a floating gate device. This consisted of an MOS device with two gates where the gate closer to the silicon surface was completely insulated (i.e., no electrical contact was made to it). By applying voltage to a second overlaying gate, they were able to capacitively couple enough voltage into the floating gate until charge from the silicon channel flowed into it. Once the programming voltage was removed, the charge into the floating gate controlled the current flowing in the MOS device. They did not provide an easy way of erasing the memory charge but the concept of an electrically programmable read-only memory was born.

1.1.5
Technology Innovation Must Go On!

By 1968, the semiconductor business had surpassed the $2 billion level and Fairchild Semiconductors had became very profitable, but soon most of the profits began flowing into other operations within Fairchild Camera and Instruments and Fairchild Semiconductors was losing control of its future. Much R&D still needed to be done and this needed money and free initiative; so, for the second

time, Noyce and Moore were on the move and this time they wanted to be in complete control. On July 18, 1968 Intel was incorporated with the two scientists at the helm, deciding the future of the company.

In the subsequent 4 years, most of the devices that represent the Electronic Industry foundation were demonstrated and commercialized, but also a much bigger technology battle was looming on the horizon.

1.1.6 Bipolar against MOS!

In very simplistic terms, bipolar transistors were vertically organized transistors; the critical charge controlling the behavior of the transistor was stored in the base region and the width and depth of the base were responsible for most of the time needed for charges to travel and switch a transistor on or off. A base depth of below 0.5 μm could be achieved at the time. This dimension was controlled by the accuracy of diffusion of the base and emitter dopants. Most of all, in bipolar transistors, the current was vertically flowing from emitter to collector completely contained within the bulk of the semiconductor. This implied that surface effects were not relevant for how well a bipolar transistor behaved. On the other hand, MOS transistors were by design horizontal transistors, with source, gate, and drain horizontally laid out as three regions adjacent to each other. As far as performance is concerned, gate lengths were on the order of at least 10 μm since they were defined by lithography and etch accuracy and by gate-to-source and gate-to-drain overlap requirements. In summary, an MOS transistor easily consumed at least five times the space of a bipolar transistor. Charges had to vertically travel 0.5 μm in a bipolar transistor and instead they had to horizontally travel 10 μm in an MOS transistor. Furthermore, MOS transistors had to deal with the quality of the silicon–silicon dioxide interface that was still a challenging element with respect to the performance of MOS transistors. Not surprisingly, in 1970 Fairchild Semiconductors delivered the first 256-bit static random-access memory (RAM) fabricated using bipolar transistors. The Fairchild 4100 required only 70 ns to either write or read 1 bit in the memory array. This product became instantaneously popular and was used in the Iliac IV computer.

Intel also launched its first product, the 3101, a 64-bit Schottky bipolar RAM rated at 35 ns access time in 1969.

However, MOS circuits required only four masks and only one diffusion step to be fabricated compared to the seven to eight masks and the multiple diffusion and oxidation steps required to fabricate bipolar devices. It was clear that the MOS could be fabricated at a much lower cost if all the remaining technical problems could be solved. In 1967, B.E. Deal, M. Sklar, A.S. Grove, and E.H. Snow published a method of how the surface states that had plagued MOS development for 20 years could be controlled and minimized (Figure 1.6). It was Intel that finally took advantage of this breakthrough and launched the world's

J. Electrochem. Soc.: SOLID STATE SCIENCE *March* 1967

Characteristics of the Surface-State Charge (Q_{SS}) of Thermally Oxidized Slicon

B. E. Degl, M. Sklar, A. S. Grove, and E. H. Snow

Research and Development Laboratories, Fairchild Semiconductor, Palo Alto, California

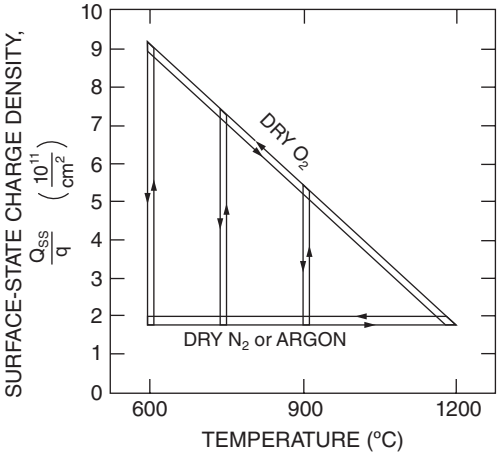


Figure 1.6 Methods used to reduce surface states.

first 256-bit MOS static memory 1101 RAM with access times on the order of 1000–1500 ns. But the real event that sealed the success of MOS technology over bipolar technology was the launch of the first 1024-bit dynamic RAM, the 1103, with access times of 300 ns, and this was entirely built with PMOS transistors! It also launched the first microprocessor, the 4004, that would eventually mark the beginning of the success for Intel. Intel closed the year with revenue of $4 million.

As you remember Kahng and Sze had demonstrated how to put charge in a floating gate and Dov Frohman-Bentchkowsky completed the task in 1971 by making the device writable and erasable; the Intel 1701 had an architecture comprising 256 long word by 8 bits. This time the device was built with a transparent quartz lid that allowed erasure by exposing it to UV light at about 257 nm – (Floating gate transistor and method for charging and discharging same, US patent 3,660,819 A).

The DRAM was actually the first product that fully utilized the main advantage of the planar process: physical and electrical properties of MOS transistors across each die were very closely matched with an unprecedented accuracy. This

led to the DRAM architectural methodology whereby the charge stored in a dynamic mode on any capacitor anywhere in the die was readily compared by means of a differential sense amplifier to the charge placed on a reference capacitor. This method allowed easily writing and reading any memory cell in a die.

Another fundamental discovery was first publicized by Robert Dennard in 1972 and further refined in 1974. He noticed that most of the equations related to an MOS device could be connected via a common scaling parameter (Figure 1.10). By setting this parameter to a specific value, it was possible to quickly derive all the properties of the scaled-down transistor. This implied that the design of the next scale-down MOS could be done in a relatively short amount of time using Dennard's equations instead of making multiple cut and try attempts.

Finally, the silicon gate process had been established as a solid method of making cost-effective ICs. Dennard's scaling laws enabled the understanding of how an MOS transistor became very accessible to anyone who had some basic knowledge of physics. Most of all, the obscure prophecy made by Moore in 1965 was becoming a very visible reality and all the sudden venture capitalists looked at the semiconductor industry with benevolent eyes. Many people (engineers and marketing types) began to realize that it was not too difficult to identify new not yet realized products and new company "start-up" mania became a major trend. About a block from Fairchild Semiconductors, there was a bar called "the Wagon Wheel." All that anybody who wanted to start a new company had to do was to walk in around 4:00 pm and wait for possible prospects to straddle in and simply whisper: "There is a new start-up . . ."

Between 1965 and 1975, more than 40 new semiconductor companies sprung out, in what came to be known as "Silicon Valley."

The competition between bipolar and MOS was beginning to take shape. Bipolar RAM were small, static, and fast; MOS DRAMs were slower but could be produced at a higher density than bipolar memories.

All the early MOS integrated circuit consisted of PMOS transistors. Even though the understanding of how to deal with surface states had progressed and it was possible to deal with the issue in a manageable way, there was still the problem of the fix positive charge residing at the oxide interface. This charge induced negative charge in the underlying silicon that actually provided some natural isolation in the case of circuit made of PMOS transistors. It was well known that electrons traveled two to three times faster than holes in silicon, but making NMOS transistors viable meant finding a way of eliminating the positive charge at the silicon dioxide interface. The induced negative charge residing under the isolation between two adjacent NMOS transistors would have shorted one transistor to the other. So this time it was up to creative engineering to solve the problem since science was not coming to the rescue.

An additional problem consisted in the fact that transistor-to-transistor isolation, even when using PMOS transistors, was accomplished by growing a thick layer of oxide (0.5–1 μm) between any two adjacent transistors. This was

equivalent to interposing a transistor with a high threshold voltage between them and this approach limited the amount of leakage between two adjacent transistors.

However, this approach created steep steps that were difficult to overcome without problems for subsequent metal interconnections creating "metal cracks" on steps. These metal defects created either open-circuit failures or reliability failures.

The use of local oxidation of silicon (LOCOS) invented by Else Kooi in 1967 largely alleviated this problem. This approach consisted in covering all the active areas with a combination of a very thin silicon dioxide layer (on the bare silicon) covered with a silicon nitride layer. Subsequently, after oxidation the silicon nitride was removed from the isolation areas. Since silicon nitride does not practically oxidize, it meant that exposing the wafer to an oxidizing ambient induced silicon dioxide growth only in the areas not covered by the silicon nitride. However, some oxygen penetrated sideway underneath the silicon nitride creating some level of oxidation. The overall result yielded a very gradual transition between the field oxide and the active regions, thus limiting metal step coverage problems.

The complete transformation of the isolation process, aimed at solving the challenges of introducing into manufacturing NMOS transistors, was accomplished by combining the LOCOS process with the introduction of a p-type dopant in the isolation regions before oxidation. Diffusion was first used as the doping technique, but it was then replaced later on by ion implantation for increased accuracy. Finally, with this new isolation method well under control, the transition from PMOS to NMOS occurred in the mid-1970s and the stage for the final confrontation between bipolar and NMOS was now set.

The showdown occurred at the 4-kbit level. Silicon gate devices could by then be fabricated with gate lengths of about 6 μm and with minimal gate overlap to either source or drain by virtue of the silicon gate process. The 2147 H static 4 K RAM was produced with access times as low as 45 ns. On the other hand, bipolar device yields were negatively affected by several defects introduced during the fabrication of the emitter, and the most deleterious of them all was called "emitter pipes."

It was established that a number of processing defects might produce emitter-to-collector shorts in double diffused bipolar transistors. However, by far the most interesting and most troublesome incidence of emitter–collector shorts was that due to "pipes." These were due to accelerated n-dopant diffusion along dislocation or contaminants that penetrated the base region and shorted the emitter to the collector with consequent catastrophic effects on yields.

While excessive phosphorus penetration through the base region was hampering progress in bipolar technology, it was phosphorous diffusion properties that were making the NMOS process more competitive. The technology used for aluminum deposition was still relegated to evaporation in the mid-1970s. This type of deposition operated by line-of-sight, very similar to light illumination, in the sense that any protruding structure created adjacent regions where the aluminum deposition was shadowed with the result that the step coverage of aluminum lines was drastically affected by the topology of the wafer. In addition,

aluminum reaction with the underlying silicon created long aluminum alloy spikes that were shorting the aluminum to the substrate underlying the source and drain regions. But here came phosphorous to the rescue. Intel pioneered a process called "dielectric reflow." It consists in placing enough phosphorous in the final dielectric and then opening the contacts to source and drain. By then placing the wafer in a furnace at temperatures on the order of 1050–1100 °C in phosphorus ambient, two things were occurring. The high phosphorous concentration in the dielectric caused it to reach a viscous state whereby the oxide underwent a partial flow. This new smooth topology eliminated aluminum step coverage problems. In addition, the phosphorus diffused in the open contacts created a very deep pocket (~2–3 μm) that fully contained any aluminum spike well within the source and drain diffusions. Of course, all was well in a hermetic package, but in a plastic package moisture penetration caused a phosphorus compound that attacked aluminum lines. Eventually, the deposition of an oxynitride layer impervious to humidity but still flexible enough to adjust to the plastic packaging process solved the problem.

In essence, better MOS yields at the 4K-bit RAM level made them readily available, whereas corresponding bipolar memories were not available or too expensive and with this accomplishment MOS claimed the complete victory over bipolar.

1.1.7 Finally It All Comes Together

Moore's prediction published in 1965 had become a verified reality by 1975. Moore was then called to present at the International Electron Devices Meeting (IEDM) and this time he predicted that the transistor growth rate was going to reduce to doubling every 2 years in the foreseeable future (Figure 1.7).

This time his prediction was officially labeled as "Moore's law." It was now time to combine Dennard scaling rules (Figure 1.8) with Moore's transistor growth rate prediction. By setting the linear scaling factor in Dennard's equations to 0.7, it automatically followed that the area reduction of a transistor was set at 50% generation to generation. Conversely, it could be said that in order to double the number of transistors placed in an integrated circuit as Moore had recommended, it was necessary to set the linear scaling factor to 0.7. Of course other factors like die size, transistor layout, and overall circuit cleverness were key contributors to doubling the number of transistors integrated on a single die. This powerful combination began to drive the growth of the semiconductor industry like no other industry had ever seen before.

Nonvolatile memory products (NVM) were indeed continuing to grow at Moore's law pace in the 1980s. DRAMs were desperately trying to keep up with the increasing demand of the computer industry that required a quadrupling of memory every 3 years consistently with the design cycle of this industry. However, technology could only provide a 2× contribution, and so additional number of transistors was needed to come from more aggressive design rules but most of

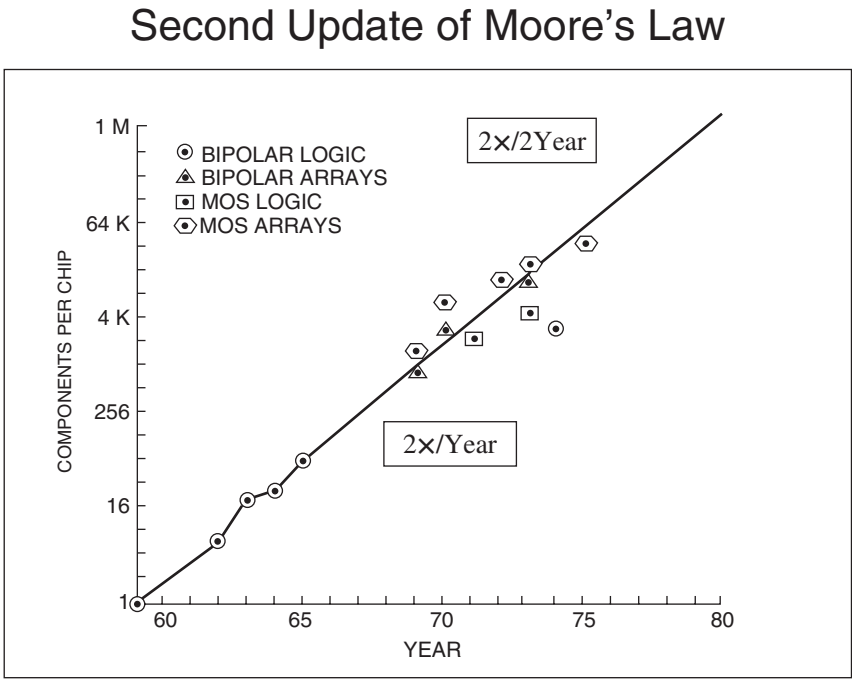


Figure 1.7 International Electron Device Meeting, December 1975.

all by increasing the dies size by 40–50%. Of course, these introductory DRAMs were much more expensive due to the enhanced consumption of silicon real estate, and within 6–12 months it was necessary to "shrink" the die to make the product economically viable. So, even though product introduction was on a 3-year cycle, the economics demanded a reduction in die size that brought the pace

Dennard Scaling

Device or Circuit Parameter	Scaling Factor
Device dimension tox, L, W	1/K
Doping concentration Na	K
Voltage V	1/K
Current I	1/K
Capacitance eA/t	1/K
Delay time per circuit VC/I	1/K
Power dissipation per circut VI	$1/K^2$
Power density VI/A	1

Figure 1.8 Robert H. Dennard scaling laws.

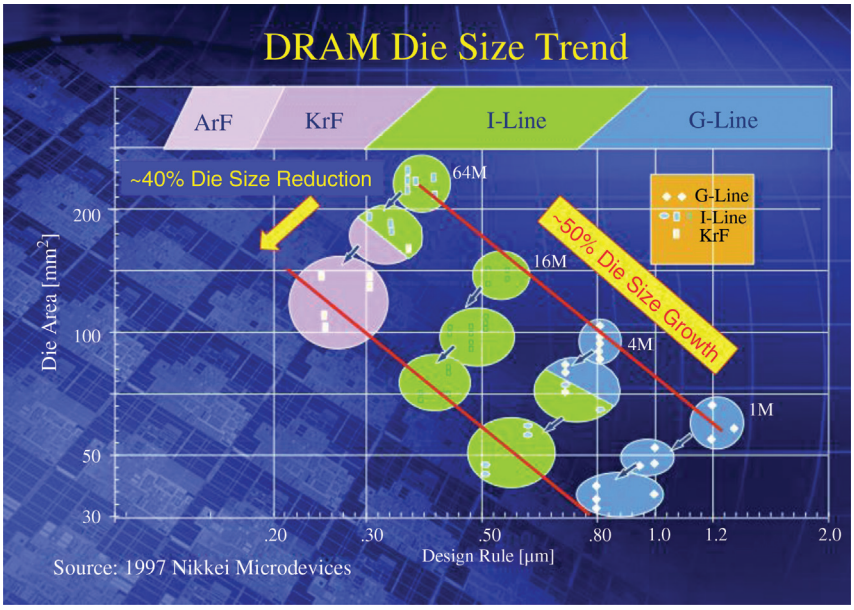


Figure 1.9 DRAM die size trend.

of real volume runners close to 4 years. In summary, 4×/3years in reality approximately corresponded to about 4×/4years that means 2×/2years (Figure 1.9).

On the other hand, microprocessors were more difficult to design, debug, and could not work without appropriate instructions set. Not to mention that practical applications were very few. For these reasons, microprocessors were introduced with 4× the number of transistors (generation to generation) on a 4 year cycle. Here it is demonstrated another way to proceed at an equivalent rate of 2×/2years.

Computational power was benefitting by all these novel devices and computing machines were becoming more and more capable. In general, these machines were rather large and required several programmers to get the job done. Only corporations could afford to have computer for multiple users.

By 1974, few hobbyists were playing with the concept of building an oversimplified computer using off-the-shelf components that could be placed over a desktop for a single user and for this reason this simple system came to be known as a "personal computer."

1.2
The Growth of the Semiconductor Industry: An Eyewitness Report

I built my first vacuum tube radio in 1958 and my first transistor radio in 1963. By 1972, I began to operate in what came to be known as the "Silicon Valley"

experiencing the research world of Stanford University, working with Fairchild Instrument R&D Labs and also working for more than three decades with Intel Corporation. During this time, I witnessed firsthand a technological revolution that has changed the world in which we live and that is still changing it at an even more accelerated pace. In the following sections I have summarized some of the main events as they have been narrated to me by some of the actual players, as I have seen them with my own eyes, and also many other events in which I played a relevant role. More could have been said but only at the expense of making the narrative too lengthy and more complicated, so I decided to keep it simple and highlight only the main events. Hope you like it!

1.2.1 The Making of the PC Industry

Magnetic-core memory was the predominant form of storing information before the advent of integrated circuits. These memories consisted in magnetic rings (cores) through which wires were threaded to write and read information. By applying the appropriate combination of electrical signals, each core could be programmed in two different states. By means of a similar process, the status of the core could be electrically read. The construction of these memories was extremely laborious. The advent of semiconductor memories was extremely fast due to the ease of fabrication and usage.

Early on, logic gates were produced at small levels of integration and used as building blocks to realize more complex systems, but the idea of building a more complex logic circuit to produce an electronic calculator was in the air. System integrators owned the rights of the circuits they commissioned to be built by semiconductor companies, so Busicom (a Japanese company) decided to make this idea a reality and asked Intel to design a set of integrated circuits for a new line of programmable electronic calculators in 1969. This request led to the invention of the first microprocessor, the Intel 4004. Busicom owned the rights to the first microprocessor, the Intel 4004, which they created in partnership with Intel in 1970 but later on these rights were also granted to Intel.

This invention opened the way to a proliferation of pocket calculators produced by many companies. In 1986, calculators still represented an estimated 41% of the world's general-purpose hardware capacity to compute information. Computer terminals were used for time-sharing access to central computers. In the early 1970s, computers were generally large, costly systems owned by large corporations, universities, government agencies, and similar-sized institutions. End users generally did not directly interact with the machine, but instead would prepare tasks for the computer on offline equipment, such as cardpunches. In some cases it could take hours or days between submitting a job to the computing center and receiving the output.

The first 8-bit microprocessor, the 8008, capable of handling 48 instructions was announced in 1971 and the company moved from 2 to 3 in. silicon wafers for manufacturing.

The Intel 8080 was introduced in 1974 as the first truly general-purpose microprocessor. In the same year, the 2107 4 k DRAM was introduced using n-channel transistors.

In 1974, Intel had sales of $140 million with about 3100 employees.

The Mark-8 was the first microcomputer design based on the Intel 8008. The Mark-8 was introduced as a "build it yourself" project in the cover article of July 1974 issue of *Radio-Electronics*. BASIC was generally recognized as the easiest programming language to learn in 1975. It automatically converted simple English-like commands to machine language, effectively removing the programming limitations. Bill Gates and Paul Allen wrote a program that actually ran on the Intel 8080 processor that became MITS BASIC. But the inventor of the first operating system for microcomputers was Gary Kildall. In order to run his operating system on the Intellec-8, he needed a floppy controller that John Toronde was able to provide to him and with this addition the CP/M operating system became a reality. By then, there were already at least 100 small companies that were producing 8080-based computer and this meant to adapt the CP/M in each case by means of a very lengthy coding process; so Gary extracted the part of the code that interfaced with the specific computer from CP/M, reducing the magnitude of the coding task. This was called Basic Input/Output System or BIOS for short. With CP/M and BIOS, the microcomputer software architecture was once for all defined.

The first successful microcomputer product was the Apple II designed primarily by Steve Wozniak and commercialized by Steve Jobs who introduced it in 1977.

However, several personal computers were introduced in the following years, but there was no clear winning application aimed at these machines. No computer can be successful without a compelling application. But Apple II had been produced between 5 and 6 millions by the end of production in 1993.

So you may ask, why did the Apple II become so popular?

Dan Bricklin was a computer programmer who went to business school. There he learned how business planners covered the whole blackboard with columns and rows of numbers to manage their business, changing the value of one cell triggered a full recalculation of the whole spreadsheet. So he decided to automate the process and invented VisiCalc. Cost of a personal computer was about $3000, so *by chance* he and his colleague Bob Frankston were able to get a *loner* Apple II. VisiCalc was introduced in October 1979 and soon became an outstanding success with business people.

The success of the Apple II was finally noticed by IBM (International Business Machines), which decided to take some action. IBM, one of the world's largest companies, had a 62% share of the mainframe computer market in 1981. Its share of the overall computer market, however, had declined from 60% in 1970 to 32% in 1980.

So IBM began planning how to enter the personal computer business. The task was given to Bill Lowe, the lab director in the company's Boca Raton, FL, facilities. Early studies had concluded that there were not enough applications to justify acceptance on a broad basis and so he realized that it could not be done

quickly in IBM. The company had become way too big and could not move fast enough to successfully execute this project.

Don Estridge was the project manager and he decided that to meet deadlines, it was necessary to adopt tested vendor technology; a standardized, one-model product; open architecture; and outside sales channels for quick consumer market saturation. This definitely was not something IBM could do!

Therefore, he decided to acquire the processor and chipsets, the operating system, and the application software. The IBM team decided the processor had to be the most advanced 16-bit processor. It was known that Gary Kildall owned the operating system and a little known company in Seattle called Microsoft could provide the BASIC language and so IBM decided to go and buy!

I was fortunate enough to have joined Stanford for a postdoc in 1972 and after an experience at Fairchild Semiconductors R&D had joined Intel in 1978. As a reward for fixing some packaging encapsulation problems, I was offered the position of technology manager for microprocessors and static RAM in 1980. This was considered a junior manager job since the big engine of semiconductor business was DRAMs, which were already worldwide produced in the hundreds of millions (Figure 1.10), while processors were produced in the tens of thousands. My instructions were very clear: Do not spend too much R&D money, stay two generations behind DRAM so that equipment depreciation does not affect costs and do not introduce more than a couple of new process steps into any new technology generation.

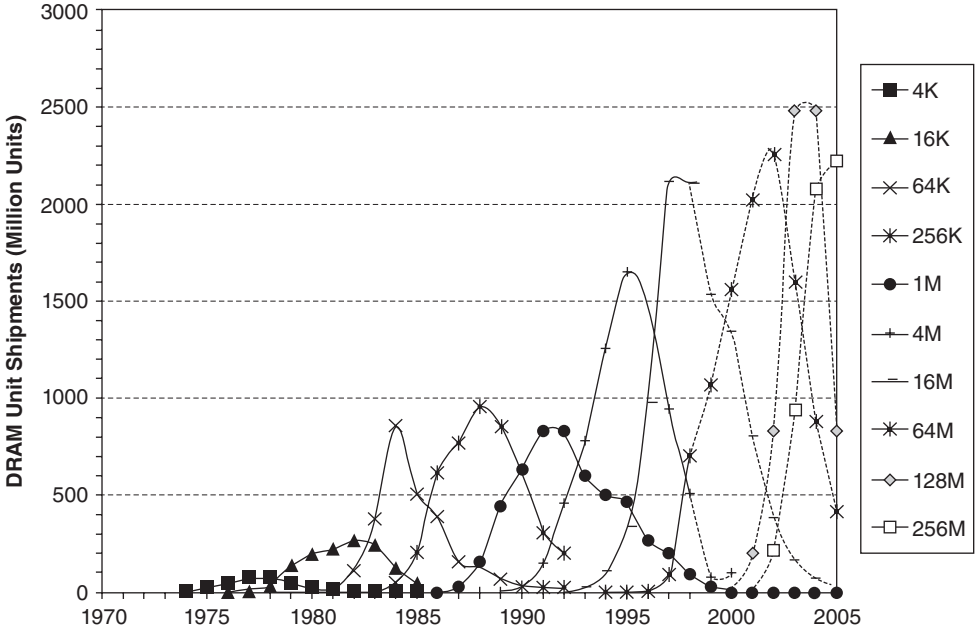


Figure 1.10 DRAM unit volume production.

In the mean time, the IBM lawyers went to meet with Gary Kildall, the owner of a popular operating system, in Pacific Grove near Santa Cruz, CA, but it was a beautiful day and even though he knew they were coming he had gone out (ah the new freedom gained with the revolution of the 1960s!) to fly his plane: no connection; so they left empty handed and went to Seattle to acquire the BASIC program from Microsoft. During the negotiations, they mentioned their failure in acquiring an operating system. At that time software was a new subject and it was not protected by copyrights and so Bill Gates bought the rights from an alternative owner of the popular operating system, renamed it MS DOS, and made "a deal" with the IBM lawyers. He would let IBM use MS DOS and Microsoft would collect a fee every time one operating system was sold.

Next came the selection of the microprocessor. Intel, Motorola, and National had just developed brand new powerful 16-bit processors that IBM wanted, but all of them lacked the chipsets that went with them. However, Intel had also introduced the 8088 processor that had an internal 16-bit architecture and bus, but with an external 8-bit bus. This meant that the chipset and the applications developed for the Intel 8080 were easily useable with the 8088. Since all the other processors were not going to have a chipset in 1981, IBM selected the 8088; overall it may not have been the best choice but processor, chip sets, operating system, and applications were ready . . . and so Microsoft and Intel got the business!

On August 12, 1981, the IBM personal computer with a price tag of $1565 was announced. Two decades earlier, an IBM computer often cost as much as $9 million, required an air-conditioned quarter-acre of space, and a staff of 60 people to keep it fully loaded with instructions. The game had been forever changed but at the time nobody really could even vaguely imagine what was going to happen in the not so distant future.

As a demonstration of the above statement, in December of that year, a small number of *us* were treated to a celebration dinner at "Au Charbartin," the fancy and most expensive French restaurant in Los Altos, no limits on wine tag! We thought this was the ultimate reward!

In the next few years, Intel indeed capitalized on this success by realizing the importance of improving the microprocessor and also devising a new method to rapidly change the pace of introduction of new microprocessors but for the time being the 80286 was introduced in 1982 and I completed the transfer into manufacturing of the 80386 in December 1985. But all the semiconductor business was soon going to change!

1.2.2 The DRAM Wars

A big recession hit in 1982. During the 1974 recession, Intel like many other semiconductor companies had been forced to lay off 30% of its employees and morale was low. During the 1981–1982 recession, instead of laying off more employees, Intel decided to reduce cost and accelerate new product

development. To accomplish these tasks, Intel asked employees to share the pain. Managers asked exempt employees to work *2 extra hours per day, without pay, for 6 months (this was called the 125% solution and we got a beer mug with this inscription for it!)*. Sales went up the following year but it did not last, and, again, instead of more layoffs, Intel imposed *pay cuts of up to 10%*. This strategy paid off and by June 1983, Intel undid all the cuts and *awarded retroactive raises*. Moreover, in December 1982, IBM paid $250 million for a 12% stake in Intel. This gave Intel much needed cash and strengthen its relationship with IBM, which would raise its stake to 20% before selling it in 1987.

These events are however just one of the consequences of major changes occurring in the semiconductor business environment at the time that led to a fundamental transformation of the whole semiconductor industry. For this reason, it is important to get into more details on this subject.

It is not possible to run a low-defect and high-volume factory without running some kind of memory in it. Once a factory is built and equipped, it is absolutely essential that capital cost be amortized. This is the fix cost of running a factory. In addition, there are costs related to personnel, power, gases, consumable materials, and so on that are related to the volume of wafers running through the factory. The total monthly cost of producing a wafer is calculated by adding all the above costs divided by the number of wafers produced in that month. Normally, during the initial years, the capital cost way exceeds the variable cost, so it is imperative to run the highest possible number of wafers to make any product economically viable. Memory products are the highest volume runners in the semiconductor industry and DRAMs were then the highest volume runners of them all. In addition, memory devices are absolutely essential to increase the yields of a factory. A memory device typically undergoes a raster scan test whereby any failing bit can be individually identified. In subsequent failure analysis, it is possible to pinpoint and analyze the failed bit and discover the reasons why it failed and take remedy action, no DRAM no avenue to high yields.

By the late 1970s, DRAM roadmap had become highly predictable. Every 3 years a new DRAM was being introduced at a memory density four times the previous generation. As explained before, the actual commercialization cycle was closer to 4 years than 3 years, but the announcement of a new technology was however loudly heralded every 3 years. Japanese companies in particular had become very efficient in systematically mapping the introduction of new DRAMs from R&D, to development pilot line, to manufacturing pilot line to high-volume manufacturing. In addition, Japanese equipment suppliers had carefully analyzed US-built equipment and had systematically developed equipment that was much better in all aspects.

It all came together in the early 1980s when a technical article pointed out that Japanese suppliers of DRAMs were outpacing US suppliers in all aspects. This announcement produced an absolute shock reaction in the semiconductor industry, but it was just the early warning of a new reality in the memory business that resulted in catastrophic implications for many US companies.

Intel, for one, had to eventually leave the semiconductor memory business. The founders were forced into that decision because it is said that Japanese competitors had developed "the 10% rule," quote a price 10% below the prices of Intel and AMD chips. The Japanese firms' market share went from under 30% in 1976 to above 50% in 1988; US companies lost the most. For Intel, memory sales fell from 90% of revenues in 1972 to about 20% in 1988. Intel's managers were dealing with a dramatic industry change, an inflection point with no obvious solution in sight.

On the technical front, another fundamental technology change was about to occur in 1985. The early DRAMs were built with PMOS transistors, but manufacturing had switched to NMOS transistors by the mid-1970s. NMOS DRAMs were many times faster than PMOS DRAMs and getting better. However, power consumption was becoming a problem not only in DRAM but also across the whole industry. The electrical solution was well known. In a 1963 conference paper, C.T. Sah and Frank Wanlass of the Fairchild R&D Laboratory showed that logic circuits combining p- and n-channel MOS transistors in a complementary symmetry circuit configuration drew close to zero power in standby mode. Wanlass patented the idea that today is called CMOS.

The question of whether CMOS implementation was beneficial was not technical but was related mainly to cost and density. Building two types of transistors required more processing steps and also in order to keep the two types of transistors isolated, it was necessary to allow extra space for isolation that negatively affected bit density. Several comparative (i.e., NMOS versus CMOS) tests were run in many companies proving that CMOS was a viable solution. In 1985, Intel had completed similar tests and reached the conclusion that conversion from NMOS to CMOS technology was a viable solution.

For this purpose, it had constructed and outfitted a completely new factory (Fab 5) capable of running 1 μm CMOS technology.

But in 1985 Intel reported a 16% loss in revenue from the previous year and essentially no net income ($1.570 million versus $198.189 million for 1984). All of these losses were due to the decline of Intel's DRAM business.

It is said that one day Andy Grove walked into Moore's office and asked the question: "Suppose this question was not related to your company but suppose somebody would ask you what you would recommend to a company that was losing money in a business without hope of turning that around?" To this it is said that Gordon simply replied: "I would recommend to shut it down!"

This announcement is in the 1985 Intel Corporation Annual Report:

> "This year we announced our decision to drop out of the dynamic random access memory business. This very competitive area has been targeted by non-US manufacturers so it is difficult to produce a return on investment required to be a participant."

The situation got even worse in 1986 when the company reported a net income loss of $173.165 million.

In the meantime, the microprocessor business was booming. So Intel decided to make a bold move and gave the brand new factory built in Oregon for DRAM (Fab 5) to the Technology development group. I was sitting at my desk when my boss Dr. Gerry Parker walked in quite excited and told me: "Grove just gave us Fab 5!" This was completely unusual and surprising.

The Fairchild experience had demonstrated to the Intel founders that the transfer from R&D to manufacturing might at times become extremely difficult if not impossible. So based on this bad experience, Moore and Noyce decided that Intel was not going to have an R&D pilot line. The R&D assignment consisted in developing new technology modules such as lithography, deposition, etch, and so on that were then integrated with the existing manufacturing lines. In this way only few new steps were introduced from one generation of technology to the next into a well-established manufacturing process and therefore this approach did not require an independent R&D wafer line.

I had indeed experienced the Fairchild problem firsthand as I had unsuccessfully tried to transfer technologies from Palo Alto (R&D) to Mountain View (manufacturing) during my staying at Fairchild R&D. Even though the distance between these two locations is only less than 10 miles, it was as if it had been 1000 miles and as if people spoke two completely different languages.

The problem consisted in the fact that the factory manager had the only and ultimate power to decide how to conduct the factory business. So he and his team felt they could change any process recipe and any equipment, as they pleased, not fully aware of the profound implications associated with these actions. It is true that the industry had developed a very robust MOS process, but it was as if we were marching a *solid but narrow path surrounded by ravines, quick sand, dense fog*, and so on; make the slightest deviation and you are completely lost. However, the manufacturing group, without the complete understanding of the consequences of their actions, made often-fundamental changes in the process flow running in the factory in Mountain View with often disastrous consequences. As a related example, one of the factory exhausts used to blow up (i.e., explode) every few weeks without any apparent reason. It was eventually discovered that both oxygen from the oxidation furnaces and hydrogen from annealing furnaces and epitaxial reactors were merged in this single exhaust. When the hydrogen reached the 4% ratio to oxygen, it became flammable. In few words, depending on which furnaces and reactors were simultaneously being utilized, eventually the hydrogen-to-oxygen ratio would reach the explosive point and boom; there goes the exhaust stack!

Going back to Intel in the 1970s, the R&D team did not have a pilot line and was compelled to work (asking favors) to manufacturing to fully process an experimental set of wafers. As part of my gift of becoming technology manager for microprocessors, I had also received a brand new clean 10,000 ft^2 room extension in Fab 3 (located in Livermore) with a lot of equipment capable of processing 100 mm wafer, not the full pilot line but at least half of it. However, the scars of the disappointing experience of Fairchild were still very real on my back and so I decided to fully engage with manufacturing to avoid a repeat of the

Fairchild negative experience. Any new engineer in my small team (17 people in total) when hired was assigned to manufacturing for 3–6 months training before he/she came on the technology development job. In addition, any wafer processing capacity not utilized by the R&D team was given to manufacturing. In a few words, the two groups were integrated as well as possible and one-by-one new technology modules transferred seamlessly from R&D to manufacturing. I was particularly proud for transferring the tungsten silicide/polysilicon gate MOS process (first in the industry) with 1.5 μm features into manufacturing with this handshake procedure between R&D and manufacturing in 1983.

1.2.3 The Introduction of New Materials

The original silicon gate process developed in the 1960s had a very clear list of materials: silicon wafers, silicon dioxide, boron and phosphorous as dopants, polysilicon as gate material, and aluminum interconnections. Polysilicon had acquired a dual function in the silicon gate process as it represented the gate electrode and also it was used for short-range interconnects.

The only variations occurred in the late 1970s consisted in the introduction of arsenic for shallow junctions. This very shallow and abrupt junction created however a new problem. The HMOS II generation (tox ~400 A) showed a brand new failure mode. Electrons were so accelerated near the drain that the most energetic (hot) electrons were able to jump into the oxide after impact with atoms near the drain region. This phenomenon caused a threshold drift that eventually stopped the transistor from properly functioning. This problem was solved by introducing a "tip implant" at the edge of the drain facing the gate that provided a smooth field transition from channel to drain that eliminated the hot electron problem. As indeed junction depth began to be reduced, it was no longer possible to utilize the deep phosphorous diffusion into contacts as a way of containing aluminum penetration and silicon was introduced into the aluminum to quench its thirst for silicon. Sputtering revealed itself as the best way of depositing aluminum/silicon as high-quality sputter target containing the correct ratio of aluminum to silicon became available. This solution was much simpler than any failed attempt of controlling dual evaporation guns. However, it was just the misalignment of one of this guns that provided a very useful technological breakthrough. Due to the misalignment of one of the electron gun, some copper from the crucible was included in the aluminum during deposition and this combination showed a remarkable resistance to electromigration.

In the early 1980s, the signal propagation delay of polysilicon lines began to be way too long and several metal silicides, due to their lower resistivity than doped polysilicon, were tried in conjunction (on top of) with polysilicon to alleviate the problem. Eventually, tungsten silicide proved the most appropriate solution. By depositing the film silicon rich, it was possible to deposit and then define both the tungsten silicide and the polysilicon underneath in a single etch step. During

the subsequent oxidation, the excess silicon in the silicide migrated to the top of the silicide forming a silicon dioxide protective film.

1.2.4
Microprocessors Introduction Cycle Goes from 4 to 2 Year

But now we are back to 1985 and Technology Development had just received a brand new 24 000 ft^2, 150 mm wafer, and 1 μm CMOS capable factory. This was real business but the next question arose: How should this be utilized? After long discussions, it was decided to gamble on the success of PC and consequently of the X86 microprocessors family.

The technology of the 80486 that was just in its infancy was transferred from Livermore (Fab 3) to the newly received Fab 5 now rebaptized Development 1 (D1). In the subsequent 3 years, a detailed plan to avoid the repetition of the DRAM demise was formed. Since the Japanese companies were much richer and bigger than Intel, the only protection of the new microprocessors jewels that we could imagine consisted in running forward faster than anybody had ever done before. This required bypassing the traditional concept of the R&D pilot line and of the premanufacturing pilot lines and instead making a direct transfer from R&D to manufacturing at high-volume manufacturing (HVM) yields. In addition, the manufacturing organization could no longer change any of the equipment or any of the process steps transferred from the R&D group in order to avoid wasting any time for the sake of re-engineering some of the process steps. This multiple steps transfer operation normally required 1–2 years during which time the manufacturing yields were substantially lower than those demonstrated by the R&D group. Most people were quite skeptical that this "skipping" methodology could ever work, but when the first technology transfer from D1 to the next manufacturing site occurred completely seamlessly and when the yields of the first wafers run in manufacturing were equal, for the first time in Intel history, to those of the D1 site, people finally believed. From then on, MPU technologies were introduced on a 2-year cycle. This technique was eventually named "copy exactly" and it is still diligently practiced nowadays.

1.2.5
The 300 mm Wafer Size Conversion

Reducing the size of individual transistors at a reasonable cost increase (~5–10%) from one technology generation to another is not the only way of reducing die cost. Wafer size conversions are also a viable means of cost reduction as long as the number of die/wafer doubles (this is due to the larger wafer area), while the manufacturing cost only increases by a modest 20–30%. This should be the rationale for any wafer size conversion, but that is not what actually happened in two most relevant cases. Both Intel and IBM led the 100/125 to 150 mm and the 150 to 200 mm wafer size conversion, in early 1980s and early 1990s, respectively. Both companies paid a premium for the development of the

equipment necessary for a full production line and also placed their leading products on the new wafer size. *However, this approach has a fundamental flaw.*

In any wafer size conversions, *all* the equipment need to be fully functional to yield a viable production line. If only one tool does not perform according to the specifications, the whole line is nonfunctional and it is shut down. Under these emergency conditions, money is spent at exponential rates to overcome these last obstacles to manufacturing. However, once all the equipment has been developed, any company can acquire it within 6–12 months and the whole production line can be quickly set up without having to pay for the equipment development cost. In both cases, Intel and IBM overall spent much more than anticipated since the introduction of one of their leading products was delayed due to some of the new equipment being not fully functional and as expected money was poured into this problem without any limits. Both companies made the same mistake and got absolutely no benefits from leading the wafer size conversion.

In 1993, the conversion to 300 mm wafer size was beginning to be at the center of many discussions. In 1995 I was part of a panel discussion at the VLSI Symposium in Kyoto where all the participants stated without any hesitation that they wanted to be the second company to convert to 300 mm manufacturing, an obvious choice! In addition, the industry had become much savvier about extravagant expenditures, so nobody wanted to be the hero leading the 300 mm wafer size conversion given the poor examples of the two companies that led the previous wafer size conversions.

The Sematech consortium had been formed in 1987 as a cooperative effort between 14 US semiconductor companies and the Defense Advanced Research Projects Agency (DARPA) to re-energize the US semiconductor industry after the demise of DRAM. After a shaky start up, this consortium had found its real purpose in rebuilding the US equipment industry that was falling from a 90% market share in the 1970s toward a projected 10% market share by the mid-1990s. Thanks to the effort of Sematech by 1993, the US equipment industry stabilized its market share around a 50% value and was able to hold from then on this market position on its own.

Leading the 300 mm conversion seemed the natural project for Sematech and indeed a working group on this subject had been formed since 1993. However, it became clear by 1994 that all the members of Sematech were using an almost even split (i.e., 50/50) of US and non-US equipment, so by mid-1994 I humbly proposed in a Sematech meeting that the formation of an independent subsidiary to evaluate equipment for 300 mm from everywhere in the world should be taken under consideration.

After about 18 months of negotiations, it was finally decided to form an independent subsidiary capable of evaluating equipment from everywhere in the world and, most of all, capable also of accepting international (i.e., non-US-based) IC companies as members.

It was clear that any funding toward this effort had to be kept completely separated from the contributions made by the Sematech members and DARPA toward US only projects.

In total three companies from Korea, three companies from Europe, one company from Taiwan, and six companies from the United States joined this effort. This kind of worldwide cooperative endeavor had never been tried before and it must be said that most people in these companies did not personally knew each other.

The first meeting of the International 300 mm Initiative (I300I) was held on April 2–3, 1996 in Munich, Germany, and I was elected Chairman of this effort mostly because at that time I was the only person in the room that everybody had met before that meeting.

Japan had also initiated a similar effort on 300 mm equipment evaluation as far back as April 1994. By June 1996, representatives of the two organizations met in Hawaii during the VLSI symposium for the first time. I must confess that from 1985 to 1995 I had spent over 2 years in Japan in installments of 2–3 weeks at a time dealing with equipment suppliers and research projects and I had learned a lot about Japanese business culture. So, I recommended at the end of the first meeting in Hawaii that we would go Karaoke. We sang and drank until late at night and after this joint experience, things really changed for the best between the two organizations. After few more meetings, the two organizations decided to develop joint equipment targets, exchange some equipment evaluations, and eventually agreed to develop a unified automated material handling approach that would become a SEMI standard. This latter goal was accomplished in the subsequent 2 years and for the first time in history, the semiconductor industry had finally agreed on a Unified Material Handling System (UMHS) and implemented a standard full factory automation approach. The first 300 mm wafer was processed in 1999 in a prototype development line by the Siemens-Motorola alliance in Dresden to produce 64 Mbit DRAM memory devices. By 2002, all the process and automation equipment was successfully deployed in high-volume manufacturing (HVM). The 300 mm wafer size conversion remains to date the most successful wafer size conversion in history as the equipment cost increase resulted in much less than 30% for a 2.25 increase in useable wafer area.

1.2.6 The 1990s: Scaling, Scaling, Scaling

In the 1990s, all the elements of scaling reached their apogee. Technologies kept on coming on a 2-year cycle as a clockwork and with them the PC business changed the corporate world, proliferated through the small business world, and finally reached the consumer world. The PC brought an increase in productivity that powered an astonishing average 17%/year growth for the semiconductor industry. Scaling according to Dennard's rules and doubling the number of transistor every 2 years according to Moore's law seemed to be bound to continue forever.

In 1991 industry, academia and national labs in the United States began collaborating together in an effort to jointly compile a technology roadmap spanning a 15 years horizon. The name selected for this effort and relative document was the National Technology Roadmap for Semiconductors (NTRS).

This first meeting was followed by the publication of three subsequent NTRS documents in 1992, 1994, and 1997. However, globalization of the semiconductor industry was making it practically impossible for a single region to generate a roadmap for the whole world. In addition, it was clear that Dennard's scaling was coming to an end and it was necessary to completely change many materials and the whole structure of the MOS transistor in the next 10 years.

Talking to Moore in 1995 made me realize the need for action even more clearly as he stated: "No exponential is forever."

Re-energized by this challenge, my first action as the newly elected Chairman of the NTRS in January 1998 consisted in proposing the internalization of the roadmap process.

The proposal was submitted to the World Semiconductor Council (WSC) representing the semiconductor industry associations of Europe, Japan, Korea, Taiwan, and the United States in April 1998. The WSC decided to give a tentative approval to this international effort with the intention of evaluating the results in a couple of years.

The first meeting of the renamed International Technology Roadmap for Semiconductors (ITRS) was held on July 11–12, 1998 in San Francisco followed in the subsequent years by spring meetings in Europe, summer meetings in the United States, and winter meetings rotating among Japan, Korea, and Taiwan.

In April 2000, I had the pleasure to report the results of the ITRS, as Chairman of this effort, to the WSC that was pleased with the results, approved this effort, and made it a permanent organization.

Since then the ITRS has produced yearly documents according to the following format: Full revisions were produced in odd years and updates, consisting in revisions of some tables as needed, were produced in even years (see www.itrs2.net).

1.2.7
Equivalent Scaling: Designers Will Never Know What We Have Done

In order to realize Moore's law, the horizontal dimensions of transistors' pitch in both x and y horizontal directions had been scaled to 70% every 2 years yielding an area reduction of 50% from one technology generation to another. Of course die size increases and design cleverness remained intrinsic elements supporting Moore's Law trend. However, the insatiable demand for performance (e.g., higher frequency of operation) had led to over-scaling to about 60% of both transistor gate length and gate oxide thickness from one technology generation to another. For instance, the graph of Figure 1.11 showed to an incredulous audience that gate oxide thickness was inevitably going to reach the monolayer level by 2005 and something needed to be done very soon.

This implied that *Dennard's scaling was coming to an end* and something completely new was required if the semiconductor industry was going to march at Moore's law pace in the next decades. In few words, what was required consisted in a new way of continuing to scale down the horizontal dimensions of the transistors while replacing one by one most of the materials and while also

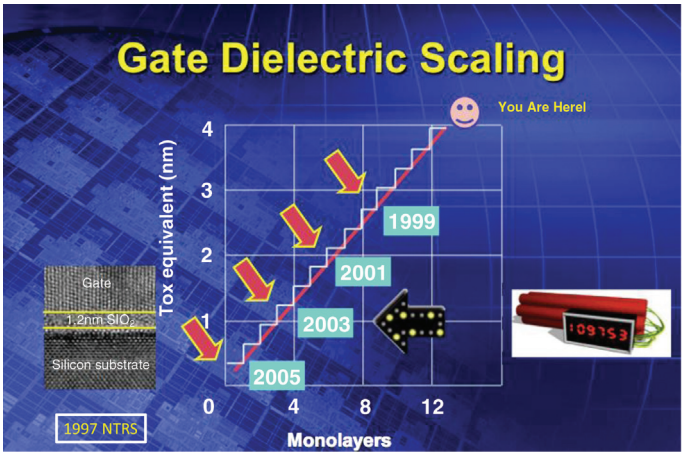


Figure 1.11 1997 NTRS: forecast of end of gate oxide scaling.

introducing new transistor structures (Figure 1.12). The new methodology was identified by the name of "Equivalent Scaling"; the goal consisted in continuously increasing transistor density and performance at historical rates without letting the design community detect that different transistor structures and materials were utilized.

In meetings held on this subject in 1997–1998, most people, remembering the difficulties of engineering the very first transistors, called for the formation of a new "Bell Labs" organization to address the problem. However, it was soon realized that no association was capable of collecting the $1 billion required to fund this kind of effort. After several meetings, it was decided that the only practical way of addressing this fundamental, highly scientific problems, consisted in the formation of clusters of universities addressing these problems. It was decided also that this organization was going to be funded by the members of the

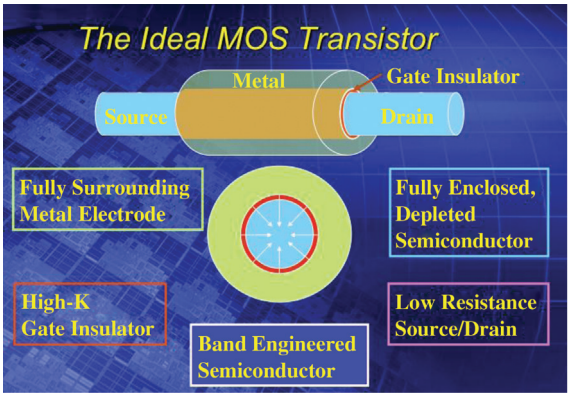


Figure 1.12 1998 ITRS: Equivalent Scaling vision of required transistor innovation.

Semiconductor Industry Association (SIA) and once again DARPA agreed to become a key member. This effort was indeed launched in 1998 and named Focus Center Research Program (FCRP). The goal of the program consisted in researching the items identified by the ITRS with a 3–7 year outlook. Similar efforts existed also in Europe under the program called ESPRIT.

Fortunately, the global outreach of the ITRS triggered many more research programs around the world that began addressing the challenges of completely re-engineering the transistor. The first result came from research initiated in the early 1990s at Stanford University by Prof. Judy Hoyt and supported also by Dr. Shinichi Takagi. This research related the effect of stress on carrier mobility. It had been demonstrated that compressive stress enhanced holes mobility while tensile stress enhanced electron mobility. This effect had never been considered in Dennard's scaling. The practical implementation of this concept was introduced into manufacturing in 2003.

Even Moore was impressed by this result and by the research activities flourishing around the world and this led him to readjust his views on the future as demonstrated by the title of his 2003 ISSCC keynote address:

No Exponential is Forever: But "Forever" Can Be Delayed!

The next challenges were however much more difficult since first of all it required the replacement of the gate oxide with a new material with a much higher dielectric constant. In essence, such material needed to behave, from an electrical point of view, as if the gate oxide thickness had been reduced by the ratio of the dielectric constants, while in reality the thickness of the new dielectric was actually increased; any increase in dielectric thickness implied also a reduced gate leakage. This fully departed from Dennard's scaling.

SRC (Silicon Research Corporation) and Sematech carried out the initial research effort on a new dielectric material, while the FCRP concentrated on solving the problems of interconnect lines. The new interconnect lines came into production between 1999 and 2001. The solution consisted in four elements. First, copper (resistivity $\sim 1.8\,\mu\Omega\,\text{cm}$) replaced aluminum (resistivity $\sim 2.7\,\mu\Omega\,\text{cm}$) due to its lower resistivity. Several contamination problems were solved in order to prevent copper from reaching the silicon substrate with deleterious effects on electrical performance (e.g., increased leakage). Second, new intermetal dielectric began to replace the traditionally deposited silicon dioxide. This replacement resulted more difficult than expected due to the poor mechanical properties of the new materials. Third, it was also recommended that interconnect lines were made as short as possible to prevent excessive signal degradation. Finally, one more layer of metal interconnections began to be introduced with each technology generation. This requirement has led to the introduction of more than 10 layers of interconnections and the number is still growing.

In the meantime, research on new transistor's materials was extensively conducted around the world and several scientific meetings were held to report progress, if any, toward solving the challenge posed by the fact that the thickness of the gate dielectric was soon reaching the value zero if traditional

scaling methodology was followed. All the researchers had agreed that the simplest solution consisted in experimenting with several materials with dielectric constant substantially larger than 5 since this was the best value that could be achieved with the slightly nitride-rich gate oxide in production at the time. By using one of such materials, the physical thickness of the gate dielectric would have increased by as much as the ratio of the dielectric constants, while the "Equivalent Oxide Thickness" (EOT) would have appeared to be reduced as if the traditional gate dielectric was still scaling down according to the historical rates. Several materials were evaluated, but it was soon clear that the solution to this problem was more difficult than expected. The same fixed dielectric charges that were responsible for the higher dielectric constant were also negatively affecting the mobility of the charges moving in the underlying channel. Results reported at IEDM around the year 2000 and similar conferences indicated that if any of these new materials were used as gate dielectric, the charge mobility in the channel would be reduced by as much as 50%, no gain at all!

It has been reported in so many cases that when searching for a solution to a problem, the winning solution may came out under purely fortuitous circumstances and so it was indeed the case for the solution of this problem as well.

In February 1999, I attended the European Industry Strategic Symposium (ISS) in Rome and went out for dinner with a good friend of mine, Dr. Gilbert Declerck. He and I had met in 1973 when we were both researchers at Stanford University in California. After that time, he had gone back to Belgium where he had become the COO of IMEC (Interuniversity Micro Electronics Center). He was later on appointed to the position of CEO of IMEC in June 1999. We decided to go out for dinner to catch up with events occurred since last time we had seen each other and selected "La Lupa," a typical Roman restaurant for our reunion. The she-wolf, that is what it means, is the symbol of Rome at it is said that she fed the orphan founders of Rome that got their character from her; could it be that eating there was going to have a positive influence on Gilbert and me?

During that dinner, I learned that actually IMEC had equipment capable of depositing dielectrics with high dielectric constant. By July 2000, I was able to promote a three-way cooperation between SRC-Sematech-IMEC that accelerated the progress in the search for the best replacement material.

It was eventually demonstrated at IMEC, under the leadership of Dr. Marc Heyns, that by replacing the silicon dioxide with hafnium oxide, whose relative dielectric constant ranges in the 15–20 values, it was possible to make functional transistors. In order to ensure good adhesion to the underlying silicon, it was still necessary to keep a monolayer of silicon dioxide in the structure. In addition, it was demonstrated that it was also necessary to introduce a metal gate material as replacement of polysilicon. It turned out that the combination of the overlaying metal gate and the underlying hafnium oxide mutually compensated any negative effect on charge mobility in the channel. This new transistor was

successfully introduced into manufacturing in 2007. This transistor demonstrated good electrical performance and reduced gate leakage.

After over 40 years, two of the materials that had marked the success of MOS (i.e., silicon dioxide and polysilicon) had been successfully replaced.

The distributed research model had produced in about 10 years the winning solution!

This success had also a fundamental effect on the future of IMEC as Gilbert decided to create a new broad research program centered on advanced transistor development ahead of industry needs that was named "Core Partners Program." Based on the accomplished results, I was so convinced of the validity of this program that I was the first Core Partner to subscribe in 2003 to this program on behalf of Intel Corporation.

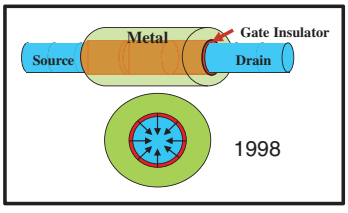
But the renovation process of fully implementing Equivalent Scaling was not over yet and one more change needed to be introduced to allow Moore's law to continue on. Leakage problems, this time related to drain to source direct interaction, were dominant when circuits were in the "off-state." In essence, electric field lines originated from the drain were reaching into the source region, more so as the channel length kept on being reduced, causing some level of current flowing from source to drain even when the gate voltage lowered to few millivolts. In essence, transistors were not completely shut off but kept on functioning in a subthreshold regime. To minimize this effect, it was necessary to deplete the semiconductor region of charges to essentially eliminate this "drain outreach" and make the electric field generated by the gate the dominant controlling factor over the source-to-drain current in the off-mode as well. The solution from an electrostatic point of view required thinning down the semiconductor as much as possible and also surrounding it with the gate electrode to maximize the effect of the metal gate potential over the effect of the drain in controlling the charges in the semiconductor. The ideal fully integrated structure had been outlined in the initial 1998 ITRS proposal (Figure 1.12), but it was now time to find a cost-effective solution. Indeed, the use of a very thin layer of silicon over a layer of insulator (SOI) had been previously proposed as a possible solution to depleting the channel region of unwanted charges, but this approach implied the adoption of an expensive starting material and still the depletion of charges was limited to the control of the gate electrode that operated only from the top side of the transistor structure.

The team at the University of California Berkeley (UCB) provided a much better solution. They suggested the formation of a "thin and tall" semiconductor structure that could be easily surrounded on three sides by the metal gate electrode. The shape of this structure suggested the term FIN-FET as an intuitively natural name. This structure provided also a better packing density as the transistor was essentially turned on a side. The need for an expensive starting material was in this way completely eliminated. This technology was introduced into manufacturing in 2011 (Figure 1.13).

This last accomplishment was finally allowing to *fully* reply with actual results, this time encompassing all the elements of Equivalent Scaling, to Moore's

Incubation Time

	Early Invention	Focused Research	Introduction Manufacturing
Strained Silicon	1992	1998	2003
HKMG	1996	1998	2007
Raised S/D	1993	1998	2009
MultiGates	1988	2000	2011



< 11 years

Figure 1.13 Mapping, planning, and introducing into manufacturing of Equivalent Scaling.

concern expressed in 1995 about the impossibility of an exponential trend to be continuing forever with a resounding: *But Forever Can Be Delayed.*

1.2.8 Is There Life Beyond the Limits of CMOS and of Von Neumann Architecture?

The experiment of using the ITRS as a means of identifying challenges and possible solutions in conjunction with the distributed research model supported by universities, consortia, and research organizations successfully demonstrated outstanding results and proved itself as a powerful method to solve the problems created by the end of Dennard's scaling. Equivalent Scaling, launched in 1997–1998, has supported the growth of the semiconductor industry since 2003 and it will continue to do so well into the next decade (i.e., beyond 2020). The ITRS, University, Consortia Research, and finally Industry cascade method has proven that it is possible to solve the hardest problems as long as *comprehensive* research is initiated 10–12 years ahead of the planned introduction into manufacturing of a new class of materials and new structural solutions (Figure 1.13). However, even though the Equivalent Scaling Method had infused new life into CMOS technology, other problems are now looming on the horizon. Assuming that it is still cost-effective to introduce new technologies at a rate close to the 2-year cycle, it is clear that by 2020–2025 the transistor dimensions will approach the few nanometers range and further reduction of transistor dimensions will reach fundamental physical limits. This emergency is similar to the alarm voiced in the 1997–1998 time frame about the vanishing of the gate oxide by the middle of the subsequent decade but this time there is even more to worry about. This time the problem is compounded by the problem the computer industry is already facing.

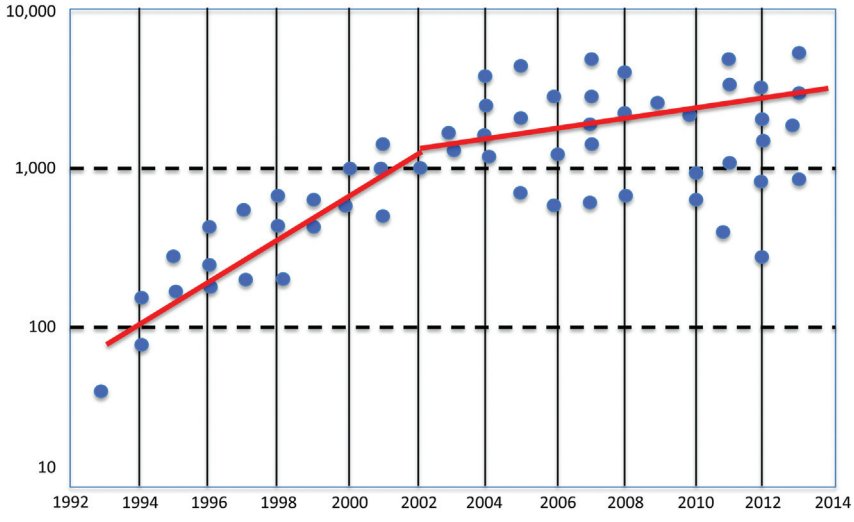


Figure 1.14 Clock frequency trend.

The computer architecture published in 1945 by Von Neumann began showing severe performance limits around 2001–2005. For more than 20 years, the performance of microprocessors had improved by operating the microprocessor at a continually higher frequency and as a result the performance of computers spanning from PC to supercomputers had kept on improving, as simple as that. The operational frequency had gone from the few megahertz of the 1980s to a few gigahertz at the turn of the century and consequently the number of executable operations per second (MIPS) had increased. Of course, many improvements in architecture and software programming had proceeded hand in hand with the technological improvements. During the same period of time, the number of transistors used by microprocessors had continuously increased in accordance with Moore's law spanning from tens of thousands to billions. However, power consumption of any MOS transistor is directly proportional to operational frequency and this fact cannot be changed. Finally, it was bound to happen that the product of the number of transistors times the power consumed by each transistor reached the ~115–130 W level in 2001 and power dissipation became a practical wall. Operational frequency could not be increased anymore due to power limitations and processor performance stalled (Figures 1.14 and 1.15). The simplistic architectural solution consisted in dividing the CPU in multiple smaller CPUs, each operating at a much lower frequency (parallel processing) and then combining the results of each unit with the others to enhance the output rate. Great idea but this multicore solution assumed that any problem could be partitioned in multiple parallel subsets. Even though this approach worked in several cases, there were also many cases where the result of an operation gated the beginning of another operation and as a result the two operations could not be processed in parallel.

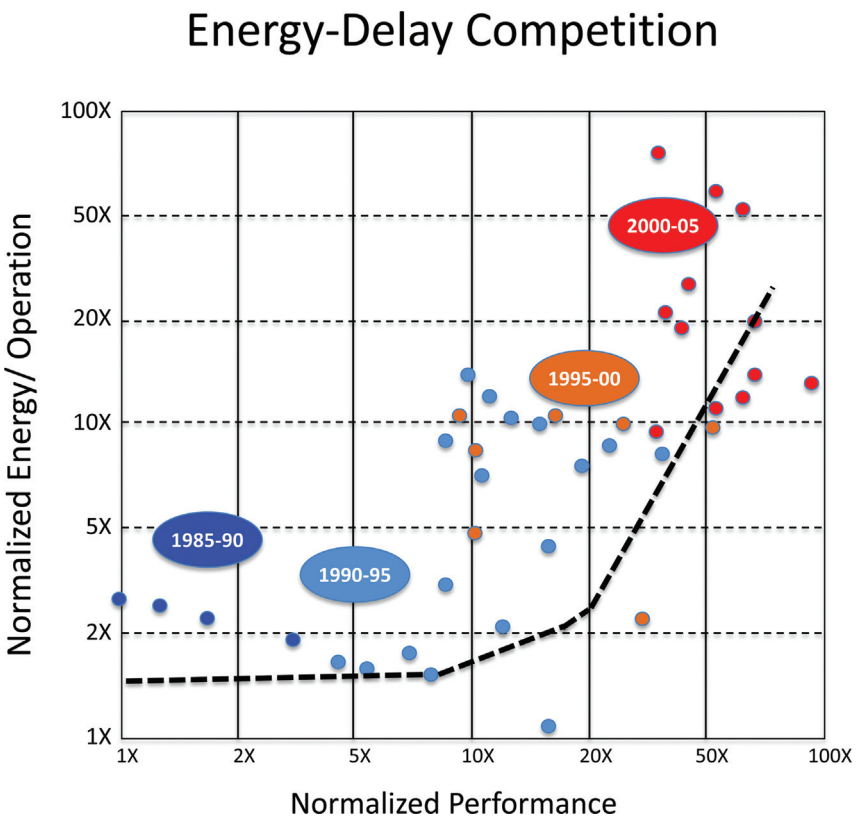


Figure 1.15 Rapid escalation in energy/operation is required for any improvement in performance.

1.2.9 Nanoelectronics to the Rescue

In January 2000, it was announced in the United States that investments in nanotechnology (~$270 million) were going to be doubled in the following year; this was going to occur under the National Nanotechnology Initiative (NNI).

> "My budget supports a major new National Nanotechnology Initiative, worth $500 million . . . the ability to manipulate matter at the atomic and molecular level. Imagine the possibilities: materials with ten times the strength of steel and only a small fraction of the weight – shrinking all the information housed at the Library of Congress into a device the size of a sugar cube – detecting cancerous tumors when they are only a few cells in size. Some of our research goals may take 20 or more years to achieve, but that is precisely why there is an important role for the federal government."
>
> *—President William J. Clinton, January 21, 2000, California Institute of Technology*

This initiative consisted in a coalition of multiple US government organizations headed by Dr. Mike Roco of the National Science Foundation (NSF). This was a great opportunity to leverage large amounts of research funds toward a new transistor concept. By 2003, I was able to organize a meeting of Academia, Government, and Industry representatives that outlined a joint strategy on how to explore a new class of phenomena occurring at the nanometer scale and use them for practical applications (e.g., discovery of a new transistor switch) for the benefit of the electronics industry.

By 2005, the search for a "new switch" that would eventually replace the transistor was supported by several members of the Semiconductor Industry Association (SIA) and jointly with NSF the Nanoelectronics Research Initiative (NRI) was launched. I initiated the first university cluster called Western Institute of Nanoelectronics (WIN) immediately followed by three more initiatives across the United States. Shortly after NIST (National Institute of Standards and Technology) also joined the NRI organization. Again, similar initiatives flourished around the world.

Following the ITRS concept about sharing information across the world, Mike Roco and I also initiated an international conference called International Nanotechnology Conference on Communication and Cooperation (INC) that was immediately accepted and supported by Europe and Japan. In this conference, regional programs from several parts of the world are reviewed (www.incnano.net).

The effort of NRI in the first 5 years consisted in exploring any kind of possible device that could operate as a logic or memory element. For this purpose, an extensive search for a new "switch" was initiated across the world under very broad guidelines. Multiple new types of switches were identified by 2010 and substantial experimental results were reported in 2015. Early on, the new desirable properties of the new "switches" were classified into three categories:

1) *The new devices were required to function in at least two separate logic states but the more, the better.*
2) *It was desirable that the devices could operate at speed comparable to CMOS with minimal power consumption (i.e., consuming less power than CMOS devices) and with essentially no power consumption at all in their standby mode.*
3) *It was also desirable that the devices had the ability to retain memory information with essentially no power consumption (i.e., retention even without a power supply).*

Based on past experience, few fundamental modes of physical operation were identified. In one case the devices still relied on the flow of electric charge, in another case the devices relied on magnetic properties and finally some devices would operate in a completely new way. While the first mode of operation could still be associated with charges flowing from one location to another, the second mode of operation was associated with stationary magnetic dipoles that did not consume any power in their standby condition. It was quickly realized that even

though it was well known that electrons carried a negative charge, it was also known that they also carried a magnetic dipole (spin) associated with them, although this latter property had never been used for construction of commercial integrated circuits.

In 2010, an extensive review of progress on novel devices and possible applications was published in the *Proceedings of IEEE* (vol. 98, No. 12, December 2010). Among these devices, the *tunnel transistor* appeared as a leading candidate with respect to operation at very low voltage. The operation can be explained in very simple terms in the following way. In a typical NMOS transistor, the charges in the conduction band of the source region are prevented from flowing to the drain region by the potential barrier generated by the NP junction existing between the source and the body of the transistor. Application of positive voltage to the gate lowers this barrier and lets electrons flow to the drain. However, the electrons in the source region have an energy distribution that spreads to values exceeding the potential barrier even when there is no voltage applied to the gate. Under these conditions, some level of leakage cannot be eliminated. Conceptually, it is possible to completely eliminate the effect of the tail of the electron energy distribution by deciding to operate via the electrons located in the valence band of the source region. Under these conditions, charges would not be able to overcome the combination of the bandgap and junction barrier voltage when operating in a standby mode! While it is true that any leakage from source to drain could be eliminated under these conditions, it is also true that in order to allow any current at all to flow from the source to the drain region it would be necessary to somehow apply a voltage that would overcome the combination of bandgap and p–n junction potentials. So, no leakage but also no current could flow under normal voltage operations.

However, Prof. Leo Esaki had demonstrated that when two energy bands are brought into very close proximity, charges could flow from one band to the other by tunneling though this band-to-band potential barrier. In 1973, he received a Nobel Prize for this invention.

Further improvements in the practical implementation of this concept for transistor fabrication have led to very promising results in recent years. It has been demonstrated that, under proper conditions, flow of current can occur at very low gate voltage and leakage current is practically negligible in the off-state. Subthreshold slopes below 30 mV/Dec have been demonstrated. The research effort is now concentrated on selecting the appropriate materials. However, while essentially steep subthreshold transitions from the off-state to the on-state have been observed when TFETs were operated at low temperatures (i.e., 28 mV/Dec at 77 K), it has also been reported that leakage increased as operation approached the temperature of 300 K. This effect is due to the existence of unwanted electronic states that allow current to flow between the two bands. The problem to be solved is not different from the problem of eliminating surface states in early MOS that scientists had to deal with for over 20 years. This time the "states" are buried somewhere between the two bands across which the charges have to tunnel. Among the many proposed solutions, the use of 2D

material seems very promising since their very own structure eliminates one set of undesirable "dangling bonds."

Several devices utilizing the spin property of electrons have also shown promising results.

In particular, the concept of operating a device with current (dynamic mode) but then storing the result by means of a magnetic state (static mode) has been demonstrated. Spin-transfer torque (STTM) is an effect in which the orientation of a magnetic layer in a magnetic tunnel junction or spin valve can be modified using a spin-polarized current. Once the magnet is polarized, no current is required to keep the magnet in this state. In essence, this device embodies and takes advantage of both the current-carrying properties of electrons and their ability of transferring magnetic information to create a permanent state in a magnetic layer. In summary, STTM offers the ability to electrically program a magnetic memory that can then permanently store information without using any energy.

These few items outlined above are just few examples indicating that by 2020 several new devices will be available to work in conjunction with or better than CMOS on some specific applications (Figure 1.16).

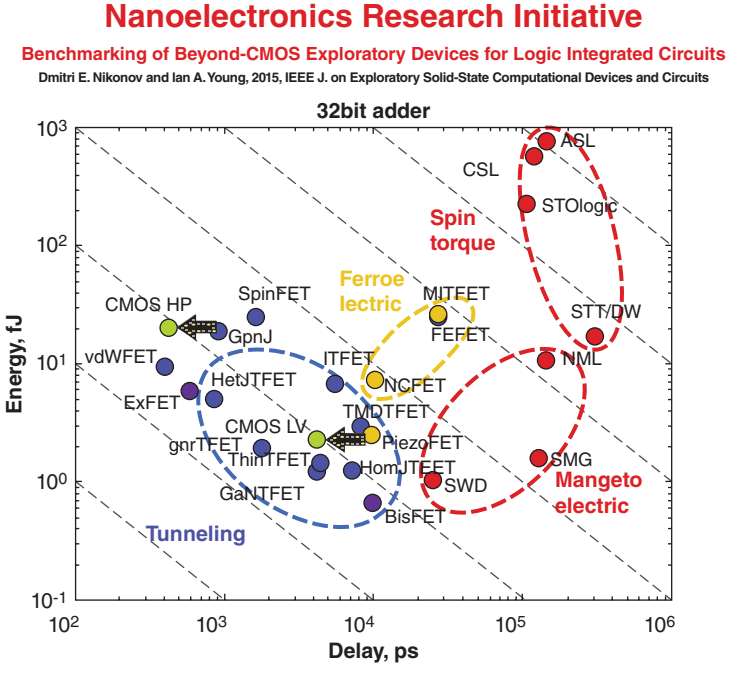


Figure 1.16 Switching energy versus delay of a 32-bit adder.

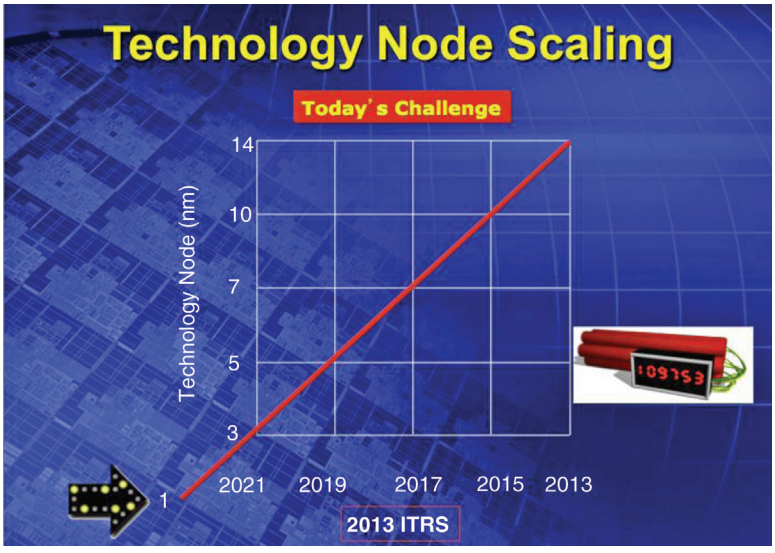


Figure 1.17 Physical limits of 2D scaling will be reached in the next decade.

1.2.10 The New Manhattan Project

The previous section should give the reader some confidence that several new very promising devices capable of operating better than CMOS in *very specific applications* are under evaluation and will likely be available to complement or replace CMOS in the 2020–2025 time frame. However, another challenge is looming in the not too distant future.

The invention of the planar fabrication process led to the invention of the integrated circuit and this invention has been and remains at the very heart of the electronics industry. A multitude of extremely accurate technological processes have allowed scaling down the features of integrated circuits from the tens of micrometers of the late 1960s to the few tens of nanometers of nowadays. By 2020–2025, device features will be reduced to a few nanometers and it will become practically impossible to reduce device dimensions any further (Figure 1.17). At first sight this consideration seems to prelude to the unavoidable end of the integrated circuit era, but once again the creativity of scientists and engineers has devised a method "to snatch victory from the jaws of defeat." The basic concept of this solution is actually rather obvious if we only observe places like Manhattan, Tokyo, Seoul, or Hong Kong; once real estate space was fully utilized, people discovered the unexplored space of the vertical dimension and began building skyscrapers.

Even though producers of logic integrated circuits have been the most prominent champions of Moore's law, it is also true that producers of memory devices have been the leaders in production of integrated circuits with the highest

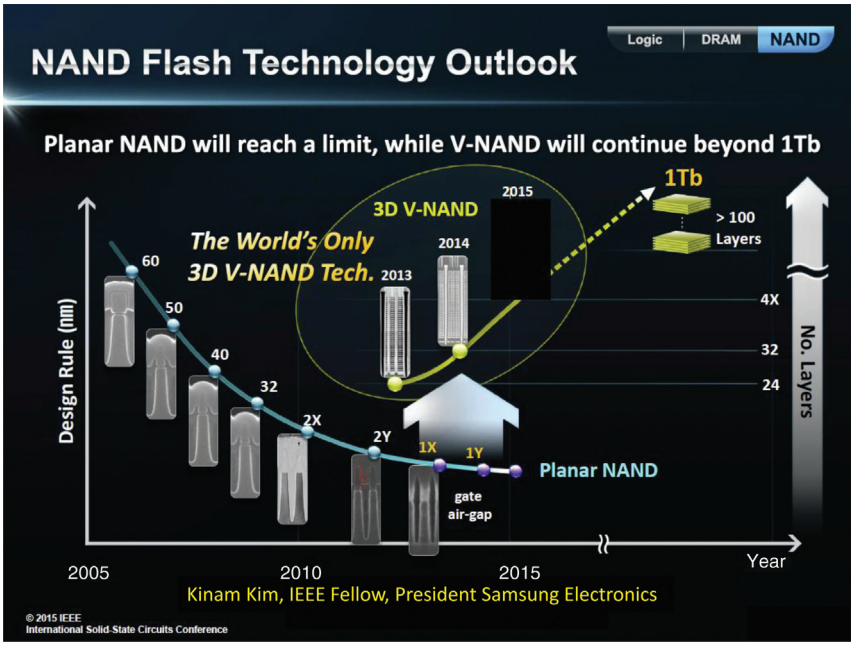


Figure 1.18 Evolution of NAND products from planar to fully vertical cell architecture.

number of transistors, the tightest tolerances and smallest dimensions. DRAM producers already dealt with space problems as far back as the "megabit" memory generation. In order to build capacitors back then with enough storage capacity, it would have required more silicon area than the area needed to build the transistors and therefore DRAM producers adopted stack capacitor or trench capacitor solutions to take advantage of the vertical dimension either above or below the surface of the silicon. Nowadays, Flash memory producers are facing a similar problem since they are running out of horizontal space, the cost of producing integrated memory circuits of small dimensions keeps on rising, and the number of stored electrons in the floating gate keeps on decreasing. To eliminate these problems, Flash memory producers have already demonstrated and announced several new products that stack multiple layers of memory on top of each other in a single integrated circuit. As many as 32 and 48 layers of Flash memory have been reported. Flash memory devices constituted by more than 100 layers have been predicted (Figure 1.18). These devices offer multiple benefits.

First, the dimensions of the critical features can be substantially relaxed since it is no longer necessary to aim for the smallest possible horizontal dimensions to reach the desired memory density. This leads to a reduced mask count and a substantially reduced cost of lithographic investments.

Second, larger vertically organized Flash cells can accommodate more charge, thereby making the device easier to operate and most likely more reliable.

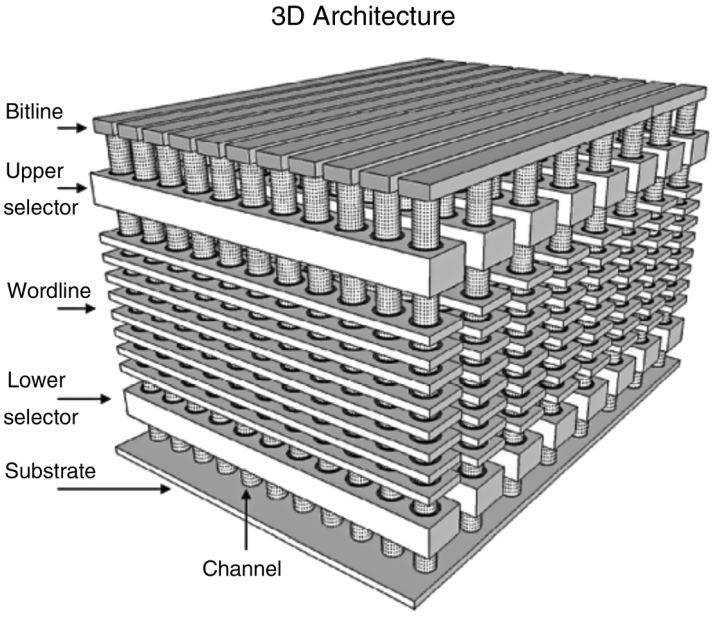


Figure 1.19 Fundamental building blocks of future integrated circuits.

Finally, memory producers are once again paving the way for logic producers who continuously crave for memory proximity to the logic to improve performance. Multiple layers of memory right above the logic part of the circuit can provide the opportunity of reducing the time signals spend traveling in interconnect lines and transferring back and forth between logic and memory circuits.

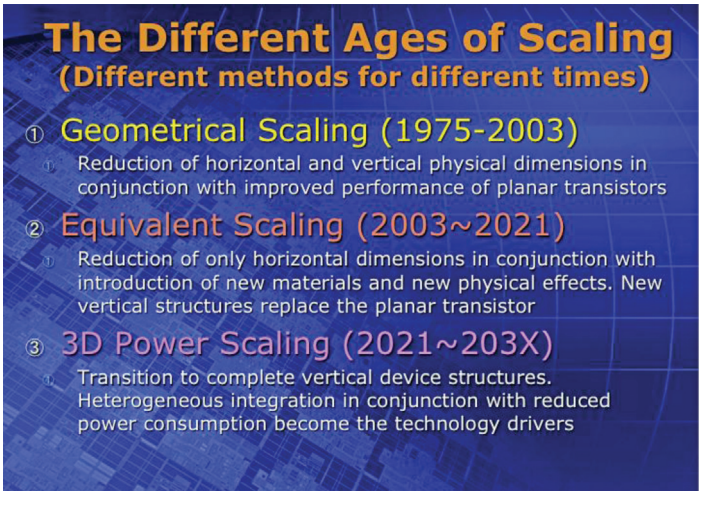


Figure 1.20 1975-203X: the three ages of scaling.

In any case, the architecture of future ICs, whether memory or processor types, will evolve beyond 2020 to a type of structure as shown in Figure 1.19.

It is clear that this new phase of scaling, capable of adding multiple layers of devices in a 3D structure, can potentially substantially accelerate Moore's law. Power reduction is also the other dominant requirement for any future technology. In summary, the new era of scaling is named "3D Power Scaling" (Figure 1.20).

1.2.11 System Requirements and Heterogeneous Integration

The semiconductor industry was created in the late 1960s with three main business drivers in mind:

1) To produce cost-effective memory storage devices. Manually assembled magnetic memories were the incumbent products to beat.
2) To produce simple standard logic gates. These parts were used as building blocks of more complex systems.
3) To produce custom parts for system houses. Concept and design rights belonged exclusively to the system houses.

It is not by accident that the first high-volume commercial products were memory devices capable of storing 64 bits, 256 bits, and eventually 1024 bits of information. The growing computer industry was the first enthusiastic adopter of these memory devices. Memory products needed to be produced in large quantities and justified additional investments were required in increasing manufacturing capacity. Memory products were the early engines of growth of the semiconductor industry.

Operational amplifiers and "flip-flops" represented families of standard products that were readily utilized by system producers to build more and more complex systems, saving money and saving space in products.

Production of custom parts began in the early 1970s. System houses building calculators and early versions of "on-button-cameras" were among the early adopters of integrated circuits. In these cases, the semiconductor companies operated as foundries and did not gain any intellectual rights on the products produced for system companies, neither had they any real influence on product specifications.

This business environment changed with the introduction of the personal computer since the semiconductor and the software companies were able to control for the first time the intellectual rights of their products. In retrospective, this unique situation was more the result of a complete underestimation of the power of the PC market by the system house(s) than the results of a specific strategy developed by the semiconductor and software producers. In essence, the PC fell into their laps and they ran as fast as they could with it. As the selling price of personal computers continued to decline and as the number of applications continued to increase, eventually the electronic industry reached the consumers.

If the system houses could be considered victims of their complacency in the way they approached the PC business, it is also true that semiconductor and software

companies fell into a similar complacency situation by underestimating the power of the combination of design (fabless) and foundry (design-less) manufactures.

By the beginning of the century, it became clear that a system house could design their own integrated circuits and have it produced by a foundry and still keeping full control of the product architecture while also retaining all the intellectual properties. Under these conditions, it was then worth asking the question: *Who needs to deal with an integrated device manufacturer?* Since system producers did not need to operate under the "PC conditions and restrictions," multiple innovative products began to emerge. MP3 players, smartphones, and tablets were born under the exclusive control of fabless system manufacturers! System houses had once again reconquered full control of their products!

Since the late 1990s, the vast technological arsenal of the semiconductor industry had also proliferated to other industries and enabled a large variety of new sensors and actuators identified as microelectromechanical systems (MEMS). The combination of sensors, displays, antennas, and radios heterogeneously integrated with more traditional integrated circuits allowed the realization of a multitude of very compact revolutionary mobile systems. These revolutionary systems could only be built by integrating in a heterogeneous way multiple dissimilar technologies.

If we also add into the equation the pervasive proliferation of the Internet plus the acquired mobility of small form-factor device plus the ubiquitous connectivity enabled by extensive coverage of wireless communications, it becomes clear that innovative systems are nowadays completely in control of the growth of the market of the electronics industry and, most of all, are *dictating the specifications of all the individual building blocks.*

1.2.12 Evolve or Become Irrelevant

To reflect and adapt to this new environment, the ITRS underwent a dramatic transformation in the 2014-2015 timeframe symbolized by operating under the name of ITRS 2.0 (www.itrs2.net) and emerging in 2016 as the new *International Roadmap for Devices and Systems* (IRDS).

Seven Focus Teams that cover top-down and bottom-up requirements and possible solutions of the new electronics industry constitute the IRDS.

The top-down Focus Teams cover system integration, heterogeneous integration, heterogeneous components, and outside system connectivity requirements and possible solutions.

The bottom-up Focus Teams cover requirements and possible solutions for CMOS to the limit called More Moore, post CMOS, and Factory Integration.

As outlined in the previous paragraphs, the improvements in performance of computers in the 1980s and 1990s had occurred by and large as a result of the availability of larger numbers of cost-effective transistors operating at higher frequency.

Performance of computing systems had all along taken advantage of this increasing frequency trend and adopted pipelined superscalar microarchitectures. These superscalar microarchitectures were enabling higher frequencies though

deeper pipelines. This meant more instructions needed to be "in flight" than was possible by waiting for branch instructions to execute. This led to *speculative execution*: predicting what path a program would take and then doing that work ahead of time, in parallel. Thus, higher frequencies meant deeper pipelines, which in turn required more and more speculatively executing instructions. But no prediction is 100% accurate. Invariably, these microprocessors did a lot of extra, wasted work by miss-speculation. The deeper the pipeline, the more the power wasted on these phantom instructions.

But performance was the name of the game and operating frequency kept on increasing with each new technology generation through the 1990s, thus encouraging the development of these complex and speculative architectural and software operations until it eventually happened, the processor power consumption reached and exceeded the 100–130 W operating range!

Under these conditions, the only practical way to prevent a "melt down" of the processor consisted in limiting the operating frequency to few gigahertz. This frequency wall has since then set up an upper limit on computing performance that has practically stalled for the past 10 years.

In order to resolve this computing performance impasse, two IEEE Fellows, Dr. Tom Conte and Dr. Elie Track, formed the IEEE Rebooting Computing Initiative (RCI) in 2013 with the goal of identifying solutions to the architectural problem of the computing industry. Several workshops were held in the 2013–2015 time frame that have identified possible architectures.

http://rebootingcomputing.ieee.org/

It is clear that any solution to overcome the computing wall hinges on the development of new computing architectures that must keep into account the practical limits at which CMOS transistors can operate under the power constraints previously outlined. In summary, the number of available transistors has continued and will continue to increase at historical rates, but most of them will have to sit idle waiting for a call to action to prevent the microprocessor from malfunctioning. Some people have called these idle transistors "Dark Silicon." In fact, nowadays many special circuits are typically spread throughout the die monitoring the power consumption of specific areas. If the temperature of any of these critical locations rises above preset values, these control circuits automatically reduce the operating frequency bringing the local temperature down to a safe range.

It is clear that the above considerations have had a profound influence on the way transistor performance is optimized. CMOS transistors have indeed been evolving under different guidelines than in the past for the past 10 years. Reduction in power consumption has become the major driver as opposed to the past when performance at any (power) cost was the overwhelming driver. As a result of these considerations, it is now clear why initiatives like NRI and other similar initiatives around the world were early on aimed at new classes of power-efficient "transistors." All the new devices under study must be capable of operating at lower power levels than CMOS. It is forecasted that around 2020, these new devices will begin to operate in conjunction with CMOS and will eventually replace CMOS transistors.

Under these conditions, it became clear in 2014–2015 that ITRS 2.0 and IEEE RCI needed to work closely together in order to avoid a major disconnect between the outcomes of the two organizations. *What if in the end the new architectures did not work with the new devices?*

For this reason, the two organizations entered into a cooperative effort agreement in February 2015 and since then have held several joint workshops.

On July 29, 2015, the Office of the White House announced the National Strategic Computing Initiative (NSCI):

http://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

One of the goals of this initiative is directly related to what IEEE RCI and ITRS 2.0

(IRDS from 2016) have already initiated

> "Establish hardware technology for future HPC systems"
>
> Computer system performance has increased at a steady rate over the past 70 years largely through improvements in the underlying hardware technology, but semiconductor technology is reaching scaling limits. There are many possible successors to current semiconductor technology, but none that are close be being ready for deployment.
>
> A comprehensive research program is required to ensure continued improvements in HPC performance beyond the next decade. *The Government must sustain fundamental, precompetitive research on future hardware technology to ensure ongoing improvements in high performance computing*."

Hopefully, this initiative will trigger similar efforts around the world!

1.2.13 Bringing It all Together

The discovery of the transistor effect and the invention of the integrated circuit that occurred between 1947 and 1959 revolutionized the electronics industry and forever changed the way we live.

The sequential introduction of new revolutionary and scalable technologies has supported the continuation of Moore's law since 1965. The density of cost-effective transistors has continued to increase at an exponential rate for more than 50 years. Technologies have come and gone (i.e., bipolar, PMOS, NMOS, Dennard's CMOS, etc.), but a new technology has always taken over where the previous one had reached its limits becoming the new supporter of Moore's law. This point seems to be the hardest to understand for many people who continue to confuse the passing of the baton from one technology to the next with the end of Moore's law. About every 10 years, alarming articles have regularly appeared predicting the imminent end of Moore's law, but no retraction has

ever been published after it became absolutely clear that a new technology was moving Moore's law forward at historical rates.

May be the best way to eliminate these recurrent misunderstandings consists in capturing in a single phrase the essence of these repeating events. This goal could be accomplished by using a well-known and familiar saying and morphing it to epitomize the most relevant historical trend of the semiconductor industry as follows:

> "Moore's Law is dead, long live Moore's Law!."

All these technologies have allowed electronics products built with integrated circuits to overtake any incumbent products or alternative technologies. New product capabilities have in turn created new markets that nobody could have predicted and changed the life of billions of people forever. How could anybody operate in any modern business without smartphones, tablets, wireless connectivity, and a variety of displays?

The Von Neumann architecture benefitted from the availability of cheap and numberless transistors operating at higher and higher frequency for over 40 years but eventually power consumption of microprocessors exceeded practical limits and computing performance reached a wall that has existed since the middle of the previous decade (~2005). Research for new "switches" began in 2005 and it has already provided several possible candidates that could operate in conjunction with CMOS and will be eventually capable of replacing it.

In the past few years, new organizations such as IEEE RCI have proposed new computing architectures and by closely cooperating with initiatives like IRDS, aimed at both promoting new devices and how they interact with systems; they are jointly laying out the foundations of a new renaissance of worldwide research activities around the world.

Based on these initiatives and on the accomplishments demonstrated by the distributed research model in the past decades, we should remain confident that all this coordinated research will lead to the emergence of new devices and new architectures that will enable innovative products and will continue to drive the growth of the electronics industry in the decades to come!

Hopefully, by the middle of the next decade we will be able to state once again:

> "No exponential is forever: But "Forever" Can Be Delayed!

Acknowledgments

The author wishes to thank Dr. Gilbert Declerck for reviewing and correcting this chapter. He is also indebted to Dr. Roger De Keersmaecker for the many insightful suggestions, corrections, and discussions that have definitely made a profound and positive impact on this chapter.