2

The Economic Implications of Moore's Law

G.D. Hutcheson

2.1 Introduction

One hundred nanometers is a fundamental technology landmark. It is the demarcation point between microtechnology and nanotechnology. The semiconductor industry crossed it just after the second millennium had finished. In less than 50 years, it had come from transistors made in mils (one-thousandth of an inch or 25.4 microns); to integrated circuits which were popularized as microchips; and then as the third millennium dawned, nanochips. At this writing, nanochips are the largest single sector of nanotechnology. This, in spite of many a nanotechnology expert's prediction that semiconductors would be dispatched to the dustbin of science – where tubes and core memory lie long dead. Classical nanotechnologists should not feel any disgrace, as pundits making bad predictions about the end of technology progression go back to the 1960s. Indeed, even Gordon Moore wondered as he wrote his classic paper in 1965 if his observation would hold into the 1970s. Semiconductors owe their amazing resilience to Moore's law. To truly understand their greater impact, one must understand Moore's law.

Moore's law is predicated on shrinking the critical features of the planar process: The smaller these features, the more bits that can be packed into a given area. The most critical feature size is the physical gate length; as shrinking it not only makes the transistor smaller, it makes it faster. But we are fast approaching the limits of what can be done by scaling. What changes are needed to keep the silicon miracle going, especially as we approach the nano era? This book examines these changes from a technical standpoint because barriers to Moore's law have always been solved with new technology. However, these barriers are ultimately expressed economically and have important ramifications far beyond the industry itself. Moore's law is not only an expression of a powerful engine for economic growth in the industry, but also for the economy as a whole. This chapter reviews Moore's law and the economic implications that it poses. It shows how the continuation of Moore's law provides

a foundation for future economic growth and as such, sets the stage for a technical treatise on the nano era.

2.2 Moore's Law: A Description

Looking back thirty years after Gordon E. Moore first published his observations which would become known as Moore's law, he mused "The definition of 'Moore's Law' has come to refer to almost anything related to the semiconductor industry that when plotted on semi-log paper approximates a straight line" [1]. Indeed, this abuse of the meaning of Moore's law has led to a great deal of confusion about what it exactly is.

Simply put, Moore's law [2] postulates that the level of chip complexity that can be manufactured for minimal cost is an exponential function that doubles in a period of time. So for any given period, the optimal component density would be:

$$C_t = 2 \cdot C_{t-1}, \tag{2.1}$$

where C_t = Component count in period t, C_{t-1} = Component count in the prior period.

This first part would have been of little economic import had Moore not also observed that the minimal cost of manufacturing a chip was decreasing at a rate that was nearly inversely proportional to the increase in the number of components. Thus, the other critical part of Moore's law is that the cost of making any given integrated circuit at optimal transistor density levels is essentially constant in time. So the cost per component, or transistor, is cut roughly in half for each tick of Moore's clock:

$$M_t = \frac{M_{t-1}}{2}, \tag{2.2}$$

where M_t = Manufacturing cost per component in period t, M_{t-1} = Manufacturing cost component in the prior period.

These two functions have proven remarkably resilient over the years as can be seen in Fig. 2.1.[1] The periodicity, or Moore's clock cycle, was originally set forth as a doubling every year. In 1975, Moore gave a second paper on the subject. While the plot of data showed the doubling each year had been met, the integration growth for MOS logic was slowing to a doubling every year-and-a-half [3]. So in this paper he predicted that the rate of doubling would further slow to once every two years. He never updated this latter prediction. Between 1975 and 2006, the average rate between MPUs and DRAMs ran right at a doubling every two years.

2.3 The History of Moore's Law

Moore's law is indelibly linked to the history of our industry and the economic benefits that it has provided over the years. Gordon Moore has tried repeatedly to dismiss

[1] The forces behind the law were still strongly in effect when Gordon Moore retired in 2001, leading him to quip to the author that "Moore's law had outlived Moore's career."

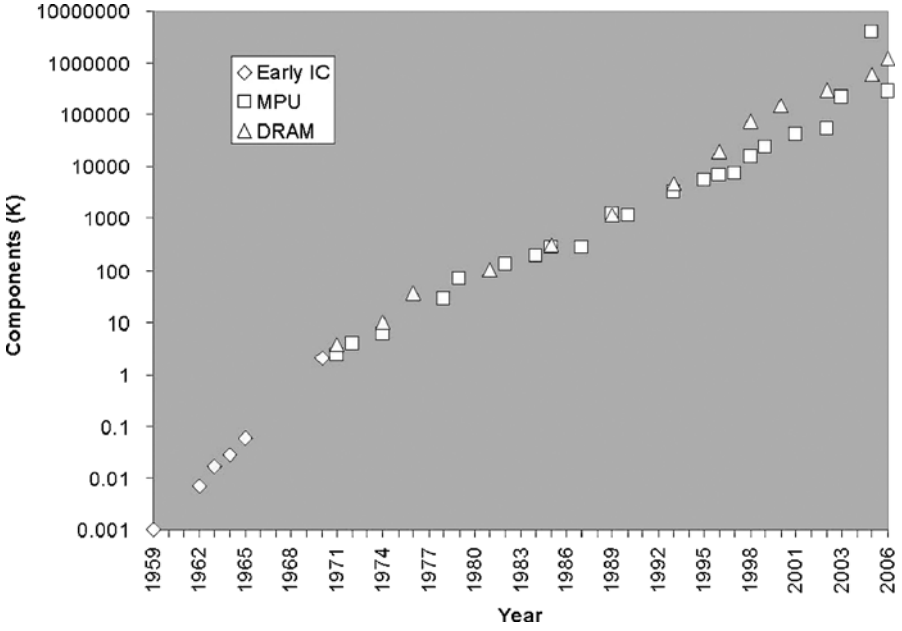

Fig. 2.1. Five decades of Moore's law

the notion that it was law, but instead just an observation. It was actually Carver Mead who first called the relationship "Moore's law." Either way, the term became famous because Moore had proved amazingly perceptive about how technology would drive the industry and indeed the world. Moore's observations about semiconductor technology are not without precedent. As early as 1887, Karl Marx, in predicting the coming importance of science and technology in the twentieth century, noted that for every question science answered, it created two new ones; and that the answers were generated at minimal cost in proportion to the productivity gains made [4]. His observation was one of the times, referring to mechanics for which the importance of the industrial age's development had been largely questioned by economists up to that point [5] (much like the productivity gains of computers in the latter twentieth century are still debated today) [6]. More important was Marx's observation that science and engineering had proved to be a reasonably predictable way of advancing productivity. Moreover, investments in science and engineering led to technology, which paid off in a way that grew economies, not just military might. Today, no one questions that science was at the heart of the industrial age, as it led to important inventions like the cotton gin, the steam engine, the internal combustion engine, and the fractional horsepower electric motor, to name a few. Nevertheless, it is the exponential growth of scientific "answers" that led to these, as well as to the invention of the transistor in 1947, and ultimately the integrated circuit in 1958, which led to

Moore's observation that became known as a law, and in turn launched the information revolution.[2]

The progress of science into the twentieth century would ultimately lead to the invention of the transistor, which is really where the story of the information revolution from the semiconductor perspective starts. Like all great inventions, it relied on the prior work of others. The solid-state amplifier was conceptualized by Julius Edgar Lilienfeld. He filed for a patent on October 8, 1926 and it was granted on January 28, 1930 (U.S. patent No. 1745175). While Lilienfeld didn't use the term Field Effect Transistor (FET), Oskar Heil did in British Patent No. 439457, dated March 2, 1934. Heil became the first person to use the term semiconductor. While neither ever gained from these patents, these works established the basis of all modern day MOS technology, even though neither author was cognizant of the concept of an inversion layer [7]. That is, the technology was not there to build it. Soon after World War II, figuring out how to make a solid state switch would become the Holy Grail of research as vacuum tube and electromechanical relay-based switching networks and computers were already proving too unreliable.

John Bardeen, Walter Brattain and William Shockley were in pursuit of trying to make workable solid state devices at Bell Labs in the late 1940s when Bardeen and Brattain invented the point-contact semiconductor amplifier (i.e., the point-contact transistor) on December 16, 1947 [7]. It was Brattain and Bardeen who discovered transistor action, not Shockley. Shockley's contribution was to invent injection and the p–n junction transistor. Bardeen, Brattain and Shockley, nevertheless, all properly shared the 1956 Nobel Prize in physics.

It was these efforts that would set the stage for the invention of the integrated circuit in 1958 and Moore's observation seven years later. The story of the integrated circuit centers on the paths of two independent groups, one at Fairchild and the other at Texas Instruments (TI), who in their collision created the chain reaction that created the modern semiconductor industry. It is more than a story of technology. It is a story about the triumph of human endeavor and the victory of good over bad management. It begins with the "Traitorous Eight" leaving Shockley Transistor in 1957 to start Fairchild Semiconductor (the eight were Julius Blank, Victor Grinich,

[2] These observations are often imitated as well. For example, Monsanto's 1997 annual report proclaimed Monsanto's Law, which is "the ability to identify and use genetic information is doubling every 12 to 24 months. This exponential growth in biological knowledge is transforming agriculture, nutrition, and health care in the emerging life sciences industry." Its measure is the number of registered genetic base pairs, which grew from nil to almost 1.2 billion between 1982 and 1997. Magnetic memory has seen a similar parallel to Moore's law as it shrinks the size of a magnetic pixel. Life sciences gains are a direct result of increased modeling capability of ever more powerful computers. Magnetic memory's gains are a direct result of chip manufacturing methodologies being applied to this field. Both are a direct result of the benefits gained from Moore's law. Indeed, Paul Allen of Microsoft fame has credited his observation that there would be a need for more increasingly powerful software as a direct result of learning about Moore's law. He reasoned that this would be the outcome of ever more powerful chips and computers and then convinced Bill Gates there was a viable future in software – something no major systems maker ever believed until it was too late.

Jean Hoerni, Eugene Kliener, Jay Last, Gordon Moore, Robert Noyce and Sheldon Roberts). They had been frustrated at Shockley because they wanted to move away from the four-layer device (thryristor) that had been developed at Bell Labs, and use lithography and diffusion to build silicon transistors with what would be called the mesa process. Fairchild was the first company to specialize exclusively in making its transistors out of silicon. Their expertise for pulling this off was a rare balance: Bob Noyce and Jay Last on litho and etch, Gordon Moore and Jean Hoerni on diffusion, Sheldon Roberts on silicon crystal growing and Gene Kliener on the financial side. The mesa process was named because cross-sectional views of the device revealed the steep sides and flat top of a mesa (it mounted the base on top of the collector). Debuted in 1958, it was the immediate rage throughout the industry, because transistors could be uniformly mass-produced for the first time.

But the mesa process would not survive because its transistors were not reliable due to contamination problems. They were also costly due to their labor intensity, as the contacts were hand-painted. It was Jean Hoerni who – in seeking a solution to these problems – came up with the planar process, which diffused the base down into the collector. It was flat and it included an oxide passivation layer. So the interconnect between the base, emitter and collector could be made by evaporating aluminum (PVD) on oxide and etching it. This was a revolutionary step that, with the exception of the damascene process, is the basis for almost all semiconductor manufacturing today. It is so important that many consider Jean Hoerni the unknown soldier whose contributions were the real seed for the integrated circuit (IC). This is because the aluminum layer would make it easy to interconnect multiple transistors. The planar process was the basis for Fairchild's early success and was considered so important that it was kept secret until 1960. Its process methodology was not revealed until after the IC had been invented. At the time, however, it was only viewed as an important improvement in manufacturing. The first work to actually interconnect transistors to build an IC was actually occurring halfway across the United States.

Jack Kilby joined TI in May of 1958 and had to work through its mass vacation in July. A new employee, with no vacation time built-up, he was left alone to ruminate over his charge of working on microminiaturization. It was then that he came up with the idea of integrating transistors, capacitors and resistors onto a single substrate. It could have been a repeat of what happened at Shockley Semiconductor. Kilby's bosses were skeptical. But instead of chasing him off, they encouraged him to prove his ideas. TI already had a mesa transistor on the market, which was made on germanium slices (TI used to call "die and wafers" "bar and slices"). Jack took one of these slices and cut it up into narrow bars (which may be why TI called chips "bars" versus the more commonly used word "die"). He then built an integrated phase-shift oscillator from one bar with a germanium mesa transistor on it and another with a distributed RC network. Both were bonded to a glass substrate and connected with a gold wire. He then built a flip-flop with multiple mesa transistors wire-bonded together, proving the methodology was universal in October of 1958. This was the first integrated circuit ever made. It was unveiled in March 1959 at the Institute of Radio Engineers show.

Back at Fairchild, during January of 1959, Bob Noyce entered in his notebook a series of innocuous ideas about the possibility of integrating circuits using Hoerni's planar process, by isolating the transistors in silicon with reversed biased p–n junctions, and wiring them together with the PVD-Litho-Etch process using an adherent layer of Al on the SiO_2. This was then put away until word of Jack Kilby's invention rocked the world later in March 1959. While many derided Kilby's work as a technology that would never yield, with designs that were fixed and difficult to change, Noyce sat up and took notice. Putting it all together, Noyce and his team at Fairchild would architect what would become the mainstream manufacturing process for fabricating integrated circuits on silicon wafers.[3] Transistors, capacitors, and resistors could now be integrated onto a single substrate. The reasons why this method was so important would be codified in Moore's 1965 paper.

In 1964, *Electronics Magazine* asked Moore, then at Fairchild Semiconductor, to write about what trends he thought would important in the semiconductor industry over the next ten years for its 35th anniversary issue. He and Fairchild were at the forefront of what would be a revolution with silicon. However, when Moore sat down to write the paper that would become so famous for its law, ICs were relatively new – having been commercialized only a year or two earlier. Many designers didn't see a use for them and worse, some still argued over whether transistors would replace tubes. A few even saw ICs as a threat: If the system could be integrated into an IC, who would need system designers? Indeed even Moore may have been skeptical early on. Robert Graham recalled that in 1960, when he was a senior Fairchild sales and marketing executive, Moore had told him, "Bob, do not oversell the future of integrated circuits. ICs will never be as cheap as the same function implemented using discrete components" [8]. In fact, Moore actually didn't notice the trend until he was writing the paper [9]. Nevertheless, by 1964 Moore saw the growing complexity and lowered cost and this understanding of the process convinced him that ICs would come to dominate. Fairchild was trying to move the market from transistors to ICs. Moore was also convinced that ICs would play an important role and he was writing the paper that would convince the world.

[3] Ironically, Kilby's method for integrating circuits gets little credit for being what is now widely viewed as one of the most important paths into the future. In practice, his invention was what would be renamed as hybrid circuits, which would then be renamed Multi-Chip Modules (MCM), then Multi-Chip Packages (MCP), and now System In a Package (SIP). It is clear today that System-On-a-Chip (SOC) is limited by the constraints of process complexity and cost; and so Jack Kilby's original integrated circuit is finally becoming a critical mainstream technology. Unfortunately, while he gets credit for the invention of the IC, few give him credit for inventing a method that had to wait 40 years to become critical in a new millennium.

Most give both Jack Kilby and Bob Noyce credit as co-inventors of the IC because of these differences; they both came up with similar ideas independently and it was Jack Kilby that prodded Bob Noyce into action. TI would go on to become an industry icon. Fairchild's early successes would turn to failure under bad management and suffer a palace revolt, similar to the one at Shockley, in 1967. It was called the Fairchild brain drain and resulted in the founding of 13 start-ups within a year. Noyce and Moore would leave to start-up Intel in 1968. But that's another story.

Titled "Cramming More Components into Integrated Circuits," Moore's paper was published by *Electronics Magazine* in its April 19, 1965 issue on page 114. Its subtitle was "With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single chip of silicon." This issue's contents exemplify how so few really understood the importance of the IC. Ahead of it was the cover article by RCA's legendary David Sarnoff who, facing retirement, reminisced about "Electronics' first 35 years" with a look ahead. Behind this were articles titled "The Future of Electronics in 1930 – Predictions from Our First Issue" and "A Forward Look at Electronics – Looking Farther into the Future" (both written by the magazine's authors). Then there appeared an article from Motorola, by Dan Noble titled "Wonderland for Consumers – New Electronic Products Will Make Life Easier." All these papers were before Moore's paper. Indeed, Moore's paper would have been at the back of the magazine had it not been for what would prove to be mundane papers titled "Changing the Nature of Research for Space," "Light on the Future of the Laser," "More and Faster Communications" and "Computers to Run Electric Utility Plants." At the time *Electronics Magazine* was the most respected publication covering the industry and it had assembled the best visionaries possible. Yet, with the exception of Moore's paper, it was mostly classic forecasting "through the rear-view mirror." His paper would be the only thing remembered from this issue. In fact, those who entered the industry in the 1990s wouldn't even recognize the magazine as it is now defunct, not surviving the Moore's law article it contained.

Moore's law paper proved so long-lasting because it was more than just a prediction. The paper provided the basis for understanding how and why ICs would transform the industry. Moore considered user benefits, technology trends, and the economics of manufacturing in his assessment. Thus he had described the basic business model for the semiconductor industry – a business model that lasted through the end of the millennium.

From a user perspective, Moore's major points in favor of ICs were that they had proven to be reliable, they lowered system costs and they often improved performance. He concluded, "Thus a foundation has been constructed for integrated electronics to pervade all of electronics." This was one of the first times the word "pervade" was ever published with respect to semiconductors. During this time frame the word "pervade" was first used by both Moore and Patrick Haggerty of TI. Since then, the theme of increasing pervasiveness has been a feature of almost all semiconductor forecasts.[4]

From a manufacturing perspective, Moore's major points in favor of ICs were that integration levels could be systematically increased based on continuous improvements in largely existing manufacturing technology. The number of circuits that could be integrated at the same yield had already been systematically increas-

[4] Pervasiveness is another misused word. Many have used it during boom times to argue that the semiconductor industry would no longer be cyclical and thus, not have a downturn. While semiconductors have been increasingly pervasive since the dawn of the industry, this fact has not been a factor in alleviating the industry's inherent cyclicality.

ing for these reasons. He saw no reason to believe that integration levels of 65,000 components would not be achieved by 1975 and that the pace of a doubling each year would remain constant. He pointed to multilayer metalization and optical lithography as key to achieving these goals. Multilayer metalization meant that single transistors could be wired together to form ICs. But of far greater import was the mention of optical lithography.

Prior to the invention of the planar process, the dominant device was known as a mesa transistor. It was made by hand painting black wax over the areas to be protected from etchant. While the tool was incredibly inexpensive (a 10-cent camel's hair brush), the process was incredibly labor intensive [10].[5] The introduction of optical lithography meant transistors could be made simultaneously by the thousands on a wafer. This dramatically lowered the cost of making transistors. This was done to the degree that, by the mid-1960s, packaging costs swamped the cost of making the transistor itself.

From an economics perspective Moore recognized the business import of these manufacturing trends and wrote, "Reduced cost is one of the big attractions of integrated electronics, and the cost advantage continues to increase as the technology evolves toward the production of larger and larger circuit functions on a single semiconductor substrate. For simple circuits, the cost per component is nearly inversely proportional to the number of components, the result of the equivalent package containing more components."

As important as these concepts proved to be, it was still not clear that the arguments would stand the test of time. Package costs now overwhelmed silicon processing costs. These costs were not subject to Moore's law and technical efforts were shifting to lowering them. Fairchild was reducing packaging costs, which were still highly labor intensive, by moving its assembly lines offshore. Texas Instruments and Motorola among others were developing and building automatic assembly equipment. Many, even those at Fairchild, were still struggling with how to make a profitable business out of ICs. While transistors could be integrated, designing and marketing circuits that customers could easily use proved more complicated. The industry had no design standards. Fairchild had developed circuits with Resistor–Transistor Logic (RTL), but customers were using Diode–Transistor Logic (DTL). Plus, Fairchild was undergoing major internal turmoil as political battles raged throughout the company. Many start-up companies were spinning out of it as senior executives left.

The most famous of these spin-offs was Intel, for which its lead founders included no lesser than Robert Noyce and Gordon Moore. Lacking the packaging infrastructure of Fairchild and having the cream of its research capability, Intel's strength was in its founders' ability to build complex circuits and their deep understanding of Moore's law. They leveraged this by focusing on memories, which Moore believed had huge market potential and could be more easily integrated – both in terms of putting large numbers of transistors on silicon and in integrating

[5] See also [11, Chap. 3, pp. 53–56].

them into customers' designs.[6] He was also convinced that the only way to compete effectively was by making the silicon more valuable than the package, offsetting the natural advantage in packaging that Fairchild, Motorola and TI had.

The strategy worked. Intel became the memory IC market leader in the early 1970s. They had started with the SRAM (Static Random Access Memory) and soon invented the DRAM (Dynamic Random Access Memory), which proved to be very integrable and was much more cost effective than the ferrite core memories used by computer makers at the time. It was at this point that Moore's law began to morph into the idea that the bits were doubling every year or two. Transistors were now old fashioned. Few argued the practicality of ICs. It was important to move on and use Moore's law as a way to demonstrate the viability of the emerging memory market. At the time, there was more to the marketing of Moore's law than most ever understood. The strategies taken would become a critical competitive advantage for Intel – enough of an advantage to keep it ahead of TI, who also focused on memories and had much more manufacturing infrastructure.

Bob Graham, another Intel founder, recalled[7] that there was a problem with Moore's law: it was too fast. Its cycle called for a doubling every year, but systems designers needed more than a doubling to justify a new design. They typically fielded a new design every three-to-four years. Graham's strategy to harness the power of Moore's law was to match the chip design cycle to the system designers'. The difference between the nodes of Moore's clock cycles and these memory nodes would lead to much confusion. But according to Graham, Intel used this confusion to keep competitors at bay when Intel's early memory strategies were plotted. It was considered highly confidential and a big secret that the real generational nodes were based on a quadrupling, not a doubling. Moore's law in his view was also a marketing head fake.

Intel introduced each new generation of memories with a $4\times$ increase in bits about every three years. Each of its generations was closely matched to customers' design cycles. Other competitors fell behind because they followed the letter of Moore's law. They tried to get ahead by introducing new chips with every $2\times$ increase. But interim generations failed. They failed from the first 64-bit memory introduced by Intel to the 64 M-bit memory. This cycle, with every $4\times$ increase in bits, was not broken until the 128 M-bit DRAM came to market three decades later in the early 2000s.

Tax law and capital depreciation also played a critical role in determining the pacing of nodes. The United States' MACRS (Modified Accelerated Cost Recovery Systems) tax code for capital depreciation specified that computers be fully depreciated over a six-year length of time. If systems makers had designed new systems every year, there would have been six generations of computers per depreciation cycle – clearly too many for customers. Customers would have over-bought and had to write-off equipment that was not fully depreciated, the annual market size would have been split into a sixth of its potential, or some compromise between the two

[6] See also [11, pp. 181–185].

[7] Private conversations with the author.

Table 2.1. Average months to double device complexity

Year	Overall	MPU	DRAM
1959–1965	12		
1966–1975	17	33	17
1976–1985	22	22	25
1986–1995	32	38	25
1996–2005	27	26	22
1976–2005	24	24	24

would have happened. Early on, systems makers paced designs so that at least one half of potential customers would replace their systems every two-to-three years. It is likely that the competitive reasons accounted for the more rapid cycling of system designs in the early 1960s. The computer market was still highly competitive then. It consolidated during the latter 1960s and IBM emerged as the dominant supplier. There are no technical reasons to explain why the node pacing slowed. But from a market perspective, the pace should have slowed naturally as IBM sought to leverage its dominance to extend the life of designs, hence having greater amortization of these costs and enhancing profitability. IBM was sued for monopolist trade practices and one of the claims was that it intentionally slowed technology. Whatever the reason, node pacing slowed to a rate of one design node every three years. Moore's clock was roughly half that.

In 1975, Moore wrote an update to the 1965 paper and revised his predictions. While technically his prediction of 65,000 components had come true, it was based on a 16 K-bit CCD memory – a technology well out of the mainstream. The largest memory in general use at the time – the 16 K-bit DRAM, which contained less than half this amount – would not be in production until 1976. Between 1965 and 1975 the pace had actually slowed to a doubling every 17 months or roughly every year-and-a-half. Later, Moore's law was widely quoted by others as being every 18 months. But, despite being widely referenced as the source, this is a periodicity that Moore never gave. The 1975 paper actually predicted the periodicity would be a doubling every two years [3]. This would turn out to be extremely accurate, though seldom quoted with any accuracy. Contrary to what many have thought, the finer points of the accuracy of Moore's law never really mattered.

The real import of Moore's law was that it had proved a predictable business model. It gave confidence in the industry's future because it was predictable. One could plan for it and invest in it on the basis that the integration scale would always rise in a year or two, obsolescing the electronics that were out there and creating new demand because the unobtainable and confusing would become affordable and easy to use. This then fed back to reinforce it, as engineers planned for it and designed more feature-rich products or products that were easier to use. As Moore later put it, Moore's law "had become a self-fulfilling prophecy [9]". But as a result, heavy international competition and technical issues would loom in the future.

It was at about this time that Japan seized on Moore's law as a planning mechanism. Without it, the industry appeared to go in random directions. But Moore's law made it easy to plan and it had been clearly proven by 1975. The DRAM explosion was already in place at TI, AMD, IBM, Intel, and Mostek. Moore's law made it all understandable. Now believers in Moore's law, they saw that since memory demand and device type were very predictable it would play to their strengths. Moreover, cost of manufacturing was critical to success – one of their key strategic strengths. Moore's law was the basis of the arguments that prompted them to start their government-funded effort called the VLSI program in 1976, for which the goal was to build a working 64K-bit DRAM. They believed that if the VLSI program could do this, their semiconductor industry could lever off the results to build their own.

Japan was already successful in 4 K-bit DRAMs and their potential with 16 K-bit DRAMs looked promising. One of the keys to their success was that they implemented multiple development teams. Each team worked on the same generation, walking it from research, through development, and into manufacturing. In contrast, the west had highly stratified walls between these three areas. Research departments threw their results over the wall to development, and so forth into manufacturing. Often, manufacturing used few of these efforts because they were seldom compatible with manufacturing. Instead they built designs coming directly from marketing because they knew they would sell. Japan got the edge because they could get new designs to manufacturing faster and their cost of manufacturing was fundamentally lower. They had lower labor rates and their line workers typically stayed with a company for several times longer. This combined with the Japanese penchant for detail. TI had a facility in Japan and this facility consistently yielded higher than its American facilities for these reasons.[8] The Japanese also had newer equipment, having invested heavily in the late 1970s. Capital was tough to get in the late 1970s for American chipmakers. They had cut their investments to low levels, which would ultimately give Japan another source of yield advantage.

But the real shocker came in 1980, when Hewlett-Packard (HP) published an internal study comparing quality between American- and Japanese-made DRAMs. It showed Japan to have higher quality DRAMs. American chipmakers cried foul – that this was tested-in quality and that Japanese suppliers sent HP more thoroughly tested parts. Indeed, independent studies did later show that Japanese-made DRAM's obtained on the open market were of no higher quality than American ones. However, it was too late to change the perception that HP's announcement had created (a perception that remains to this day).

Whether or not the quality was tested-in, the one clear advantage the Japanese had was yield. It was typically 10–15% higher than equivalent American fabs and this gave the Japanese a fundamental cost advantage. When the downturn hit in 1981, these yield differences allowed Japanese companies to undercut American companies on 16-Kbit DRAMs. This, combined with the downturn, caused American producers to make further cuts in capital investment, and put them further behind. At

[8] Conversations with Howard Moss, a Board Member of Texas Instruments at the time, 1985.

the time, the Chairman of NEC noted that they had no fab older than five years. The average American fab was 8 years old. So by the early 1980s, Japan came to dominate 64-Kbit memories. By 1985, America's giants were bleeding heavily. Intel was among the worst when it came to manufacturing memories. It was forced out of memories.

The technical issues with the 64-Kbit DRAM proved to be enormous. The most commonly used lithography tool of the day, the projection aligner, would not make it to this generation because it lacked the overlay alignment capability. Something new had to be developed and the industry was not ready. The transition to stepping aligners proved much more traumatic than anyone expected. Steppers had the potential for very high yields. But unless the reticle was perfect and had no particles, yield would be zero because the stepper would repeat these defects. The result was a three-year hiatus in Moore's law. 64-Kbit DRAMs did not enter volume production until 1982 – a full three years after they should have arrived – taking a full six years to pass from the 16-Kbit node.

Another transition occurred in the early 1980s that favored Intel. NMOS began to run out of steam and couldn't scale well below one micron. Some even predicted that we had hit Moore's wall. But the industry escaped this by transitioning to CMOS. One of the benefits of CMOS was that performance also scaled easily with shrinks. An engineer in Intel's research organization observed this and recognized its importance to microprocessors. Moreover, having exited memories it was important that Intel not lose the brand value of Moore's law it had, having its discoverer as chairman of the company. So marketing morphed Moore's law a second time. It had started as the number of transistors doubling, then the number of bits, and now it was speed, or more comprehensively, performance. This new form would serve Intel and the industry well.

In the early 1990s, the pace of integration increased again. There were several factors driving this change. It could be argued that the manufacturing challenges of the early 1980s had been overcome. Yet there was significant fear that new hurdles looming in the near future would be insurmountable. America's semiconductor industry had just instituted the roadmap process for planning and coordinating semiconductor development. As it turned out, this focused pre-competitive research and development like it had never been before. Instead of hundreds of companies duplicating efforts, it allowed them to start from a common understanding. The result was an increased pace of technology development. At the same time, efforts to reinvigorate competitiveness led companies to adopt time-to-market measures of effectiveness. This naturally accelerated the pace of development. Also, the shift from mainframe and minicomputers to personal computers had dramatically altered the competitive landscape in the 1980s.

IBM had quickly come to dominate the PC market in the early 1980s. Unlike all earlier IBM computers, the PC had been designed with an open architecture. Their dominance might never have been challenged. However on August 2, 1985, the IBM senior executives who ran and had developed its PC business violated a major corporate rule and boarded Delta Airlines flight 191 to Dallas, Texas. The flight encountered wind-shear and crashed on landing. IBM's understanding of how

the market was evolving as well as its leadership capability in the still-emerging PC market perished. Unlike all earlier IBM computers, the PC had been designed with an open architecture, which meant it could be easily cloned. Outside of this small group, IBM really didn't understand how to respond to the hoards of clone makers who had entered the market. Applied to pricing and self-obsolescence, the clone hoard's slash-and-burn strategies made IBM's defenses about as useful as France's Maginot line in World War II. As they lost their leadership, the pace of technical change accelerated again to limits set primarily by technical developments.

At the 1995 Semiconductor Industry Association (SIA) forecast dinner, Gordon Moore gave a retrospective on 30 years of Moore's law. He claimed to be more surprised than anyone that the pace of integration had kept up for so long. He had given up on trying to predict its end, but commented that it was an exponential and all exponentials had to end. While it did not stop the standing ovation he received, he concluded that "This can't go on indefinitely – because by 2050… we're everything."

2.4 The Microeconomics of Moore's Law

The essential economic statement of Moore's law is that the evolution of technology brings more components and thus greater functionality for the same cost. This cost reduction is largely responsible for the exponential growth in transistor production over the years. Lower cost of production has led to an amazing ability to not only produce transistors on a massive scale, but to consume them as well. Each year's new production alone amounts to roughly 40% of the total transistors ever produced in every year before it. It has crossed 12 orders of magnitude since the industry's inception. By comparison, auto production did not cross a single order of magnitude over this period. The other amazing aspect of this is that anomalies such as the millennial boom have no effect on production. In any given year since the industry's birth, the production of transistors has averaged 41% of the cumulative total ever produced up until then.

So what makes Moore's law work? The law itself only describes two variables in the equation: transistor count and cost. Behind these are the fundamental technological underpinnings that drive these variables and make Moore's law work. There are three primary technical factors that make Moore's law possible: reductions in feature size, increased yield, and increased packing density. The first two are largely driven by improvements in manufacturing and the latter largely by improvements in design methodology.

Design methodology changes have been significant over the years. They have come as both continuous and step function improvements. The earliest step function improvements were the reduction in transistor counts to store memories. The industry started with 6-transistor memories. In the late 1960s, designers quickly figured how to reduce this to four, then two, and, finally, the 1-transistor/1-capacitor DRAM cell, developed by R.H. Dennard [12]. While this did not affect Moore's law as measured in transistors, it did when measured in bits, because a 4-Kbit memory (with a 6-T

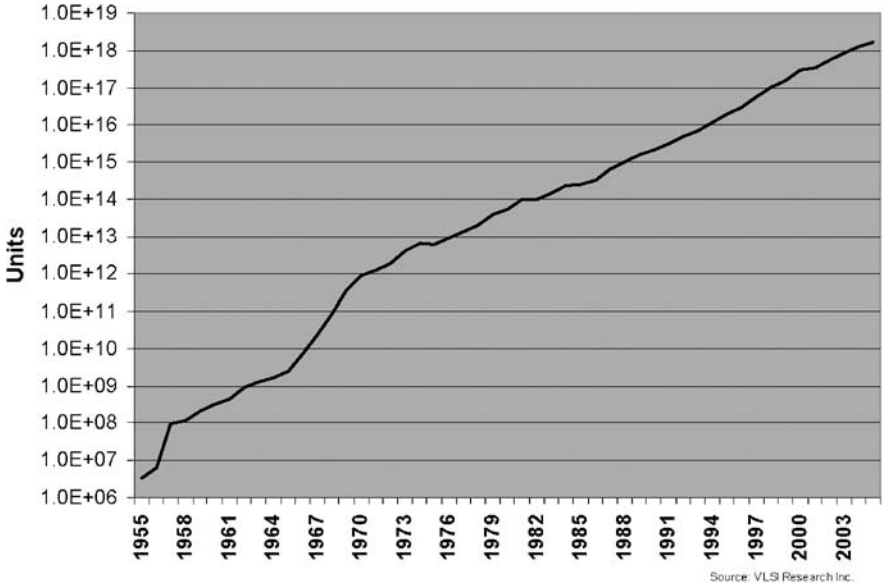

Fig. 2.2. Worldwide transistor production (all, including foundry, merchant, and captive producers)

cell) needed 24-K transistors and could now be made with only 4-K transistors with a 1 T/ 1 capacitor cell. This was an enormous advance. Cost-per-bit plummeted and it further added to the mythical proportions of Moore's law, as customers saw little real difference between transistors and bits. What they were most interested in was reductions in cost-per-function and designers had delivered. There were less well-known additions as well. The development of Computer-Aided Design (CAD) in the early 1980s was a significant turning point. Now with Electronic Design Aids (EDA), CAD's first contribution was to prevent the ending of Moore's law. As the industry progressed from MSI to LSI levels of integration, the number of transistors to be wired together was becoming too large for humans to handle. Laying out the circuit diagram and cutting the rubylith[9] for wiring 10,000 transistors (with three connections each) together, with 3-connections each, by hand had to have been a daunting task. The 64-Kbit DRAM loomed large in this picture as the decade turned. With just over 100,000 K transistors, it would be the first commercial VLSI grade

[9] In those days, masks were made by drawing the circuit out on a large piece of paper that sometimes covered the floor of a large room. Then a laminated piece of Mylar called rubylith was used to make the initial mask pattern. Rubylith was made of two Mylar films, one clear and one red. A razor-edged knife was used to cut strips of the red Mylar away. Initially this was done by hand. One had to be careful not to cut the clear Mylar underneath so the red Mylar could be pulled away, leaving the final mask pattern. This pattern was then reduced to circuit dimensions to make a mask master.

Table 2.2. Integration scale measures used in the 1970s

SSI	Small Scale Integration	<100 Transistors
MSI	Medium Scale Integration	101–1,000 Transistors
LSI	Large Scale Integration	1,001–10,000 Transistors
VLSI	Very Large Scale Integration	>100,000 Transistors

chip produced in volume – and it was a point of hot competition. So being able to automate this process would be a great leap forward.

The next step for design was to improve the layout capability of these tools. This improved the packing density. Early layout tools were fast, but humans could lay out a circuit manually in 20–30% of the area. Today, no one would manually lay out an entire circuit with millions of transistors. Even today, EDA tools do not offer the most efficient packing density. Designers who want the smallest die will "handcraft" portions of a circuit. This is still commonly done when a market is large and the die-size reduction can justify the cost of handcrafting.

Performance improvements are another way that design has directly affected Moore's law. It is particularly important to the third version of Moore's law, which measures the gain in circuit performance over time. Scaling theory states that transistor switching speed increases at a rate that is inversely proportional to the reduction in physical gate length. However, a designer can improve on this by using the switching of transistors more efficiently. These transistors switch with the clock of the circuit. Early processor architecture required several clock cycles per instruction. So a 1-GHz clock might only perform at a rate of 300 Millions of Instructions Per Second (MIPS). Using techniques like parallel processing, pipelining, scalar processing, and fractional clocks, designers have systematically improved this so that three-to-five instructions per clock cycle can be achieved. Thus, a processor with a 1-GHz clock can exhibit run rates of 3,000-to-5,000 MIPS. Considering 1-MIP was considered state-of-the-art for a circa-1980 mainframe processor, subsequent architectural gains have been quite significant.

Design tools are having further impacts today; one is their ability to improve testing. Without these tools test costs would explode or worse, the circuits would be untestable, making further gains in integration scale pointless. Another impact is the ability to automatically lay out the patterns needed to make reticles with optical proximity correction and phase-shifting. This substantially reduces feature sizes. But, it is important to realize that none of these gains would have been possible without ever more powerful cost-effective computers. All of these benefits were made possible by Moore's law. Hence, instead of running down, Moore's law is a self-fulfilling prophecy that runs up. Indeed, many of the manufacturing improvements since the 1980s have come only because Moore's law had made computing power so cheap that it could be distributed throughout the factory and in the tools, be used to design the tools, and even perform the engineering and economic analysis to make more efficient decisions.

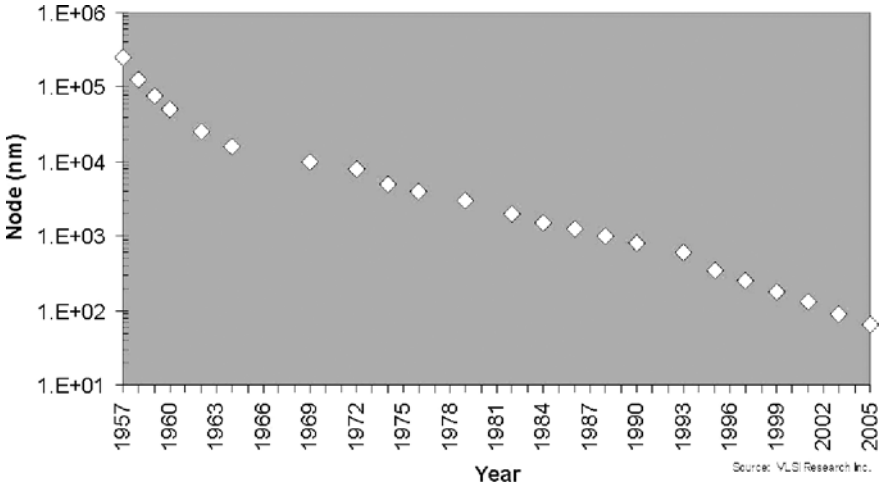

Fig. 2.3. Five decades of critical dimension shrinks (in nanometers)

Reductions in feature sizes have made the largest contributions by far, accounting for roughly half of the gains since 1976. Feature sizes are reduced by improvements in lithography methods. These enable smaller critical dimensions (CDs, which are also known as Minimum Feature Sizes or MFSs) to be manufactured. If the dimensions can be made smaller, then transistors can be made smaller and hence more can be packed into a given area. This is so important that Moore's first paper relied entirely on it to explain the process.

Improvements in lithography have been the most significant factor responsible for these gains. These gains have come from new exposure tools; resist processing tools and materials; and etch tools. The greatest change in etch tools was the transition from wet to dry etching. In etching, most of the older technology is still used today. Wet chemistries used for both etching and cleaning are the most prominent of these. Improvements in resist processing tools and materials have generally been incremental. Resist processing tools have remained largely unchanged from a physical perspective since they became automated. The changes have mostly been in incremental details changed to improve uniformity and thickness control. Resist chemistries have changed dramatically, but these changes are easy to overlook. Moreover etch and resist areas have relatively small effects on cost. Exposure tools have gone through multiple generations that followed the CD reductions. At the same time they have been the most costly tools and so generally garner the most attention when it comes to Moore's law. Moreover, without improvements in the exposure tool, improvements elsewhere would not have been needed.

Exposure tools were not always the most costly tools in the factory. The camel's hair brush, first used in 1957 to paint on hot wax for the mesa transistors, cost little more than 10 cents. But since that time prices have escalated rapidly, increasing roughly an order of magnitude every decade-and-a-half. By 1974, Perkin–Elmer's newly introduced projection aligner cost well over $100,000. In 1990, a state-of-the-

Table 2.3. Evolution of lithography technology used to manufacture semiconductors

Year first used in manufacturing	CD (microns)	Lithography technology	Etch
1957	254.000	Camel's hair brush, hand painting	Wet etching
1958	127.000	Silk screen printer	
1959	76.200	Contact printer W/emulsion plates	
1964	16.000	Contact printer W/chrome plates	
1972	8.000	Proximity aligner	
1974	5.000	Projection aligner	Barrel plasma
1982	2.000	g-line (436 nm) stepper	Planar plasma
1984	1.500		Reactive ion etching
1988	1.000		High density plasma
1990	0.800	i-line (365 nm) stepper	
1997	0.250	248 nm scanner	
2003	0.100	193 nm scanner	

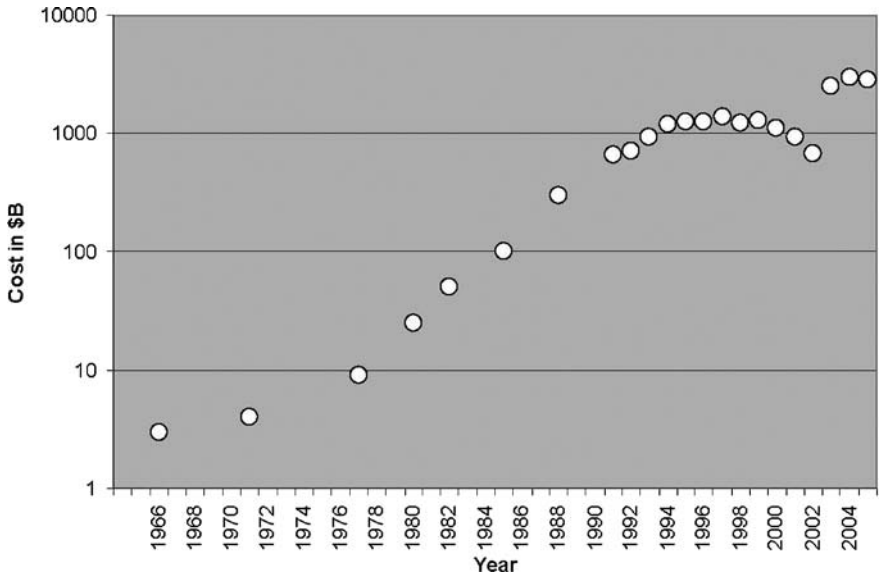

Fig. 2.4. History of average cost to build & equip a wafer fab

art i-line stepping aligner cost just over $1 million. When this was first written in 2002, 193-nm ArF excimer laser scanning aligners were about to enter manufacturing and they cost a shocking $10 million each. As of this writing in late 2006, they cost upwards of $50 million.

Over the decades, these cost increases have been consistently pointed to as a threat to the continuance of Moore's law. Yet the industry has never hesitated to

adopt these new technologies. It is a testimony to the power of this law that these costs have been absorbed with little effect.

Lithography tools have become more productive to offset these increases. Moreover, they are only part of the rising cost picture. The increase in the cost of semiconductor factories had been a recurring theme over the years. In fact it was first noted in 1987 that there was a link between Moore's law and wafer fab costs [13]. Between 1977 and 1987, wafer fab costs had increased at a rate of 1.7× for every doubling of transistors.

In the 1980s, the cost of factories was offset primarily by yield increases. So rising tool costs were offset by lower die costs. However, this relationship stalled in the 1990s, when the rise in tool prices began to be offset by increases in productivity. So as effective throughputs rose, the unit volumes of tools in a fab needed to produce the same number of wafers declined. This did change somewhat with the introduction of 300 mm wafers. However, the doubling in fab costs, shown above starting in 2004, is due to the fact that the typical output of a fab in wafers doubled. When normalized for output, fab costs have been constant since the 1990s and really have ceased to be an issue.

In any case, Moore's law governs the real limit to how fast costs can grow. According to the original paper given in 1965, the minimal cost of manufacturing a chip should decrease at a rate nearly inversely proportional to the increase in the number of components. So the cost per component, or transistor, should be cut roughly in half for each tick of Moore's clock (see (2.1) and (2.2)). However, since this paper was first given, it has generally been believed that industry growth will not be affected if the cost per function drops by at least 30% for every doubling of transistors. This 30% drop would allow the manufacturing cost per unit area of silicon to rise by 40% per node of Moore's law (or by twice the cost-per-function reduction ratio requirement) (Appendix A). This includes everything from the fab cost to materials and labor. However it does not take yield or wafer size into account.

Thus if cost per function needs to drop by 30% with each node, wafer costs can also theoretically increase by 40%, assuming no yield gains (see Appendix A for proof). Yield is a function of die size and so is directly dependent on component counts and CD reductions. There are many equations for calculating yield, the most basic of which is the classic Poisson probability exponential function:

$$Y = \exp -(ADN),$$

where A = die area, D = defect density per mask layer, N = number of masks.

Note that this equation also accounts for increased process complexity as component counts rise. It would seem that this effect would be the most significant cost-reducing factor. In the early days of the industry, it was. In the 1970s, yield typically started at 15% when a device was introduced and peaked at 40% as the device matured. Gains of two-to-three times were typical over the life of a chip and gains of four-to-five times were not uncommon for devices that had long lives. Improvement in manufacturing methods increased these gains dramatically during the 1980s and 1990s. This was primarily due to better equipment and cleanroom technology. For example, the switch to VLSI equipment technology such as steppers and plasma

etchers caused initial yields for the 64-Kbit DRAM to come in at 40% in 1982. It matured at 75% three years later. Today, devices typically enter production at 80%; rise to 90% within six months; and can achieve close to 100% at maturity. But at best the gain is only a quarter of initial yields.

Wafer size has been another cost-reducing factor used over the yields. Larger wafers have typically cost only 30% more to process and yet have had an area increase of 50-to-80%. The move to 300 mm wafers from 200 mm will yield an area increase of 125%! Larger wafers have always brought a yield bonus because of their larger "sweet spot" – yielding a relatively larger number of good chips in the inner regions of the wafer – and the fact that they require a new generation of higher-performing equipment. Like the sweet spot of a tennis racket, wafers tend to have the lowest defect density at their centers and highest at their edges where they are handled most. In addition, process chamber uniformity tends to suffer the most at the edges of the wafer. There are also gains in manufacturing efficiency that occur over time. The result is a continued decrease in manufacturing costs per die.

However, the continuation of Moore's law via reduction in CDs, increased yields, larger wafer sizes, and manufacturing improvements has taken its toll in other areas. Costs have risen significantly over time as seen in the rise of wafer fab costs. Moreover, the CD reductions have caused a need for increasing levels of technical sophistication and the resultant costs. For example, the camel's hair brush used for lithography in the 1950s cost only 10 cents; early contact aligners cost $3,000–5,000 in the 1960 and $10,000 by the 1970s; a projection aligner in the late 1970s cost $250,000; and the first steppers cost $500,000 in the early 1980s. By 2000, a 193-nm (using an ArF excimer laser) stepper cost $10 million and the latest tools can cost upward of $50 million. That is an increase of more than eight orders of magnitude over five decades.

Moreover, the cost increases are prevalent throughout the fab. Increased speeds have forced a transition from aluminum to copper wiring. Also, silicon-dioxide insulation no longer works well when millions of transistors are switching at 2 GHz, necessitating a switch to interlevel dielectrics with lower permittivity. At the gate level, silicon-dioxide will no longer be useful as a gate dielectric. Scaling has meant that fewer than ten atomic thicknesses will be used and it will not be long before they fail to work well. The solution is to replace them with high-k dielectrics so that physical thicknesses can be increased, even as the electrical thickness decreases. These new materials are also causing costs to escalate. An evaporator, which could be bought for a few thousand dollars in the early 1970s, now costs $4–5 million. Even diffusion furnaces cost $1 million per tube. As costs have risen, so has risk. There has been a tendency to over-spec requirements to ensure a wide safety margin. This has added to cost escalation.

At some point the effect of these technologies translating into high costs will cause Moore's law to cease. As Gordon Moore has put it, "I've learned to live with the term. But it's really not a law; it's a prediction. No exponential runs forever. The key has always been our ability to shrink dimensions and we will soon reach atomic dimensions, which are an absolute limit." But the question is not if,

it's when will Moore's wall appear? "Who knows? I used to argue that we would never get the gate oxide thickness below 1000 angstroms and then later 100. Now we're below 10 and we've demonstrated 30-nm gate lengths. We can build them in the 1000's. But the real difficulty will be in figuring out how to uniformly build tens of millions of these transistors and wire them together in one chip" [14]. In fact, we routinely build them today in cutting-edge manufacturing. In the end, it is more likely that economic barriers will present themselves before technical roadblocks stop progress [15].

2.5 The Macroeconomics of Moore's Law

Moore's law was more than a forecast of an industry's ability to improve, it was a statement of the ability for semiconductor technology to contribute to economic growth and even the improvement of mankind in general. It has a far richer history than the development of semiconductors, which to some extent explains why Moore's law was so readily accepted. This history also explains why there has been an insatiable demand for more powerful computers no matter what people have thought to the contrary.

The quest to store, retrieve, and process information is one task that makes humans different from other animals. The matriarch in a herd of elephants may be somewhat similar to the person in early tribes who memorized historical events by song. But no known animal uses tools to store, retrieve, and process information. Moreover the social and technological progress of the human race can be directly traced to this attribute. More recent writers have pointed to this as a significant driving force in the emergence of western Europe as the dominant global force in the last millennium [16].

Man's earliest attempts to store, retrieve, and process information date back to prehistoric times when humans first carved images in stone walls. Then in ancient times, Sumerian clay tokens developed as a way to track purchases and assets. By 3000 B.C. this early accounting tool had developed into the first complete system of writing on clay tablets. Ironically, these were the first silicon-based storage technologies and would be abandoned by 2000 B.C. when the Egyptians developed papyrus-based writing materials. It would take almost four millennia before silicon would stage a comeback as the base material, with the main addition being the ability to process stored information. In 105 A.D. a Chinese court official named Ts'ai Lun invented wood-based paper. But it wasn't until Johann Gutenberg invented the movable-type printing press around 1436 that books could be reproduced cost effectively in volume. The first large book was the Gutenberg Bible, published in 1456. Something akin to Moore's law occurred, as Gutenberg went from printing single pages to entire books in 20 years. At the same time, resolution also improved, allowing finer type as well as image storage. Yet, this was primarily a storage mechanism. It would take at least another 400 years before retrieval would be an issue. In 1876, Melvil Dewey published his classification system that enabled libraries to store and retrieve all the books that were being made by that time. Alan Turing's "Turing Ma-

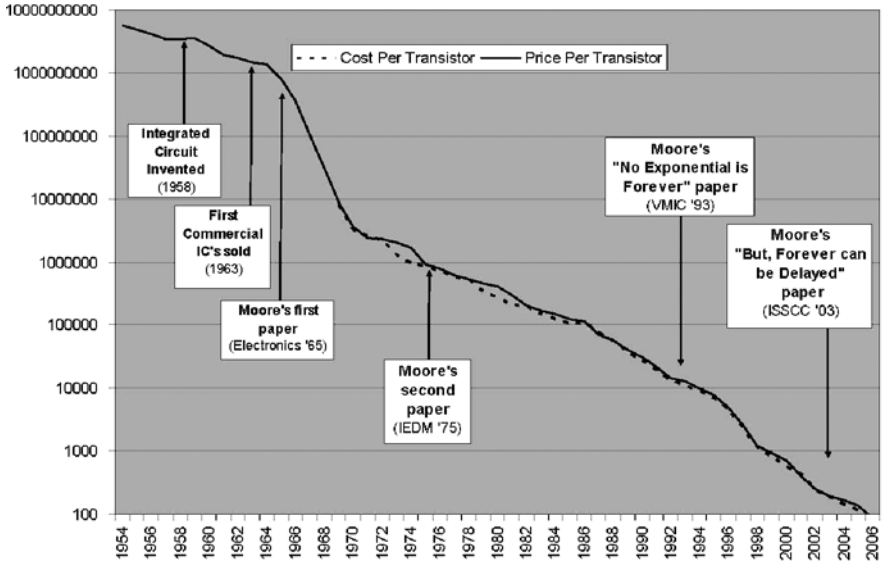

Fig. 2.5. Average price and cost per transistor for all semiconductors (in nanodollars)

chine," first described in 1936, was the step that would make the transformation from books to computers. So Moore's law can be seen to have a social significance that reaches back more than five millennia.

The economic value of Moore's law is also understated, because it has been a powerful deflationary force in the world's macro-economy. Inflation is a measure of price changes without any qualitative change. So if price per function is declining, it is deflationary. Interestingly, this effect has never been accounted for in the national accounts that measure inflation adjusted gross domestic product (GDP). The main reason is that if it were, it would overwhelm all other economic activity. It would also cause productivity to soar far beyond even the most optimistic beliefs. This is easy to show, because we know how many devices have been manufactured over the years and what revenues have been derived from their sales.

Probably the best set of data to use for analyzing the economic impact of Moore's law is simply price and cost per transistor. It is exceptionally good because it can easily be translated into a universal measure of value to a user: transistors. Transistors are a good measure because in economic terms they translate directly into system functionality. The more transistors, the greater the functionality of the electronic products consumers can buy. This data is shown in Fig. 2.5.

This data includes both merchant and captive production, so it is a complete measure of industry production. The constancy of this phenomenon is so stunning that even Gordon Moore has questioned its viability. The implications of this data are even more stunning: Since a transistor's price in 1954 was 64 million times more than it is as of this writing, the economic value the industry has brought to the world is unimaginable. If we take 2006's market and adjust for inflation, the value of to-

day's integrated circuit production would be 13 peta-dollars – or $13,000 trillion. That is far larger than Gross World Product, which measures the value of output of all the world's economies. Moreover, that doesn't include the value of all semiconductors! So it is hard to understate the long-term economic impact of the semiconductor industry.

2.6 Moore's Law Meets Moore's Wall: What Is Likely to Happen

Moore's law meets Moore's wall and then the show stops, or the contrary belief that there will be unending prosperity in the twenty-first century buoyed by Moore's law, have been recurring themes in the media and technical community since the mid-1970s. The pessimists are often led by conservative scientists who have the laws of physics to stand behind. The optimists are usually led by those who cling to "facts" generated by linear extrapolation.

The problem with the optimists is that the issues that loom are not easily amenable to measurement by conventional analysis. Eventually real barriers emerge to limit growth with any technology. Moreover, as Moore himself has often quipped, "No exponential goes on forever." But so far, the optimists have been right.

The problem with the pessimists is that they typically rely too much on known facts and do not allow for invention. They don't fully account for what they don't know, leaving out the "what they don't know" pieces when assembling the information puzzle. Yet it is the scientific community itself that expands the bounds of knowledge and extends Moore's law beyond what was thought possible. History is replete with many really good scientists and engineers who have come up with new things to constantly expand the boundaries of our knowledge, and as noted above, this is not likely to stop. When anyone asks me about Moore's wall, my first response is "Moore's wall is in Santa Clara, just outside Intel's Robert Noyce building. If you look closely, you will find the engraved names of people who made career-limiting predictions for the end of Moore's law." This has certainly been the case for those who have predicted the coming of Moore's wall in a five-or-ten-year span over the years. Yet, Moore himself said in 1995 that the wall should be finished and in place somewhere around 2040, when he poignantly pointed out that otherwise, "we'll be everything" if things continue at historical growth rates.

Herein lies the real dilemma. If our industry continues to grow unbounded, it really will become as large as the global economy in the first half of the twenty-first century. This leads to the historical view that as this occurs our industry's growth will become bounded by macroeconomic growth. However, if you look at history, it dispels this idea. At the beginning of this millennium rapid advances in agricultural techniques did not slow economic growth. Instead, they buoyed it as they freed-up human resources to work on other things, which in turn kicked off the High Middle Ages. Ultimately, this made possible the industrial age in the latter part of the millennium. As industry grew to be a larger part of the economy it did not slow to the 1% annual economic growth of agricultural economies. While it did slow, it also pushed economic growth up to an average of about 3%. Mechanized transportation

allowed centralized manufacturing, so factories could achieve greater economies of scale. This combined with the mechanization of the factory and greatly improved productivity; thus allowing greater non-inflationary growth levels. Since the latter half of the 1990s, the United States has been able to achieve regular non-inflationary growth of 4–5%. It is non-inflationary because of productivity gains. These gains are made possible by information technology.

Another factor driving the non-inflationary growth potential of the economy is that information technology tends to be energy saving as well. One of the real limits to the agricultural age was the fact that the primary fuel was wood. Entire forests were decimated in the Middle East and then Greece and Italy. The industrial age was prompted with the discovery of fossil fuels. This stopped deforestation to a great degree, but from an economic perspective, it also allowed for greater growth potential. Fossil fuels were easier to transport and use, so they too increased productivity. This, combined with the ability to transport materials to centralized manufacturing locations and then back out with trains, led to massive improvements in productivity. The information age takes the next step and relies on electricity. More important, it replaces the need to transport people, materials, and products with information. For example, video teleconferencing allows people to meet without traveling great distances. The voice and image information at both ends is digitized into information packets and sent around the world so that people can communicate without being physically near. At the same time, products can be designed in different places around the world and the information can be sent so products can be produced either in low-cost areas or, where transportation costs are high, locally. For example, for semiconductors being designed in the United States in close cooperation with a customer in Europe it is now a daily event to have the designs sent over the Internet to Texas for reticles to be made, to California for test programs, then to Taiwan to convert wafers into ICs, then to Korea for packaging, and finally the product is shipped to the customer in Europe. In the case of beer, transporting liquids is far too expensive. So a company in Europe can license its process to brewers in the United States and Japan, where they are manufactured locally. Using the Internet, the original brewer can monitor production and quality with little need to leave the home factory. The productivity effect seen in the transition from the agricultural to the industrial age is really happening as we move into the information age.

It can be argued that macroeconomic growth could rise to as high as 8% while creating a similar growth cap for our industry. What happens when this occurs? It is inevitable that the semiconductor industry's growth will slow from the 15–20% range it has averaged over its history in the last half of the twentieth century. The barriers that will limit it will be economic not technical, as Moore's law is a statement of powerful economic forces [15]. Technology barriers first show up as rising costs that go beyond the bounds of economic sense. Transportation speed limits could exceed the speed of sound. But economic limits make private jet ownership unattainable for all but a very few. Economic limits make the automobile the most commonly used vehicle in major industrialized countries and the bicycle in others. But even here, the economic limits of building infrastructure limit average speed to less than 20 MPH in

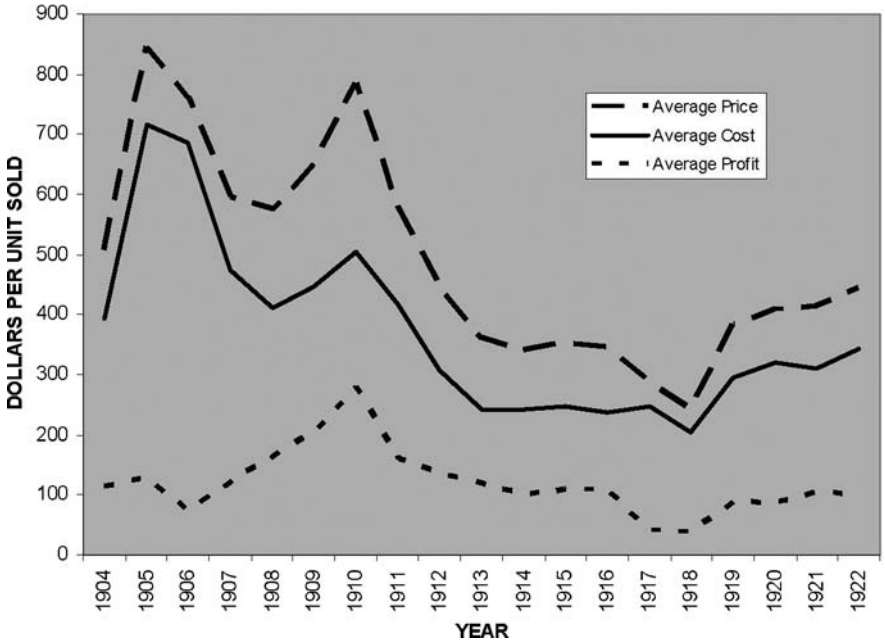

Fig. 2.6. Ford motor company's equivalent to Moore's law (the early years of the auto industry)

industrial countries (which is one reason why the bicycle has become such a popular alternative).

If we look to the auto industry for guidance, similar declines in cost during its early years can be found. At the turn of the century, cars were luxury items, which typically sold for $20,000. They were the main frames of their day, and only the ultra-rich could afford them. Henry Ford revolutionized the auto industry with the invention of the assembly line. Ford's efforts resulted in a steady reduction in costs, quickly bringing the cost of manufacturing a car to under $1000. But even Ford's ability to reduce costs had bottomed out by 1918, when the average hit a low of $205 (see Fig. 2.6, which has not been adjusted for inflation).

While these efforts pale in comparison to gains made in semiconductors, the lesson to be learned is that cost gains made on pushing down one technical river of thought will eventually lead to a bottom, after which costs rise. Science and engineering can only push limits to the boundaries of the laws of physics. Costs begin to escalate as this is done because the easy problems are solved and making the next advance is more difficult. At some point, little gains can be made by taking the next step, but the cost is astronomical. In the case of automobiles, the gains were made by the development and improvement of assembly line technology. In the case of semiconductors it has largely been lithography where the gains were made.

These are not "economies of scale" as taught in most economics classes, where increased scale drives cost down to a minimum – after which, costs rise. Instead, technology is driving cost. These are economies of technology and are some of the

most important underlying factors that make Moore's law possible and will ultimately result in its demise when gains can no longer be made.

Similar things are happening in semiconductors. Fab equipment prices have risen steadily at annual rates above 10%. This was fine as long as yields rose, giving added economic boost to the cost of steadily shrinking transistors to stay on Moore's curve. But yields cannot go up much further, so gains will have to come from productivity improvements.

It is important to note that as these economic barriers are hit, it does not mean the end of the semiconductor industry. The industry has lived with Moore's law so long that it is almost of matter of faith, as exemplified in the term "show stopper." The term has been used extensively to highlight the importance of potential limits seen in the industry's "road mapping" of future technologies. Yet it is unlikely that the show will stop when the show stoppers are finally encountered. Just think of the alternatives. Moreover, the auto industry has been quite healthy in the eight decades since it hit its show stoppers. People did not go back to horses as a means of regular transport. As the gains from automation petered out, auto manufacturers shifted their emphasis from low-cost one-size-fits-all vehicles to many varieties – each with distinct levels of product differentiation. The other hallmarks of the industrial age trains and planes also found ways to go on after they hit technical and economic limits. For this to happen in semiconductors, it means manufacturing will have to be more flexible and design will continue to become more important.

2.7 Conclusion

Moore's law has had an amazing run as well as an unmeasured economic impact. While it is virtually certain that we will face its end sometime in this century, it is extremely important that we extend its life as long as possible. However well these barriers may be ultimately expressed economically, barriers to Moore's law have always been overcome with new technology. It may take every ounce of creativity from the engineers and scientists who populate this industry, but they have always been up to the task. Moore's law is predicated on shrinking critical features. Since the 1970s, it has always seemed that we are fast approaching the limits of what can be done, only to find someone had come up with a new idea to get around the barrier. The "red brick wall" has proved more imaginary that real – it's real effect having been to spur innovation.

So what advice would Gordon give us? I had the chance to ask him several years ago on the day he entered retirement.[10] One thing he wanted to point out was that he never liked the term Moore's law: "I've learned to live with the term. But it's really not a law; it's a prediction. No exponential runs forever. The key has always been our ability to shrink dimensions and we will soon reach atomic dimensions, which are an absolute limit." But the question is not if, it's when will Moore's wall appear? "Who knows? I used to argue that we would never get the gate oxide thickness below

[10] Personal conversations with Dr. Gordon Moore and the author, May 24, 2001.

1000 angstroms and then later 100. Now we're below 10 and we've demonstrated 30-nm gate lengths. We can build them in the 1000's. But the real difficulty will be in figuring out how to uniformly build ten's of millions of these transistors and wire them together in one chip." The key is to keep trying. Keep trying we did: As of this writing 30-nm gates are just making it to production.

Gordon felt that any solution must champion manufacturing because "there is no value in developing something that cannot be built in volume. Back at Fairchild the problem was always in getting something from research into manufacturing. So at Intel we merged the two together." He advised, "Always look for the technical advantage (in cost). I knew we could continue to shrink dimensions for many years, which would double complexity for the same cost. All we had to do was find a product that had the volume to drive our business. In the early days that was memories. We knew it was time to get out of memories when this advantage was lost. The argument at the time was that you had to be in memories because they were the technology driver. But we saw that DRAMs were going off in a different technical direction because problems in bit storage meant they had to develop all these difficult capacitor structures."

He also pointed to the need to avoid dependency on specific products. "I've never been good at forecasting. I've been lucky to be in the right place at the right time and know enough to be able to take advantage of it. I always believed in microprocessors but the market wasn't big enough in the early days. Ted Hoff showed that microprocessors could be used for calculators and traffic lights and the volume could come in what we now call embedded controllers. I continued to support it despite the fact that for a long time the business was smaller than the development systems we sold to implement them. But just when memories were going out, microprocessors were coming into their own. Success came because we always sought to use silicon in unique ways."

So what did Gordon have to say about his contribution and the future of our industry: "I helped get the electronics revolution off on the right foot … I hope. I think the real benefits of what we have done are yet to come. I sure wish I could be here in a hundred years just to see how it all plays out."

The day after this discussion with Gordon, I knew it was the first day of a new era, one without Gordon Moore's oversight. I got up that morning half-wondering if the sun would rise again to shine on Silicon Valley. It did – reflecting Gordon Moore's ever-present optimism for the future of technology. As has Moore's law, which continues to plug on, delivering benefits to many who will never realize the important contributions of this man and his observation.

Appendix A

Moore's law governs the real limit to how fast costs can grow. Starting with the basic equations (2.1) and (2.2), the optimal component density for any given period is:

$$C_t = 2 \cdot C_{t-1},$$

where C_t = Component count in period t, C_{t-1} = Component count in the prior period (Also please note the "−1" here and below is symbolic in nature and not used mathematically.)

According to the original paper given in 1965, the minimal cost of manufacturing a chip should decrease at a rate that is nearly inversely proportional to the increase in the number of components. So the cost per component, or transistor, should be cut roughly in half for each tick of Moore's clock:

$$\begin{aligned} M_t &= \frac{M_{t-1}}{2} \\ &= 0.5 \cdot M_{t-1}, \end{aligned}$$

where M_t = Manufacturing cost per component in period t, M_{t-1} = Manufacturing cost component in the prior period.

However, since this paper was first given, it has generally been believed that industry growth will not be affected if the cost per function drops by at least 30% for every doubling of transistors. Thus:

$$M_t = 0.7 \cdot M_{t-1}$$

since,

$$\begin{aligned} M_t &= \frac{\text{Tdc}_t}{C_t} \quad \text{and,} \\ M_{t-1} &= \frac{\text{Tdc}_{t-1}}{C_{t-1}}, \end{aligned}$$

where Tdc_t = Total die cost in period t, Tdc_{t-1} = Total die cost in the prior period.

Thus,

$$\begin{aligned} \frac{\text{Tdc}_t}{C_t} &= \frac{0.7 \cdot \text{Tdc}_{t-1}}{C_{t-1}}, \\ \frac{\text{Tdc}_t}{2 \cdot C_{t-1}} &= \frac{0.7 \cdot \text{Tdc}_{t-1}}{C_{t-1}}, \\ \text{Tdc}_t &= \frac{2 \cdot C_{t-1} \cdot 0.7 \cdot \text{Tdc}_{t-1}}{C_{t-1}} \end{aligned}$$

Simplified it reduces to:

$$\begin{aligned} \text{Tdc}_t &= \frac{2 \cdot 0.7 \cdot C_{t-1} \cdot \text{Tdc}_{t-1}}{C_{t-1}}, \\ \text{Tdc}_t &= 1.4\text{Tdc}_{t-1}. \end{aligned}$$

If the cost-per-function reduction ratio is different than 0.7, then

$$\text{Tdc}_t = 2 \cdot \text{Cpfr} \cdot \text{Tdc}_{t-1},$$

where: Cpfr = Cost-per-function reduction ratio for every node as required by the market.

In general, the manufacturing cost per unit area of silicon can rise by 40% per node of Moore's law (or by twice the cost-per-function reduction ratio requirement. This includes everything from the fab cost to materials and labor. However, it does not take yield or wafer size into account. Adding these two:

$$\mathrm{Twc}_t = 2 \cdot \mathrm{Cpfr} \cdot \mathrm{Twc}_{t-1}.$$

So,

$$\mathrm{Tdc}_t = \frac{\mathrm{Twc}_t}{\mathrm{Dpw}_t \cdot Y_t} = \frac{2 \cdot \mathrm{Cpfr} \cdot \mathrm{Twc}_{t-1}}{W \cdot \mathrm{Dpw}_{t-1} \cdot \mathrm{Yr} \cdot Y_{t-1}},$$

where Twc_t = Total wafer cost requirement in period t, Twc_{t-1} = Total wafer cost in the prior period, Dpw_t = Die-per-wafer in period t, Y_t = Yielded die-per-wafer in period t, W = Ratio of die added with a wafer size change, Dpw_{t-1} = Die-per-wafer in the prior period, Yr = Yield reductions due to improvements with time. Y_{t-1} = Yielded die-per-wafer in the prior period.

References

1. G.E. Moore, Lithography and the future of Moore's law. SPIE **2440**, 2–17 (1995)
2. G.E. Moore, The future of integrated electronics. Fairchild Semiconductor (1965). This was the original internal document from which *Electronics Magazine* would publish "Cramming more components into integrated circuits," in its April 1965 issue celebrating the 35th anniversary of electronics
3. G.E. Moore, Progress in digital integrated electronics, in *International Electron Devices Meeting* (IEEE, New York, 1975)
4. K. Marx, *Capital* (Progress, Moscow, 1978). Chap. 15, Sect. 2
5. J.S. Mill, *Principles of Political Economy* (London, 1848)
6. G. Ip, Did greenspan push high-tech optimism on growth too far?. The Wall Street Journal, December 28, 2001, pp. A1, A12
7. H.R. Huff, John Bardeen and transistor physics, in *Characterization and Metrology for ULSI Technology 2000*. AIP Conference Proceedings, vol. 550 (American Institute of Physics, New York, 2001), pp. 3–29
8. G.D. Hutcheson, The chip insider. VLSI Research Inc., September 17, 1998
9. Scientific American Interview: Gordon Moore, in *Scientific American*, September 1997
10. W.R. Runyan, K.E. Bean, *Semiconductor Integrated Circuit Processing Technology* (Addison-Wesley, Reading, 1990), p. 18
11. C.E. Spork, *Spinoff* (Saranac Lake, New York, 2001)
12. R.H. Dennard, IBM – field-effect transistor memory. US Patent 3,387,286, Issued June 4, 1968
13. G.D. Hutcheson, The VLSI capital equipment outlook. VLSI Research Inc., 1987
14. G.D. Hutcheson, The chip insider. VLSI Research Inc., May 25, 2001
15. G.D. Hutcheson, J.D. Hutcheson, Technology and economics in the semiconductor industry. Sci. Am. **274**, 54–62 (January, 1996)
16. J. Diamond, *Guns, Germs, and Steel* (W.W. Norton, New York, 1997)

http://www.springer.com/978-3-540-74558-7

Into The Nano Era
Moore's Law Beyond Planar Silicon CMOS
Huff, H. (Ed.)
2009, XXVIII, 348 p. 136 illus., Hardcover
ISBN: 978-3-540-74558-7