#2.

Establishing Moore's Law

Ethan Mollick
MIT Sloan School of Management

Every field has brief formulas or relationships that are useful for back-of-the-envelope calculations. Rarely do these maxims become popular knowledge; even more rarely do they become as ubiquitous and influential as Moore's law, the 40-year-old prediction that the speed of computers will double every year or two. Here, a look at the way in which the legendary law evolved into a self-fulfilling prophecy.

The seemingly unshakeable accuracy of Moore's law—which states that the speed of computers, as measured by the number of transistors that can be placed on a single chip, will double every year or two—has been credited with being the engine of the electronics revolution, and is regarded as the premier example of a self-fulfilling prophecy and technological trajectory in both the academic and popular press.[1] Predictions made using the law become the basis for future production goals, which in turn reinforces the validity of the law as a measurement of industry progress. In a rapidly changing environment, Moore's law has been described as "the only stable ruler" on which companies can rely.[2]

Glowing statements such as these seem curiously out of place when compared with the law's actual performance. Gordon E. Moore, a physical chemist who cofounded both Intel and Fairchild Semiconductor, first articulated what would later be identified as his law in a 1965 paper for *Electronics* magazine's 35th anniversary (for a biographical sketch of Moore, see the Biographies department in this issue). In his original 1965 paper, Moore asserted that the number of components that could be placed on a chip could be expected to double every year. Just 10 years later, Moore revised his prediction, stating that "the new slope might approximate a doubling every two years, rather than every year."[3] Yet another form of Moore's law can be found in many recent technical and popular publications. A good representative of this version of Moore's law is the statement that appeared in one prominent business and technology journal: "Moore's law, celebrated as the defining rule of the modern world, states that the computer chip performance doubles every eighteen months."[4] The apparent differences among these definitions make it difficult to view Moore's law as a single prediction, let alone an influential concept in the semiconductor industry.

It is precisely these changes that have made the law seem accurate, however. The semiconductor industry has undergone several dramatic transformations over the past 40 years, rendering irrelevant many of the original assumptions embodied in Moore's law. These changes in the nature of the industry have coincided with periodic revisions of the law, so that when the semiconductor market evolved, the law evolved with it. This has had the effect of making Moore's law seem to be a single, accurate statement, especially for the first 20 years of its existence, when, in fact, it is a number of different laws that have replaced each other in succession.

The seeming power of Moore's law has attracted innumerable popular commentaries but little scholarly examination. Other authors have covered some of the history of Moore's law, specifically Probir K. Bondyopadhyay's description of the law's origins[5] and Robert Schaller's excellent general overview of the law and its importance.[6] But there has yet been no detailed examination of how the law has evolved as a sequence of related laws, or how it eventually became fixed as a fact in the industry through foreign competition.

The history of Moore's law is more than a curiosity, since it is important not just in the semiconductor industry but also as a key example of technology trajectories and self-fulfilling prophecies. Donald MacKenzie, in his work on technical trajectories, pointed out that understanding Moore's law could shed light on how technical patterns form:

> Yet in all the many economic and sociological studies of information technology there is scarcely a single piece of published research … on the determinants of the Moore's law pattern. Explicitly or implicitly, it is taken to be a natural trajectory whose effects economists and sociologists may study but whose causes lie outside their ambit …

 Published by the IEEE Computer Society 1058-6180/06/$20.00 © 2006 IEEE

> until such a study of Moore's law is done, we cannot say precisely what intellectual opportunities are being missed, but it is unlikely that they are negligible.[7]

By examining how Moore's prediction became the immutable Moore's law as a dynamic, rather than static, process, it is possible to begin to bring to light some of determinants discussed by MacKenzie.

Technical components of Moore's law

The law itself may have changed, but one constant throughout the varying formulation of Moore's law is that all formulations observe that the number of transistors per chip, also known as complexity, has been growing rapidly. Four main factors are responsible for most of this increase in chip density: die size, line dimension, technical cleverness, and technical innovation. Together, they serve to explain most of the exponential improvements in transistors per chip described by Moore's law.

Die size refers to the surface area of a silicon wafer and the chips that are made out of it. As size increases, more components can be placed on a chip, but the chance of flaws in the silicon rendering the device unusable is also increased. Line dimension, on the other hand, refers to the depth, width, and density at which circuitry can be etched onto silicon. The progress in producing new photolithographic techniques and the finer line dimensions that result are the second major factor in determining chip density.

The third factor is "technical cleverness," a category that describes trends in the chip design process, rather than in the technical production of chips, as the first two factors do. The first element of technical cleverness is, to borrow a term coined by Moore in 1975, *device cleverness*. Device cleverness refers to the ability of engineers to design chips that use increasingly more of a chip's space through clever circuit design, rather than through technological advances in chip production techniques. The second part of technical cleverness is the contribution of computer-aided design (CAD) to the development of complex chips. Both factors have led to the creation of technically clever integrated circuit (IC) designs that more fully use advances in line dimension and die size.

The final element that contributes to increasing chip density is technical innovation, which stems from the nature of digital electronics. This term describes advances in materials and methods of chip production, such as improvements in logic families and the development of new substrates. Although other improvements have occurred in the semiconductor industry, die size, line dimension, technical cleverness, and technical innovation continue to be the main determinants of chip density, the increase of which Moore's law is supposed to measure as a function of time.

The final part of the framework required to analyze the law is to examine the "time" portion of the Moore's law equation. The evolution of the semiconductor industry is, by convention, separated into several periods, based on the number of components that can be placed on a chip, called the level of integration. The period from 1960 to 1965 is considered the time of small-scale integration (SSI) when chips contained fewer than a hundred components. It was followed by medium- and large-scale integration (MSI and LSI), with chips of up to 100,000 components, a period that lasted until the mid-to-late 1970s. Finally, there is the current period of very large-scale integration, VLSI, with chips of greater than 100,000 components. While the literature, almost without exception, takes the existence of these periods for granted, the divisions between them can vary greatly, depending on the source. That is because the periods were not originally defined in any precise way, but rather grew from the nature of technological progress in the semiconductor business.

Despite the seemingly linear progression of period names (from small to medium scale and so on), the transitions between integration levels were not smooth, but indicate fundamental shifts in the way the industry operates. In the IC business, change tends to come in large, somewhat foreseeable waves because of long R&D periods, which force companies to anticipate future needs.

As a result of long-range planning, firms can anticipate approximately when current production methods will no longer be adequate to manufacture the predicted products. Usually, companies can plan to improve production to meet future requirements through simple, evolutionary improvements to current methods. Occasionally, a truly innovative improvement in one factor, such as line dimension, will be necessary to keep products on schedule. In certain rare situations, however, a consensus grows within the industry that, in the relatively near future, semiconductor manufacturers will reach a barrier that will be completely insurmountable using simple variations on current techniques.

The shift between different production mechanisms mark the transition between periods, and they are often characterized by one

particular, symbolic product that the industry hopes to achieve. For example, the SSI period was defined by the challenge of placing a single logic gate on a chip; MSI was reached when a register—a basic means of electronic storage—was built using an IC. The beginnings and ends of periods are seldom precisely delineated and definitions can vary, especially in the years right after a new period has begun. Still, there is a clear, if unspoken, consensus among engineers and planners about when a new period will begin, and on what challenges characterize it.

Although the transitions between periods are derived from technical factors, they can also be described in terms of dates and integration levels. Of course, deciding that the date by which 100,000 transistors were placed on a chip marked the beginning of VLSI ignores the economic and technical reasons behind the VLSI barrier. From a historical perspective, however, the period in which 100,000 component chips were developed coincides with the years in which the technical difficulties arising with VLSI were solved. The challenges marking other periods are similarly associated with their specific level of integration. Thus, levels of integration are a useful indicator for a wide variety of critical challenges to the semiconductor industry.

In the context of integration levels, the timing of Moore's three major papers on what would eventually be called Moore's law is especially significant.[8] Each paper was published during the transition between one of these periods. The original article was written in 1965, at the end of the SSI period; the second in 1975, around the time that VLSI was coming into use; and the final in 1995, preceding a decade of rapid change in approaches to VLSI technology. All three papers give subtly different versions of Moore's law. The changes in the form of the law reflect the nature of the transition between integration levels. The result is a law that seems accurate, but remains so only because it has been modified to keep the law from one period relevant in another. The law's evolution has traced the growth of the semiconductor industry, beginning with its origins in the SSI period.

Moore's law in the SSI period

Moore's law was first developed early in the history of the semiconductor industry. When Moore's aforementioned paper was published in 1965, it was less than two decades since the invention of the transistor. At the time of Moore's original paper, ICs were being used in functions that could also be performed by discrete electronics. ICs were generally acknowledged as having the advantages of "lower costs and greatly simplified design," but otherwise were not fundamentally different from components wired together by hand.[9] That attitude implied that, at least in the short term, the microelectronics industry saw the discrete circuit as its competition, rather than viewing the IC as a unique product capable of doing things impossible with discrete elements alone. The economics of integrated versus discrete circuits was therefore a primary concern of semiconductor manufacturers.

To Moore and other high-level researchers at that time, technical concerns were secondary to these economic issues. It was generally believed that the industry could develop the production techniques needed for advanced products, as long as these products were economically justified. Moore explained in his paper, for example, that the building of 65,000 component chips would be possible "with the dimensional tolerances already being employed in integrated circuits," leaving the question of "under what circumstances we should do it."[9] This view matched the conditions of the semiconductor industry of the time: the MSI/LSI period was starting, with its major technical challenge of placing a logic gate on a chip having been achieved in the laboratory two years earlier. Still, the industry had yet to see a large commercial success. The new period was seen as one in which economic justifications were needed to continue technical advances.

Moore's paper must be understood in light of this unstated assumption, because the original version of his law was based on economic, rather than technical, reasoning. Moore noted that the cost of producing a chip is based on the expense of producing and etching a silicon wafer, and not on the number of components placed on a chip. Thus, cost per component decreases as chip density increases. A countervailing factor is "as more components are added, decreased yields more than compensate for the increased complexity, tending to raise the cost per component."[9]

These decreasing yields came from the random defects that tainted the silicon wafers. The more wafer that was used, the greater the likelihood that a defect would mar an active part of the chip, rendering it useless. The combination of decreasing cost per component and increasing defects with complexity described a U-shaped cost curve, with the minimum point representing the component density at which the overall price per component, as determined by both factors, was lowest (see Figure 1).

Moore took this discovery one step further. Noting that the ability to manufacture defect-free chips at low cost increases with time, he plotted out these U-curves of cost per component versus the number of components per chip, using historical data from the past few years. When the minimum points of these curves were connected, these points formed a line corresponding to a doubling of complexity every year (see Figure 2). Moore, in what would be the basic statement on Moore's law, noted the following based on his five-point graph:

> Certainly, over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years. That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000.[9]

This forecast was the first to be associated with Moore's law, and it proved quite accurate. Although Moore predicted a certain technical level of achievement by a certain date, he did not arrive at it by the examination of technological factors. Moore confirmed this point in an interview, stating that his paper was an attempt to show the cheapest way to produce microchips, and that he was surprised when it had implications beyond his original intentions.[10] This subordination of technical achievement as a function of economics is characteristic of the SSI to MSI/LSI transition.

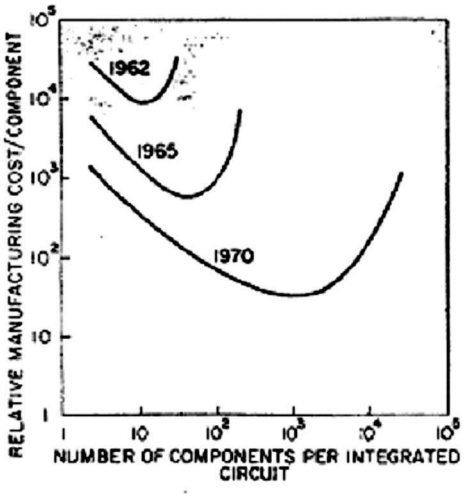

Figure 1. Number of components per integrated circuit. (Courtesy Penton Media. Reprinted from G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 19 Apr. 1965, pp. 114-117.)

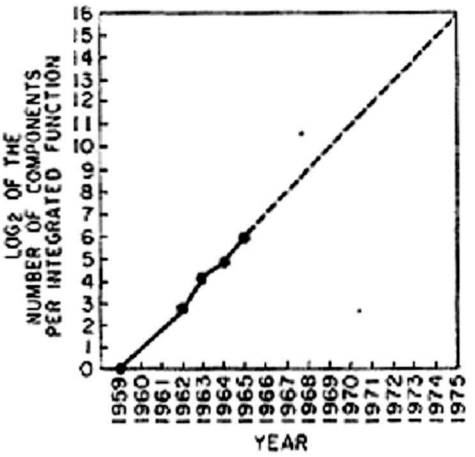

Figure 2. Log_2 of the number of components per integrated function. (Courtesy Penton Media. Reprinted from G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 19 Apr. 1965, pp. 114-117.)

Moore's law in the MSI/LSI period

By the time that the MSI and LSI periods reached their conclusion in the mid-1970s, the basic assumptions behind Moore's law had collapsed, even though its prediction came true. Instead of economics driving technical advancement, the situation had changed so that technical concerns were of similar importance to economic ones for the semiconductor industry. Dramatic advances in technology during the MSI/LSI era had changed the economics of the production of semiconductor products. Additionally, new innovations had altered the market for ICs. No longer were they substitutes for discrete components, because new products had been developed that could only be implemented using integrated functions. Moore recognized the growing importance of technical factors in his speech to the 1975 Electron Devices Meeting about Moore's law. At the beginning of his talk he stated that "Many aspects of processing and design technology have contributed to make the manufacture of such functions as complex single chip microprocessors or memory circuits economically feasible."[11] The economic argument at the heart of the original law—that integrated functions needed to be competitive with discrete circuits—had vanished, so the law was subtly, perhaps unconsciously, modified to better reflect the nature of the transition between LSI and VLSI.

To accomplish this, Moore began to refer to technical factors in his discussion of the past and future of the law. He decomposed the trend of increasing complexity into a variety of process improvements, and used advances in technology, rather than economics, to explain the progress of the semiconductor industry.[12] Even more dramatic than the changes in the paper were the shifts in the context in which Moore placed his law.

These changes, covered by the technical innovation factor, occurred in both chip production and the uses of ICs. Three major new uses were found for microelectronics: dynamic random access memory (DRAM), microprocessors, and universal asynchronous receiver-transmitters (UARTs), which provide a means of communication between electronic devices. DRAMs and microprocessors rapidly became the two most critical semiconductor products, a change that was reflected in the new understanding of Moore's law.

The DRAM is a significant product in a number of ways. First, because DRAM used a large number of regular, repeating cells, the chips were easy to design. That allowed DRAM manufacturers to use the most advanced production techniques to make the chips as complex as possible. In addition, DRAM became the critical market for semiconductor manufacturers. The demand for memory chips has historically proved almost bottomless, and DRAM sales are often a significant portion of the IC business as a whole. Because DRAMs were an important product and tended to be the most complex chips available, they were often used when indicators of the state of the semiconductor industry were needed. Moore was no exception. By the beginning of the VLSI period, he used DRAM chips as data points in graphs of the Moore's law curve.[10]

DRAMs and microprocessors became critical to the semiconductor industry, yet were unknown during the original formulation of Moore's law. To accommodate them, the law began an unstated evolution. Microprocessors, although complex circuits, tended to provide more design challenges than production problems. They were rarely as transistor-dense as DRAM chips, and thus would appear below the law's curve. The characteristics of DRAMs, however, made them ideal indicators of the chip complexity frontier. The result was a displacement of microprocessors from complexity curves in favor of DRAMs. By the early 1980s, the Moore curve was viewed as a rule about "the densities of dynamic RAMs," not as a general principle. With time and the increasing importance of microprocessors in the VLSI era, this would change again, but in the late 1970s, DRAMs became the most important chip produced, and therefore the focus of Moore's law.

During the MSI/LSI period, technical innovation occurred not just with new types of circuits but also in the development of new logic families. MOS (metal oxide semiconductor), CMOS (complementary MOS), and charge-coupled devices (CCDs), were all introduced. CCDs were seen as especially significant because they were expected to be the key to new kinds of memory chips.

Against this background of radically changing technical innovation, there was also significant progress in the other three factors that contribute to increasing complexity. Two of these, die size and line dimension, were identified by Moore in his paper as areas in which progress had led to increased chip complexity. Both factors improved significantly, although not as much through any particular breakthroughs as through a slow and steady evolution. Moore graphed these changes and identified the degree to which each element had contributed to the increases in chip complexity since 1965. Together, according to Moore, they accounted for a little over two-thirds of the increase in transistor count.

The remainder Moore attributed to the final factor of technical cleverness. During the MSI and LSI periods, circuit designers became much more adept at using the silicon chip's full area. This was a function of the advent of CAD, along with the development of memory chips, which used a simple, easy-to-design cell structure. In Moore's 1975 analysis of chips' increasing density, he said of technical cleverness that "this contribution to complexity has been more important than either increased chip area or finer lines."[13]

Despite the significant progress in technical cleverness, Moore predicted that this factor would be much less significant in determining chip density in the future. His reasoning was based on the CCD memories that he knew were in development. CCDs "can approach closely the maximum density practical," leaving little room for improvements due to clever design.[3] Because of his knowledge of the industry, he knew that CCD chips were expected to become widely available by the end of the decade.[10] When CCDs were released, the contribution from technical cleverness would disappear, and the slope of the law would change, from doubling every year to once approximately every two years by 1980.[3]

What Moore did not realize was that CCD memories would turn out to be a failure owing

to their sensitivity to natural background radiation. The failure of CCDs meant that, instead of the law changing within five years, it changed almost immediately. As a result, the reformulated Moore's law of 1975 was wrong, although the slope Moore predicted for the new curve turned out to be correct.[14] By the late 1970s, the doubling time was around two years, just as Moore had forecast, even though the actual number of components per chip was lower than Moore had predicted, because of the more rapid flattening of the curve. In the end, the CCD experience served less to demonstrate the law's weakness than to point out two important strengths.

The first strength was that Moore could successfully change the slope of his law to better reflect the nature of the transition to VLSI. Effectively, this created a different Moore's law curve, but without the historical precedent that supposedly formed the basis for believing the original Moore's law. Moore was able to do this because he phrased his modification as a forecast, rather than by adapting the law after the slope had already changed. Instead, it appeared that Moore's changed forecast was remarkably prescient, because it was made before the slope of the curve decreased. In this way, Moore was able to maintain the law's credibility, even after it was modified.

Moore's modifications to the law were made casually rather than out of an attempt to inflate the apparent accuracy of his predictions, because, even in the 1975 presentation, he did not see his estimates as significant. In an interview, Moore said of his 1975 effort, "I don't think anybody was planning their business around it, probably because I was basking in the glow of the first prediction being right. I don't think anyone was paying any attention to it."[15] This lack of concern about the law's significance is part of what allowed Moore to modify it without worrying that he was doing anything more than continuing a tradition of prediction that he had started in a paper a decade earlier. The law's growing significance in the coming decade would make the multiple predictions seem part of a single, unavoidable phenomenon, rather than a set of different estimates.

The CCD incident also illustrated a second important strength of the law, one related to the first. Moore's mistake over CCDs indicated that he was dependent on inside-industry knowledge to make predictions. CCDs were not yet on the market, but Moore already had a good idea of when they would be released and what their effect on the industry would be. He was as surprised as anyone in the industry when these expectations failed to materialize. Moore has acknowledged that his insider's view gives him industry knowledge "a year or two" ahead of most people.[10] Similarly, the last point in his 1965 paper was based on a product being developed at Fairchild but not yet released.[15] Given this foreknowledge, the prediction of a leveling of the complexity curve is less impressive, because Moore, just like other industry insiders, could have known that the engineers at Intel were unable to develop more technically clever devices at the same level of economic efficiency. The result was that some portion of the accuracy of Moore's law, at least from outside the semiconductor industry, had to do with insights from within Intel, rather than from pure prediction.

Moore's modifications to the law were made casually rather than out of an attempt to inflate the apparent accuracy of his predictions.

The combination of these elements meant that, post-1975, Moore's law was as much a product of the VLSI transition as the Moore's law of 1965 was a result of the shift to MSI/LSI. The law's new version no longer assumed that the rate of technical progress was solely the result of economic factors; instead it recognized that technological advancements in areas such as line density had their own, independent rates of improvement. Indeed, the law's entire context, derivation, and meaning had changed, but it still had the historical validity of a 10-year-old, empirically proven observation. The law was now ready to enter the VLSI period.

Moore's law in the VLSI period

By the time Moore wrote his 1995 paper, his law, as it had come to be defined, had proven to be a remarkably accurate indicator of chip complexity for the VLSI era, which started in the late 1970s. With this success came increased notice, and the law became accepted as one of the few stable rules in the otherwise unpredictable computer industry and also became, to some extent, a dependable guide for the future. Following the traditional pattern of

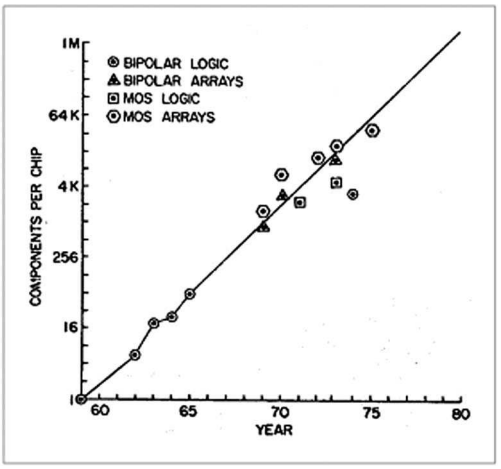

Figure 3. Circuit complexity versus time of introduction. (Reprinted with permission from R. Noyce, *Science,* vol. 195, 1977, pp. 1103-1105.)

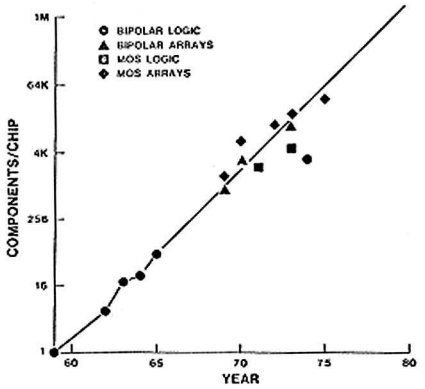

Figure 4. Approximate component count for complex integrated circuits vs. year of introduction. (© 1975 IEEE; G. Moore, "Progress in Digital Electronics," *Technical Digest of the Int'l Electron Devices Meeting,* IEEE Press, 1975, vol. 13.)

Moore's law, at least part of this success came from a further evolution in the meaning of the law, which kept it relevant in the VLSI period.

One of the major contributing factors to the widespread impression that Moore's law was accurate was a series of articles at the beginning of the VLSI era that both distorted and popularized its historical validity. In the late 1970s, ICs began to attract public attention. The invention of the microprocessor and the advent of personal computers made the chip an icon for the "computer revolution," prompting widespread interest. On the early part of this wave, the first major publications that printed generally accessible features on semiconductors were *Science* and *Scientific American*, both in 1977. In these publications, articles about Moore's law appeared that were written by Robert Noyce, Moore's partner in founding Intel and, earlier, Fairchild Semiconductor. These two articles and the particular way in which they interpreted the law were significant in establishing Moore's law as a seemingly infallible predictor of progress.

The pieces were similar in many respects, although the *Science* article was slightly more technical than the one in *Scientific American*. Both, however, discussed the trends in the semiconductor industry that led to increasing complexity, the same factors that Moore's law addressed. One way the articles explained these trends was by using graphs of complexity, minimum dimensions, and die area, all plotted against time. These were the same variables plotted in Moore's 1975 paper on the future of the semiconductor. The similarity goes beyond simply the topics of the charts. The graphs used by Noyce (see Figure 3) are identical to those Moore used (see Figure 4), and are not even updated for chips developed between 1975 and 1977.

Despite their use of the same graphs, the nature of the law changed between the two authors' treatments. Noyce's primary graph, showing the trend of increasing components with time, is labeled "circuit complexity versus time of introduction."[16] It is a reproduction of Moore's 1975 graph labeled "Approximate component count for complex integrated circuits vs. year of introduction."[11] The main difference between these descriptions is in the phrasing "complex integrated circuits," and it is significant. For his paper, Moore was selecting his data points using specific criteria, graphing only "complex integrated circuits," rather than serving as a representative sample of the circuits available, as seems to be implied in Noyce's article.

The key to understanding Moore's definition of "complex" circuits can be found in the 1975 graph. Moore's charts show a regression line linking together four points in the early 1960s with a large cluster of points in the early 1970s. Between these two groups of points, in the years 1965 through 1968, there are no ICs graphed. This is unusual because great advances in semiconductors were occurring at that time, including the entire MSI period of integration.

The explanation of this graph can be found in another paper by Moore on his law, this one from 1979. There, he once again reproduces his famous graph, stating, "Note the gap between 1965 and 1968. This occurred because it was difficult at the time to identify any semiconductor products whose complexity came close

to the potential limit."[17] This quote reveals two things. First, that Moore viewed his law as expressing the upper potential bounds for complexity at any given time, and, second, that Moore did not graph points that did not approach this expected complexity level.

This editing of data made sense from Moore's perspective. He was trying to demonstrate the way in which the most economically efficient complexity level for chips increased with time. For Moore, economic efficiency meant the efficiency of production. He was concerned about producing circuits at the lowest cost per component. This view ignored the marketability and design aspects of the semiconductor industry. Simply because it is possible to build a complicated chip does not mean that one should build it, as Moore himself noted in his 1965 paper. At times, designers have been unable to develop commercial products that take full advantage of the most efficient means of production available. For Moore's purposes, as he stated during an interview, he was interested in the commercially available technology frontier.[15]

This is exactly what happened in the late 1960s. ICs were relatively new, and advances in production were occurring faster than new chip designs could apply them. Partially, this was because basic, foundational work in the production of digital electronics was still being done, and important new advances, such as the development of the logic gate, were made without using a large number of integrated components.[18] Moore did not include chips of that type in his graph, because they failed to illustrate the increasing production efficiency of ICs.

In Noyce's article, the context of production efficiency is lost entirely, yet it uses the same graph as the Moore paper. This makes it appear, at least in Noyce's article, that there has been an uninterrupted growth in chip complexity over almost 20 years. In fact, a graph that included the points from 1965 to 1968 would look much less regular, because they would fall below the predicted trendline.[19] Since Noyce's articles were popular and often cited, it was his view of a universal and accurate Moore's law that became fixed in the public's perception rather than Moore's more qualified predictions. The result was the widespread belief in the historical fact of a law that, at various times, has been far from exact. Also obscured by the Noyce article and future papers on the subject were the shifts in meaning that occurred over the life of the law's evolution.

Although it is difficult to prove definitively that the popular conception of Moore's law stems from these two articles, substantial circumstantial evidence points to that conclusion. Noyce's articles were the first popular ones on the subject, occurring at a time when interest in microelectronics was growing. They also quickly became the basis for other discussions on the subject. This was evident as early as 1981, when one industry analyst described the *Scientific American* article as "perhaps the most famous of all statements on microelectronics ... Noyce's graphs illustrating the basic 'laws' of microelectronics have been widely copied and quoted."[20] Noyce was also the first person to use the phrase "Moore's law" in a general publication, although Moore credits Carver Mead of the California Institute of Technology with coining the term a year or two before.[21] Although these points do not conclusively demonstrate that it was Noyce who popularized Moore's law, they do at least indicate the significance of his articles.

The accuracy of Moore's law was thus partially the result of historical reinterpretation, but it was contributed to by another factor of the law in the VLSI period—its flexibility. Moore's law had not yet been precisely defined for the VLSI era; Moore's latest paper only indicated that the new doubling time would be closer to two years than to one year.[13] As long as the actual complexity growth rate approximated this curve, it would be adopted as Moore's law for the VLSI era. As a result, the actual doubling time of the law, which approximated a 26-month cycle, was fixed only after the law was first recognized. This phenomenon, combined with historical revisionism, ensured that the law would appear to be successful in the VLSI era.

These factors assisted the transformation of the law from what was essentially a statement about how economics causes technical improvement to a perceived driver of the industry in its own right. The supposed historical accuracy of Moore's law meant that the curve was seen as an indicator of a technological trajectory that seemed at times to be a law of nature. For that reason, the law was often believed to be an important component of industry planning, making it a self-fulfilling prophecy, as firms built their production schedule around the trend of increasing complexity.

The law also underwent a simultaneous change in what it was seen as predicting. The two most important products of the VLSI era, the microprocessor and the memory chip, became the focus for the law. In his 1995 article, Moore explicitly used only DRAMs and

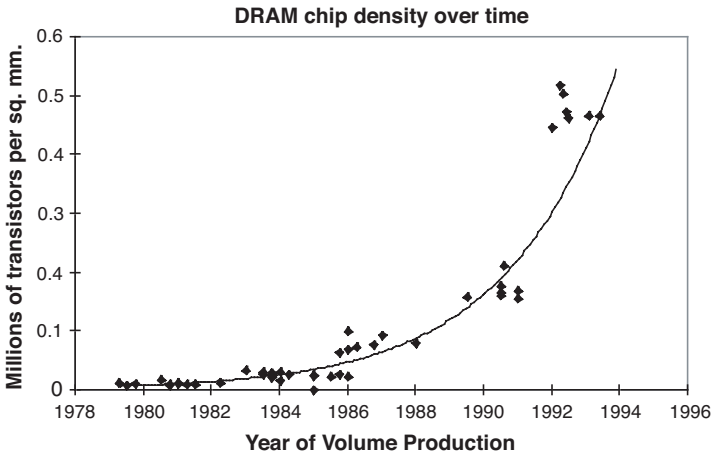

Figure 5. Transistors per mm² vs. year of volume production in DRAM.[22]

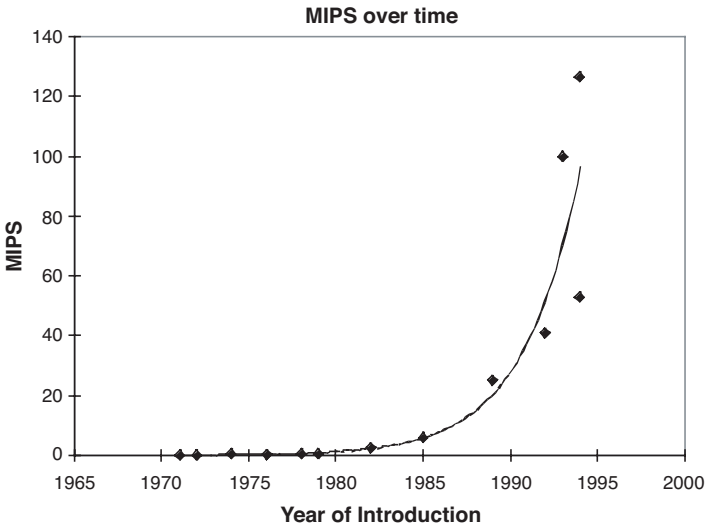

Figure 6. Processor speed in MIPS vs. year of introduction.

processors for his data points. This significantly increased the law's simplicity, especially because DRAMs are based on a stable technology. Technical innovation was therefore not a factor in increasing DRAM complexity, and technical cleverness had already ceased to be an issue since the failure of CCDs. That meant that DRAM progress was solely driven by increases in line density and die size, both of which grew at a fairly constant pace.

The standardization around DRAMs had the added advantage of allowing the pace of the law in the VLSI era to be more easily measured. In earlier periods, differing logic families and chip types made it difficult to compare ICs, a problem that does not affect DRAMs in the VLSI era. As a result, it is possible to directly calculate the accuracy of Moore's law for DRAM (see Figure 5).

Forty-three points, representing the transistor density of DRAM chips in transistors per square millimeter, were graphed. These points included chips from five generations of DRAM, with dates of commercial introduction from the late 1970s until the mid-1990s. The points fell into an exponential curve that approximated a 31 percent improvement per year, or a doubling around every 26 months. The curve fit the data fairly well, with "Moore's law," as characterized by this 26-month cycle, accounting for 76 percent of the variation in the data points. This implies that, at least through much of the VLSI era, the 26-month cycle called Moore's law is an accurate indicator of technical progress.[23]

In addition, with the increasing importance of microprocessors, they, too, were incorporated into the VLSI conception of Moore's law. The emphasis on microprocessors also changed the nature of the Moore's law measurement. For the first time, the law was used to refer to trends other than increasing complexity. This was because microprocessors are not as transistor-dense as DRAM chips, so they could not be measured on the same scale as memory chips. Instead, they are usually measured by either the total number of transistors on a chip, or else by their speed of operation, measured in the number of millions of instructions per second (MIPS) that a processor can carry out.

Typically, between microprocessor generations, a company will release a series of faster, improved microprocessors based on the same design. An example would be Intel's original 486 chip, a 20-MIPS model released in 1989, which was followed by an improved 27-MIPS chip in 1990, and a 41-MIPS chip in 1991. All these had the same transistor count as the first 486 chip, but achieved better performance thanks to a variety of design improvements. For this reason, microprocessors are often graphed by performance, in MIPS, instead of by transistor count (see Figure 6).

The graph in Figure 6, too, indicates a curve that approximates the VLSI Moore's law exponential, with 30 percent improvement per year, and a fit of 97 percent. The graph does seem to indicate two diverging sets of points, however. The lower two points represent improved versions of older chip models, sold as the basis for "economy" computers. When they are removed, the fit is much better.

These three measurements—DRAM complexity, processor transistor count, and processor speed—are all considered part of Moore's law in the VLSI era, which indicates that the law's current form states that chip performance

improves at approximately 30 percent a year. The fact that the law is accurate is not surprising, given that it is only a label assigned to these trends after the fact. What is unusual about Moore's law in the VLSI era, however, is that it has remained consistent and accurate for so long. In previous periods, the law has only operated for a few years at a time before being modified.[24] Currently, the law has been consistent for almost two decades, indicating a change in the semiconductor industry that institutionalized the Moore's law pace of innovation.

Changes in the industry

Although many factors have kept Moore's law as an industry benchmark since the 1975 paper, it is the entry of foreign competition that seems to have played a critical role in maintaining the pace of Moore's law in the early VLSI transition. Before the VLSI period, the semiconductor industry was fragmented and diverse. Many different kinds of chips used many competing logic families. This diversity and the lack of standards it implied was one of the main causes for the continually shifting nature of Moore's law during the first 10 or 15 years of its existence. In the late 1970s and early 1980s, however, the industry standardized both its products and its processes for producing them. This not only allowed the law to remain more consistent in its definition, it also intensified competition internationally. Driving this change was a shift from product to process technologies.

By the beginning of the VLSI era of integration in 1979, the processes used to produce semiconductor products had settled down to a few stable technologies that have remained basically unchanged for long periods.[25] These processes were continually refined and improved, but few major breakthroughs were needed to maintain the technical progress of the semiconductor industry. In many cases, the sustained success of techniques developed in the transition to the VLSI era were a surprise to semiconductor companies, which were accustomed to a continually shifting array of production methods.

A good example of this is the lithography process, which drives increases in line density. It was expected that photolithography, the widely used type of lithography developed in the late 1970s, would cease to be useful by the mid-1980s, causing some firms to develop expensive new ways of etching lines on silicon.[26] Improvements in the standard photolithography process allowed the basic technology to remain applicable long past that expected barrier, however, and the technique is still used today.[27] This pattern of evolutionary, rather than revolutionary, advancement in processes caused semiconductor manufacturers to standardize around a set of production methods that made it easier to enter the industry, paving the way for increased competition. The same sort of standardization occurred in the products produced using these techniques.

The LSI period had seen a proliferation of new logic types, each with its own advantages and capabilities. At least 11 different logic families were in use during that time.[28] One by one, however, each family turned out to have a flaw that ruled out its use beyond a certain complexity level. By the time that the VLSI era was under way, almost all important products used MOS technology, first a variation called PMOS, and later CMOS.[29] This standardization of logic families was, according to Moore, part of the reason why it became easier to apply Moore's law in the late 1970s.[10]

It was more than just types of logic that became standardized across the industry, however. The semiconductor business produced a wide variety of products, and it continues to do so. In the early years of the industry, when it was difficult to produce complex devices, any new product could drive technological development. By the beginning of the VLSI period, this was no longer true. Chip production technology had become so advanced that only two products pushed the technological envelope and moved the complexity capability of the industry forward.

These two chip types were microprocessors and DRAMs, and each became critical to progress in some of the four factors that led to increased complexity: circuit density, die size, technical innovation, and technical cleverness. DRAM is made up of a series of identical, repeating cells. As a result, the key to its manufacture is the ability to make very fine and densely packed circuits. Microprocessors, on the other hand, require a large surface area to contain the large number of specialized sections needed for their complex operation. DRAM, then, has usually been a driver of line dimension while microprocessors drove increases in die size.[30]

These traditional roles have made it easier for the industry to anticipate future developments. Producers of DRAMs could borrow wafer manufacturing techniques from microprocessor firms and concentrate on achieving ever finer line dimensions.[31] This made predictable technological improvements a possibility, because advancements tended to be needed in the production of only a few, well-

known chip types. These standardized ICs have also changed the nature of competition in the semiconductor industry, as a large portion of the industry's profits became linked to a small number of product types.

DRAMs, in particular, became the critical chip type of the VLSI period. Almost every computer-related product requires memory chips, creating a vast market. The generational nature of memory also added to this phenomenon, as the industry typically came out with new DRAMs with four times the storage capacity every two to three years.[32] A large proportion of consumers upgraded with each generation, creating a continual market for the newest chips. Historically, the demand for new DRAMs has proven to be almost bottomless, and producers typically cannot meet the market's requirements. These factors made DRAM sales the largest single portion of the IC business as a whole. All of these factors combined to make DRAM not just an important product, but also the primary driver for the semiconductor industry, both in terms of technology and economics.

The market for DRAMs has the additional interesting feature of being, at least by late in the VLSI era, one of the best examples of a nearly economic perfectly competitive market.[33] Demand is high and the goods are substitutes, making price the primary distinguishing characteristic for memory chips.[34] DRAMs were a relatively new product at the beginning of the VLSI period, however, and there were still comparatively few firms participating in the market. In addition, until the standardization of production processes that occurred around the same time, entry into the business was made difficult by the variety of specialized manufacturing methods that needed to be developed.

Before the beginning of VLSI, semiconductor firms depended on reputation and specialized R&D to sell products to niche markets. As the industry matured toward the production of a few products, such as DRAMs, it became possible for firms that had not previously specialized in semiconductors to enter parts of the now-standardized market using standard processes. Innovation in processes and products still took place, and at a rapid pace, as evidenced by Moore's law. The difference was that, in the VLSI era, progress occurred in an evolutionary, rather than revolutionary, manner, creating an opening for the entry of international competitors. The dawn of the VLSI period was also the beginning of semiconductors as big business because international companies were drawn to the industry.

Japanese entry and Moore's law

The most noticeable and commented-upon change in the industry between the LSI and VLSI era was the sudden success of Japanese IC firms in the late 1970s, especially in the production of memory chips. The semiconductor industry was almost exclusively based in the US until the VLSI period, when the Japanese gained vast market shares in DRAMs from their surprised American competitors. By 1983, the Japanese controlled three-fourths of the world DRAM market.[35]

This sudden Japanese entry into the market was due in large part to the emphasis on process, rather than product, improvements by Japanese firms as they approached the semiconductor market.[36] The difference in emphasis between product and process improvements is important, because it distinguishes between two very different strategies of innovation in the semiconductor market. Product improvements tend to produce market leaders, while process innovations improve manufacturing ability, which can have an effect on price and quality.[37] Firms cannot ignore one area or the other, but the priority they assign each type of innovation, and the order in which they pursue them, can have major impacts on business performance.[38]

Japanese firms' concentration on process technologies, for example, gives companies an advantage in competitive, fairly mature markets. Once a dominant set of products have emerged, an emphasis on manufacturing techniques allows a firm to better compete in a market in which price and quality are the major distinguishing characteristics among goods.[36] Process innovation, then, although not as noticeable as product development, is an important part of any industry strategy.

It was market standardization that really opened up the semiconductor industry to Japan, because Japanese firms were several years behind American firms in late 1970 in both product and process development.[39] The Japanese hoped to overcome this disadvantage by entering a relatively stable sector of the semiconductor market and using that wedge to enter the broader industry. Japanese firms thus planned to make a concerted research effort to target a profitable, standardized product in order to remain competitive. DRAMs provided this market, both because they were relatively simple in design and were profitable in high volumes, giving Japan the opportunity to catch up with American semiconductor manufacturers. This was done through an organized program from 1976 to 1979 conducted by Japan's

Ministry of International Trade and Industry (MITI), which gathered the five major Japanese electronic companies, Fujitsu, Hitachi, Toshiba, NEC, and Mitsubishi into a series of research partnerships collectively called the VLSI program. The Japanese government paid a quarter of a billion dollars toward the project, covering 40 percent of the cost and making the VLSI program the largest subsidy that the Japanese semiconductor industry ever received. The program was successful in increasing the competitiveness of Japanese firms, especially in improving production processes.[40]

For the first time, Japanese companies beat the American firms to market, and by a wide margin—almost two years.

Institutionalizing the law

Even before the conclusion of the VLSI program, the Japanese started to make a strong showing in the DRAM market. Japanese memory chips first achieved a significant market share at the end of the LSI era, with the introduction of the 16K (16-Kbyte) DRAM generation. The 16K chip was first introduced by three American companies in 1976, which were joined by two more US firms and two Japanese firms manufacturing American-designed chips in 1977.[41] By the late 1970s, the Japanese would gain significant market share in the 16K DRAM market, and be poised to move into the next generation. The success of Japanese firms in the 16K generation was mostly due to the differences in business environments, higher quality of Japanese chips, and the captive Japanese domestic market. The combination of these three elements gave Japan a dominance in the 16K generation that would later prove to be critical to the success of the Japanese semiconductor industry, and by doing so help institutionalize the 26-month growth cycle. Still, it was not the 16K but rather the 64K generation that would determine world leadership in DRAM production. The 64K generation was viewed as the critical transitional device to the VLSI generation, and, as such was a key, not just to a large profitable market, but also to the technical leadership of the next generation of microchips.[42]

This generation also demonstrated the way in which international competition enforced the dictates of Moore's law. The nature of the DRAM market was such that the costs of producing chips fell rapidly as the number of chips being manufactured increased.[32,43] The companies that were first to market a generation of chips took advantage of this phenomenon. They could quickly increase production because there was little competition, and thus rapidly lower their costs and prices. Firms whose chips came to market later faced higher costs, because they could not take advantage of economies of scale, and also had to match rapidly falling chip prices. In this way, increasing competition created pressure to innovate at the Moore's law pace, because companies that lagged behind the curve would be lost to firms that were faster to market. This is precisely what happened to the American firms with the 64K generation.

For the first time, Japanese companies beat the American firms to market, and by a wide margin—almost two years. American manufacturers were late in developing 64K chips because they attempted to get a jump on the next generation by applying product and process innovations that would prepare the way for the 256K memories.[44] In the days before a global market, this would have been a wise strategy, but the increased level of competition meant that companies could not innovate faster than the Moore's law curve, just as they could not develop technologies at a slower pace. This was because expensive innovations that delayed the production of a generation would put a company at a disadvantage relative to faster competitors, who could undercut their prices by the time they reached the market. Success required the correct mix of innovations so that a firm could keep up the pace of the industry, but would not have to delay production to develop improved technologies. American firms did not engage in such broad planning in the 64K generation, because almost every domestic effort met unforeseen problems causing delays that, in the days before international competition, would simply have pushed back the release date for the 64K generation, slowing the Moore's law curve.

With Japanese competition, however, the 64K generation was released at the expected time. The Japanese firms, on technical par with the Americans because of the VLSI program, chose a different approach from the Western manufacturers. Instead of using experimental

techniques, the Japanese simply improved and upgraded designs from the 16K generation, which they released even while working on more advanced 64K chips.[42] As a result, the Japanese beat American firms to market, although two US-based companies, Motorola and Texas Instruments, were able to keep pace by quickly redesigning their chips at enormous cost.[45] By the time the American companies had chips on the market, the Japanese controlled the 64K generation.

The entry of foreign firms did more than just increase competition; it also provided a greater variety of innovative approaches to chip design and production. Countries have varying economic structures, with varying wage rates, investment laws, and comparative advantages. These dissimilarities can lead to different rates of innovation and, perhaps more importantly, different choices as to which types of innovation to undertake.[46] The entry of the Japanese into the semiconductor industry thus increased the diversity of innovative approaches, in addition to increasing competitive pressures.

It was this new business atmosphere that forced the exponential increases in complexity that would be attributed to Moore's law. Previously, most semiconductor companies worked on a variety of products, with relatively few firms producing similar, compatible chips. Under this system, if a few companies delayed the release of a particular chip, as most US firms did with the 64K DRAM, it was possible for the entire generation to be postponed, since those few firms capable of developing the chip on time could not meet the market's volume requirements. With the advent of international competition, this situation changed in two ways. First, it was less likely that all the firms in all of the nations would use the same approach to chip production, thus ensuring that technical problems were not as widespread. Second, any delays by one group of manufacturers would be seized upon by firms in other countries as a chance to expand market share. As a result, the increased competition led to a faster, self-regulating pace of development.

International competition, then, was a key factor in maintaining the Moore's law curve during a time when American firms might have otherwise failed to meet that pace. While the exact date at which the industry's pace became attributed to Moore's law is not known, by the mid-1980s the two concepts were closely identified with each other.[47] Since then, other factors have reinforced the Moore's law trajectory, including the Microsoft–Intel alliance, the National Semiconductor Roadmaps, and many other economic drivers that have yet to be analyzed. By the 1980s, however, Moore's law had adapted to the VLSI era, and this time had achieved a form that would remain stable for over two decades.

References and notes

1. See, for example, D. MacKenzie, *Knowing Machines*, MIT Press, 1996, for a discussion of Moore's law as a self-fulfilling prophecy.
2. M.S. Malone, "Chips Triumphant," *Forbes ASAP*, Feb. 1996, p. 70.
3. G.E. Moore, "Progress in Digital Electronics," *Technical Digest of the Int'l Electron Devices Meeting,* IEEE Press, 1975, p. 13.
4. M.S. Malone, "Chips Triumphant," p. 68. Eighteen months is the standard time given in recent publications, both technical and nontechnical. See, for example, T. Lewis, "The Next $10,000_2$ Years: Part II," *Computer*, May 1996), p. 78.
5. P.K. Bondyopadhyay, "Moore's Law Governs the Silicon Revolution," *Proc. IEEE*, vol. 86, no. 1, Jan. 1998, pp. 78-81.
6. R. Schaller, "Moore's Law: Past, Present, and Future," *IEEE Spectrum*, June 1997, pp. 52-59.
7. D. McKenzie, *Knowing Machines,* chapter 3, pp. 49-66.
8. Although the 1965 and 1975 papers are generally considered the critical papers on Moore's law, I have included Moore's 1995 work since it contains the technical material on which he has based all of his recent interviews and presentations.
9. G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics,* vol. 38, no. 8, 1965, pp. 114–117.
10. G. Moore, personal interview with author, 17 Dec. 1996.
11. G.E. Moore, "Progress in Digital Electronics," 1975, p. 11.
12. G.E. Moore, "Progress in Digital Electronics," 1975, p. 11-12.
13. G.E. Moore, "Progress in Digital Electronics," 1975, p. 12.
14. See for example, J.G. Linvill and C.L. Hogan, "Intellectual and Economic Fuel for the Electronics Revolution," *Science*, 28 Mar. 1978, p. 1107, which uses the faulty prediction.
15. G. Moore, personal interview with author, 16 Feb. 2005.
16. R.N. Noyce, "Large Scale Integration: What Is Yet to Come," *Science,* vol. 195, 18 Mar. 1977, pp. 1103-1105. In *Scientific Am.*, this graph is simplified, but it uses the same slope as the *Science* graph; see p. 37.
17. G. Moore, "VLSI: Some Fundamental Challenges," *IEEE Spectrum*, April 1979, p. 32.
18. G. Moore, personal interview conducted by R. Schaller, 13 June 1996.

19. I was unable to find precise transistor counts of the most complex chips from those years, but Moore's article makes it clear that the number of transistors was less than what was expected.
20. T. Forester, ed., *The Microelectronics Revolution,* MIT Press, 1981, p. 29.
21. Noyce wrote: "We have yet to see any significant departure from Moore's law"; R. Noyce, "Microelectronics," *The Microelectronics Revolution*, T. Forester, ed., p. 34.
22. Some data courtesy of M. Iansiti, Harvard Business School, the rest gathered by the author.
23. Note that the 26-month doubling time differs from some published versions of the law. Eighteen months is the number given in much of the popular press, but no authors that I spoke to could trace this number back to its original source. Intel's Web site gives "18 to 24 months," a range that is close to the 26-month cycle.
24. The SSI version in Moore's original paper lasted from 1959 until 1966, when the gap in Moore's graphs begin. The MSI/LSI version as developed in 1975 lasted from approximately 1968 until 1976, ending with the failure of CCDs.
25. US Office of Microelectronics and Instrumentation (OMI), *A Competitive Assessment of the US Semiconductor Manufacturing Equipment Industry,* US Document Printing House, 1985, p. 60.
26. M.S. Malone, *The Microprocessor: A Biography,* Springer-Verlag, 1994, p. 69.
27. G.D. Hutchenson and J.D. Hutchenson, "Technology and Economics in the Semiconductor Industry," *Scientific Am.,* Jan. 1996, p. 55.
28. R. Zaks, *From Chips to Systems,* Sybex, 1981, p. 23.
29. D. Manners and T. Makimoto, *Living with the Chip,* Chapman & Hall, 1995, pp. 102-103.
30. "Processes of the Future," *Solid State Technology,* vol. 38, no. 2, 1995, p. 42.
31. M. Iansiti, *Technology Integration: Making Critical Choices in a Dynamic World,* Harvard Business School Press, 1998, p. VII-10.
32. H. Gruber, "The Learning Curve in the Production of Semiconductor Memory Chips," *Applied Economics,* vol. 24, no. 8, 1992.
33. M. Iansiti, personal interview with author, 7 Dec. 1996.
34. N.S. Dorfman, *Innovation and Market Structure,* Ballinger, 1987, p. 213.
35. H. Queisser, *The Conquest of the Microchip*, Diane Crawford-Burkhardt, trans., Harvard Univ. Press, 1988, p. 112.
36. W.R. Boulton, E.S. Meieran, and R.R. Tummala, "Japan's Product Development Strategy," *Electronic Manufacturing and Packaging in Japan*, Japanese Technology Evaluation Center 1994; www.wtec.org/loyola/ep/toc.htm.
37. H. Gruber, *Learning and Strategic Product Innovation,* North-Holland, 1994, p. 82.
38. Economist Harald Gruber developed a product-process matrix that models the advantages and distinguishing characteristics of these differing approaches in the semiconductor industry. Gruber, *Learning and Strategic Product Innovation*, pp. 82-95.
39. The president of Nippon Telephone and Telegraph stated in 1974 that Japan was behind the US by three years in VLSI technology. M. Anchordoguy, *Computers Inc.* Harvard Univ. Press, 1989, p. 139.
40. See K. Flamm, *Mismanaged Trade?*, The Brookings Institution, 1996, p. 98, for a list of statements on the importance of the VLSI program.
41. Electronics Panel of the National Academy of Engineering (NAE), *The Competitive Status of the US Electronics Industry,* Nat'l Academy Press, November 1989, p. 48.
42. M.G. Borrus, *Competing for Control,* Ballinger, 1988, p. 147.
43. This is often attributed to the so-called learning curve, in which costs fall as experience producing a chip increases. In fact, economies of scale, in which higher levels of production result in greater efficiency, cause this effect.
44. D.I. Okimoto, T. Sugano, and F.B. Weinstein, eds., *Competitive Edge,* Stanford Univ. Press., 1984, p. 48.
45. "A Chance for US Memories," *Business Week*, 15 Mar. 1982, pp. 126-127.
46. G. Dosi, *Technical Change and Industrial Transformation,* St. Martin's Press, 1984, pp. 272-275.
47. See, for example, R.W. Wilson, P.K. Ashton, and T.P. Egan, *Innovation, Competition, and Government Policy in the Semiconductor Industry,* Charles River Associates, 1984, pp. 34-35.

Ethan Mollick is a PhD candidate at the MIT Sloan School of Management in the Technology, Innovation and Entrepreneurship group where he studies how multiple actors interact to shape the innovation process. He holds an MBA from the same institution and a BA from Harvard University.

Readers may contact Ethan Mollick about this article at MIT Sloan School of Management, 50 Memorial Dr., E52, 5th Floor, Cambridge, MA 02142; emollick@MIT.edu.

For further information on this or any other computing topic, please visit our Digital Library at http://computer.org/publications/dlib.