Moore's Law: The History and Economics of an Observation that Changed the World

by G. Dan Hutcheson

This article is an updated and abridged version of "The Economic Implications of Moore's Law," from the Book "High Dielectric Constant Materials," Springer Series in Advanced Microelectronics, Volume 16, Springer-Verlag, New York, 2005; excerpted with permission.

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. 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. 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. Some believe that Moore's Wall is near. Yet there is a rich history to indicate this is not the case. This perspective examines this history and its implications.

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] This abuse of the meaning of Moore's law has led to a great deal of confusion about what it is exactly.

Simply, 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. 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.

These two functions have proven remarkably resilient over the years as may be seen in Fig. 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 metal-oxide-semiconductor (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 2001, the average rate between microprocessor units (MPUs) and dynamic random access memories (DRAMs) ran right at a doubling every two years.

The History of Moore's Law

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 to the industrial age 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]). Nevertheless, it is the exponential growth of scientific answers that led 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.

Before integrated circuits could be invented, transistors had to be made manufacturable. It takes only a glance at early transistors to reveal their inherent manufacturing difficulty (www.chiphistory.org). Making transistors manufacturable was as much a story of technology as it was the triumph of human endeavor and the victory of good over bad management. The Traitorous Eight left Shockley Transistor in 1957 to start Fairchild Semiconductor because they wanted to move away from the pnpn thryristor device that had been developed at Bell Labs, and build silicon transistors using lithography and diffusion techniques. This would ultimately make mass production of transistors efficient. But before this could happen, a further development was needed that would take full advantage of lithography.

Transistors were unreliable and 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, in which the transistor parts were lithographically patterned and diffused into the silicon surface, coated with an oxide passivation layer, and wired together by evaporating aluminum 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.

The next step came when Jack Kilby and Bob Noyce co-invented the integrated circuit (IC). More important to what would become Moore's law was Noyce's recognition that Hoerni's planar process could be used to wire together multiple transistors, capacitors, and resistors, integrating them efficiently onto a single substrate. Kilby's method did not become critically important until the multichip package was needed. Noyce's method quickly became the basis for all modern ICs. The reasons why this method was so important were codified in Moore's 1965 paper.

Moore's law was more than merely a prediction. Moore's paper provided the basis for understanding how and

(continued on next page)

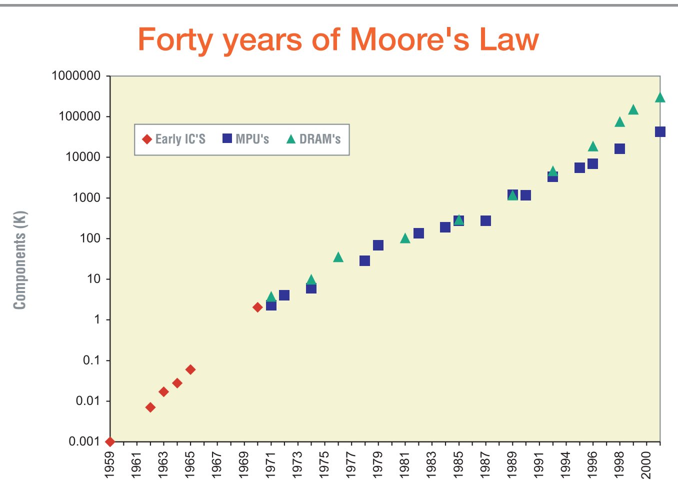

FIG 1. The forces behind the law were still strongly in effect when Gordon Moore retired in 2001, leading him to quip to the author that "Moore's law had outlived Moore's career."

Hutcheson

(continued from previous page)

why ICs would transform the industry. Moore considered user benefits, technology trends, and the economics of manufacturing in his assessment. Thus he described the basic business model for the semiconductor industry, a business model that lasted through the end of the millennium.

In 1975, Moore wrote an update that revised his predictions, as noted earlier. While technically, his prediction of 65,000 components had come true, it was based on a 16-Kbit charge coupled device (CCD) memory, a technology well out of the mainstream. The largest memory in general use at the time, the 16K-bit DRAM, which contained less than half this number of transistors, was not 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. So in the 1975 paper, Moore predicted the periodicity would slow to a doubling every two years.[3] This turned out to be extremely accurate, if seldom quoted with any accuracy. But contrary to what many have thought, the finer points of the accuracy of Moore's law never were that important.

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 to it and invest in it on the basis that the integration scale would always rise in a year or two, making the electronics that was out there obsolete 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 to 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."[7]

The Microeconomics of Moore's Law

So what makes Moore's law work? The law itself describes only two variables in the equation: transistor count and cost. Behind these variables 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.

Reductions in feature sizes have made the largest contributions by far, accounting for roughly half the gains since 1976. Feature sizes are reduced by improvements in lithography methods, which make things smaller. 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, as well as etch tools. 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.

Exposure tools were not always the most costly items in the factory. The camel 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 $100K. In 1990, a state-of-the-art i-line (365 nm) stepping aligner cost just over $1M. By 2002, 193 nm ArF excimer laser scanning aligners cost on the order of $10M, and only two years later were over $20M.

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. Lithography tools have become more productive to offset these increases. It is testimony to the power of this law that these costs have been absorbed, while the cost structure per transistor has actually declined. The increase in the cost of semiconductor factories had been a recurring theme over the years. In was first noted in 1987 that there was a link between Moore's law and wafer fab costs.[8] Between 1977 and 1987, wafer fab costs had increased at a rate of 1.7 times for every doubling of transistors.

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 Ghertz, necessitating a switch to inter-level 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 are being 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 four to five million dollars. Even diffusion furnaces cost a million dollars per tube. As costs have risen, so has risk. So 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. Nevertheless, it is more likely that economic barriers will present themselves before technical roadblocks limit progress.[9]

Moore's law actually governs the real limit to how fast costs can grow. The full version of this article goes through the mathematics to illustrate that if the cost per function must drop by 30% with each node, wafer costs can increase by 40%. So far, the economic barrier to Moore's law has proved just as malleable as the technical ones.

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. This 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. The social and technological progress of the human race can be traced directly to this attribute.

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 were abandoned by 2000 B.C. when the Egyptians developed papyrus based writing materials. It took almost four millennia before silicon staged 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 was

(continued on page 20)

(continued from page 18)

not 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. So 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 took at least another 400 years before retrieval was an issue. In 1876, Melvil Dewey published his classification system that enabled libraries to store and retrieve all the books that were being made until that time. Alan Turing's, "Turing Machine," first described in 1936, was the step that made 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 macroeconomy. 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. Using DRAMs alone, 2000's market adjusted for inflation would be $5328.9T – or just over one hundred times gross world product. Moreover, that does not include the value of all semiconductors! So it is hard to understate the long-term economic impact of the semiconductor industry.

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 21st 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 Gordon 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 do not fully account for what they do not 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 tongue-in-cheek response is to say, "Moore's wall is in Santa Clara, just outside Intel's Robert Noyce building. If you look close, 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 poignantly pointed out, in 1995, that otherwise, "we'll be everything" if things continue at historical growth rates.

However, if you look at history, it dispels this idea. At the beginning of the last millennium rapid advances in agricultural techniques did not slow to meet 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 real limit to the agricultural age was 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 importantly, 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 adjacent to each other. At the same time, products can be designed in different places around the world, the information sent, so products can be produced in low cost areas or, where transportation costs are high, locally. For example, it is a common event for semiconductors being designed in the United States in close cooperation with a customer in Europe, to have the designs sent over the Internet to Texas for the reticles to be made, to California for the test programs, and then to Taiwan to make the wafers, then to Korea for packaging, and finally 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. So, the same productivity effect seen in the transition from the agricultural to the industrial age is also happening as we move into the information age.

It may 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 its growth will be economic not technical, as Moore's law is a statement of powerful economic forces.[9] The reason is that technology barriers first appear as rising costs that go beyond the bounds of economic sense. Highway congestion could be eased by air travel. 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 cost of building roads limits average speed to less than 20 mph in industrial countries, as the ability to build them is far outstripped by demand (which is one reason why the bicycle has become the most 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 $20K. They were the main frames of their day, and only the ultrarich 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 $204.96.

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 push limits to the boundaries of the laws of physics only so far. Costs begin to escalate as this is done because the easy problems are solved, making the next advance more difficult. At some point, little gains can be made by taking the next step, but the cost is astronomical. In the case of autos, 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 economies of technology are a most important underlying factor that makes Moore's law possible and will ultimately result in its demise when gains can no longer be made.

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 a 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. Incidentally, 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, namely, 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.

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 to do this, but they have always been up to the task.

So what advice would Gordon give us? I had the chance to ask him just that during the process of putting together the original version of this chapter. It was on the day he entered retirement.[10] As to the question when Moore's wall would appear, "Who knows? I used to argue that we would never get the gate oxide thickness below 1000 Å and then later 100 Å. Now we're below 10 Å and we've demonstrated 30 nm gate lengths." He gave up on predicting it. The key is not to ask when, but just to keep trying.

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. Moore's law has continued to plug on, delivering benefits to many who will perhaps never appreciate the important contributions of this man and his observation. ■

References

1. G. E. Moore, *SPIE*, **2440**, 0-8194-1799-2/95.
2. G. E. Moore, "The Future of Integrated Electronics, Fairchild Semiconductor," 1965. This was the original internal document from which *Electronics* magazine published "Cramming more components into intergrated circuits," in its April 1965 issue celebrating the 35th anniversary of electronics.
3. G. E. Moore, *Tech. Dig. Int. Electron Devices Meet.* (1975).
4. K. Marx, *Capital*, Chap. 15, Sec. 2, Progress Publishers, Moscow (1978).
5. J. S. Mill, *Principles of Political Economy*, London (1848).
6. G. Ip, *The Wall Street Journal*, p. A1, Dec 28, 2001.
7. Interview: Gordon E. Moore, *Sci. Amer.* (Sept 1997).
8. G. D. Hutcheson, *The VLSI Capital Equipment Outlook*, VLSI Research Inc. (1987).
9. G. D. Hutcheson and J. D. Hutcheson, *Sci. Amer.*, p. 54 (Jan 1996).
10. G. E. Moore, Personal conversation, May 24, 2001.

About the Author

Dan Hutcheson is CEO of VLSI Research Inc. During his career, Dan has authored numerous publications, developed many industry models, and researched most aspects of the semiconductor industry. He is probably best known for being the first to forecast an industry recession and for having developed the industry's first cost-of-ownership model in the early1980s. He also built the first factory cost-optimization model in 1984. Similar models are now the mainstay of capital decision-making.

(continued on page 24)