REBOOTING COMPUTING

It's Time to Redefine Moore's Law Again

Erik P. DeBenedictis, Sandia National Laboratories

The familiar story of Moore's law is actually inaccurate. This article corrects the story, leading to different projections for the future. Moore's law is a fluid idea whose definition changes over time. It thus doesn't have the ability to "end," as is popularly reported, but merely takes different forms as the semiconductor and computer industries evolve.

A larger-than-life idea called Moore's law powered an information revolution and worldwide economic growth. Moore's law is certainly a statement about technological advancement, but it's also a symbol of computer technology's positive impact on innovation and the economy.

For many years, Semiconductor Industries Associations (SIAs) across the globe produced an International Technology Roadmap for Semiconductors (ITRS; www.itrs2.net/itrs-reports.html) to project the progress of integrated-circuit (IC) performance and its consistency with Moore's law. This was seen as a way to manage industry-wide IC development. However, the US SIA recently stopped supporting the road map. In addition, US president Barack Obama proposed government-funded research for what he called the "post–Moore's law era,"[1] suggesting that the US government sees the principle as outdated. So what happened?

This article will track the changing meaning of Moore's law over the last 50 years and then use its collective meaning to assess current technology. We will see that industry is moving computer technology forward at the expected rate but that Moore's law needs another redefinition to recapture past bullish attitudes.

PROJECTIONS VERSUS POSSIBILITIES

Moore's law is based on a 1965 *Electronics* magazine article[2] by Gordon Moore that makes a research contribution to IC manufacturability but also includes statements that could be considered either technical projections or tantalizing possibilities. A projection becomes a part of Moore's law for a person who sees it that way. As part of Moore's law, it can be assessed as being correct or incorrect, with Moore's law ending for that person when a projection becomes incorrect. However, if a statement is seen as a tantalizing possibility, it contributes to the article's overall sense of optimism even if it turns out to

0018-9162/17/$33.00 © 2017 IEEE

TABLE 1. Dennard scaling.[a]

Device or circuit parameter	Scaling factor
Device dimension t_{ox}, L, W	$1/\kappa$
Delay time/circuit	$1/\kappa$
Power density VI/A	1

(a) R.H. Dennard et al., "Design of Ion-Implanted MOSFETS with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, 1974, pp. 256–268.

be not completely accurate. A person can alter the meaning of Moore's law by categorizing the statements one way or the other.

The most widely understood projection from Moore's 1965 paper is that the number of transistors on an IC would double every few years. The article said "shrinking dimensions on an integrated structure makes it possible to operate the structure at higher speed for the same power per unit area."[2] This critical property enabled improvements in IC-based products. Rising device counts and higher speeds or clock rates allowed more features in successive product generations while retaining the product's package and power source.

In 1974, Robert Dennard showed how to realize Moore's law with metal-oxide semiconductor field-effect transistors (MOSFETs).[3] Table 1 shows three of Dennard's scaling expressions, based on the scaling factor κ.

The first row shows a transistor gate oxide's thickness (t_{ox}), length (L), and width (W) scaling down from 1 to $1/\kappa$. Because devices are 2D, the number of transistors per unit area will be κ^2 in the next generation.

In the second row, delay time is the inverse of clock rate. Thus, as delay time scales down from 1 to $1/\kappa$, the clock rate would increase by a factor of κ, which is the square root of the transistor count's κ^2 increase. In his 1965 article, Moore said speed or clock rate would increase but he did not identify the precise square root relationship, albeit perhaps because he was not writing specifically about MOSFETs.

In the third row, Dennard calculated that the decrease in power consumed per unit area (volts multiplied by current [VI] divided by unit area [A]) caused by using smaller transistors and lower voltage on chips would precisely offset the consumption increase caused by the utilization of more transistors and higher speeds, as indicated by the scaling factor of 1. Moore projected this in his paper using words instead of equations.

Dennard scaling no longer holds for reasons beyond the level of detail that Moore wrote about, including one evident in the table's first row. To maintain the properties of Moore's law, the thickness of a MOSFET's gate oxide would have to decrease at the same rate as the length and width, which would lead to impractically thin gate oxides. This challenge was addressed by using high-K dielectric material.

THE CHANGING DEFINITION OF MOORE'S LAW

Does Moore's 1965 article include an early version of Dennard's 1974 scaling rule, or was it just a tantalizing possibility? History shows Moore's law is a fluid concept, as opposed to an unchanging one as found in a physical or political law. In fact, Moore's article never used the phrase "Moore's law" or even the word "law." The article did include a cartoon of computers being sold as consumer items in the future, a projection that came to pass for microprocessor-based products. Of course, microprocessors weren't invented until 1971[4] and it wasn't until 1979 that Carver Mead even coined the phrase "Moore's law."[5]

The ITRS specifically projected the progress of MOSFET-based microprocessors and memory in relation to Moore's law. However, higher speeds for them were no longer technically feasible after Dennard scaling ended in 2003. Even though Moore's statement about this is no longer correct, it can be considered as just a tantalizing possibility. And Moore's law continues to be an accurate projection of the number of devices per chip.

The broad scope of Moore's article and the imprecise language of wide-audience magazine articles gave the community the flexibility to redefine and thereby "save" Moore's law in 2003. Although Moore's law could be redefined, there wasn't enough flexibility in Dennard's mathematical expressions to save his rule.

MOORE'S LAW AND CHANGING RESEARCH DIRECTIONS

The central part of Figure 1 shows the changing scope of research based on Moore's law. The scope mapped by the ITRS is shown in red and is limited to CMOS microprocessors and DRAM. Moore's original article had a wider scope that includes the adjacent purple areas.

The stagnation of microprocessor clock rates after Dennard scaling ended created a crisis for the semiconductor industry out of fear that product families could no longer grow in features and speed. Thus, government leaders adopted a post–Moore's law era research agenda,

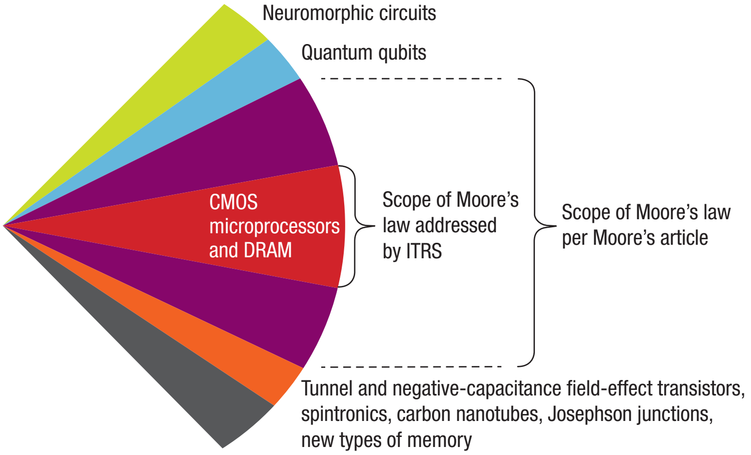
Neuromorphic circuits
Quantum qubits
CMOS microprocessors and DRAM
Scope of Moore's law addressed by ITRS
Scope of Moore's law per Moore's article
Tunnel and negative-capacitance field-effect transistors, spintronics, carbon nanotubes, Josephson junctions, new types of memory

Figure 1. Research directions' changing scope. Using the scope of Moore's original paper as a baseline, the International Technology Roadmap for Semiconductors (ITRS) projected the future more narrowly, for just CMOS microprocessors. In contrast, the post–Moore's law era is being defined for entirely new devices. The changing definition of Moore's law creates a "research gap" shown in purple.

including the technologies outside the red and purple areas in Figure 1. Some of these devices will be incorporated into logic gates—such as AND, OR, and NOT—and thus preserve current architectures. Other devices will need new architectures.

The purple areas in Figure 1 comprise topics that have shifted from being part of Moore's law to part of the post–Moore's law era. Topics in these areas are currently in a "research gap" because they aren't associated with either. To the extent that the gap is due to a misunderstanding of how Moore's law changes meaning over time and not technical limitations, it's worth examining.

Research directions: architecture

One approach represented in the research gap is the GPU. GPUs on computers and smartphones evolved from devices embedded in PC graphics cards to their current status as general-purpose coprocessors.

The creation of multicore processors with very large numbers of cores represents another research-gap architecture. Companies began moving to multicore processors in about 2003 but needed time to figure out how to program them. The subsequent programming success, along with higher-density chips, has enabled companies to build systems with dozens and even hundreds of cores, such as the Intel Xeon Phi family.

IBM's TrueNorth is a conventional CMOS chip in this category of diverse architectures, but in place of cores, it uses neuron- and synapse-like structures.

Research directions: programming

Parallel programming for multicore chips is still difficult, but a new approach gives better results. Dual- and quad-core processors almost always run existing code, because there is not enough performance gain to justify rewriting it. The new systems, on the other hand, have enough cores or other computational elements to justify such a rewrite. The emerging software approach for doing this is for a few specialist programmers to create highly optimized subroutines for components using the new architectures, while a greater number of programmers create the larger body of application code for standard processors.

Hewlett-Packard's R. Stanley Williams and I suggested in this department's October 2016 edition[6] that these approaches are examples of an emerging type of hybrid computer with multiple architectures, each designed to solve a different problem. This makes them general-purpose systems in the aggregate.

Research directions: 3D chips

Shifting from 2D to 3D chip manufacturing has led to new architectures that continue to improve performance. High-bandwidth memory and the hybrid memory cube are two approaches for assembling multiple DRAM chips and a CMOS logic chip into a single module. DRAM chips have already been stacked on GPUs and multicore processors, demonstrating increased memory bandwidth with decreased latency and power consumption.

University researchers have embraced a long-term vision for 3D manufacturing in which a chip is built with many layers of logic, memory, and interconnects without requiring an additional assembly step. Figure 2 is an example of this.[7]

The technical community has adequately addressed the engineering challenges of staying on the optimistic path that Moore called for in his famous article. However, we must redefine the phrase "Moore's law" to use its name recognition to encourage continued progress.

The strategy for properly redefining Moore's law is for organizations to cite, in their charter or in the background material for a product or process, Moore's original article as the starting point for a redefinition, perhaps as follows:

> *Industry is again on the path of Moore's law.*[2] *New architectures have been developed that do not require higher clock rates,*

and 3D manufacturing enables the number of transistors per chip to continue rising without requiring unrealistically small feature sizes. Due to the evolving requirements of applications, most of the new devices layered in the third dimension will be memory. Power per chip will thus stay constant because memory is naturally low power.

However, maintaining Moore's law based on this definition will require tightly integrated research involving semiconductors, architectures, and software.

Creating a "post–Moore's law era" may have been premature, but it has helped bring neuromorphic and quantum computing into the limelight. These useful approaches are so different from existing approaches that they are likely to coexist with today's "programmed" computing.

The ITRS joined IEEE Standards Association's Industry Connections program under the new name International Roadmap for Devices and Systems, which has goals[8] similar to the ideas in this article. IRDS leadership supports this article. ■

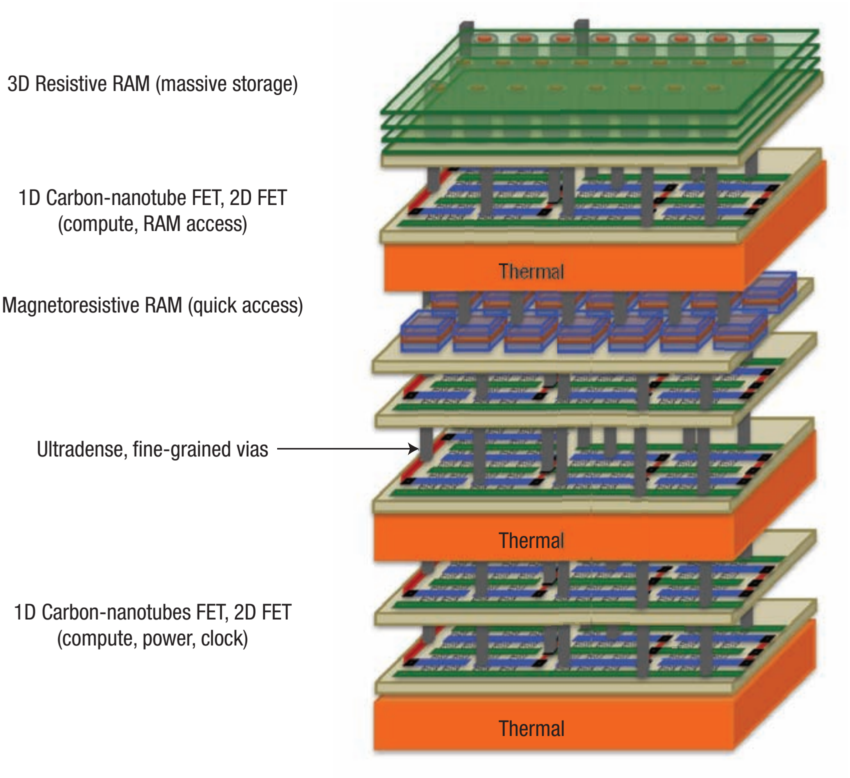
3D Resistive RAM (massive storage)
1D Carbon-nanotube FET, 2D FET (compute, RAM access)
Thermal
Magnetoresistive RAM (quick access)
Ultradense, fine-grained vias
Thermal
1D Carbon-nanotubes FET, 2D FET (compute, power, clock)
Thermal

Figure 2. 3D chip. University researchers propose building chips in three dimensions, with multiple layers of logic and memory, including field-effect transistors (FETs), connected by ultradense vias.

ACKNOWLEDGMENTS

Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the US Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

REFERENCES

1. B. Obama, "Creating a National Strategic Computing Initiative," Executive Order 13702, *Federal Register*, vol. 80, no. 148, 3 Aug. 2015, p. 46177.
2. G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 1965, pp. 114–117.
3. R.H. Dennard et al., "Design of Ion-Implanted MOSFETS with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, 1974, pp. 256–268.
4. K. Shirriff, "The Surprising Story of the First Microprocessors," *IEEE Spectrum*, vol. 53, no. 9, 2016, pp. 48–54.
5. E. Mollick, "Establishing Moore's Law," *IEEE Annals of the History of Computing*, vol. 28, no. 3, 2006, pp. 62–75.
6. E.P. DeBenedictis and R.S. Williams, "Help Wanted: A Modern-Day Turing," *Computer*, vol. 49, no. 10, 2016, pp. 76–79.
7. M.M.S. Aly et al., "Energy-Efficient Abundant-Data Computing: The N3XT 1,000 x," *Computer*, vol. 48, no. 12, 2015, pp. 24–33.
8. T.M. Conte and P.A. Gargini, "On the Foundation of the New Computing Industry beyond 2020," *Preliminary IEEE RC-ITRS Report*, Sept. 2015; rebootingcomputing.ieee.org/images/files/pdf/prelim-ieee-rc-itrs.pdf.

ERIK P. DEBENEDICTIS is a technical staff member at Sandia National Laboratories' Center for Computing Research. Contact him at epdeben@sandia.gov.