DEVICE TECHNOLOGY

Moore's law: The journey ahead

High-performance electronics will focus on increasing the rate of computation

By **Mark S. Lundstrom** *and* **Muhammad A. Alam**

The transistor was invented 75 years ago, and the integrated circuit (IC) soon thereafter. The progress in making transistors smaller also led to them becoming cheaper, which was famously noted as Moore's law (*1*). Today's sophisticated processor chips contain more than 100 billion transistors, but the pace of downsizing ("scaling") has slowed and it is no longer the only or even main design goal for improving performance in particular applications. How can Moore's law continue on a path forward? New approaches include three-dimensional (3D) integration that will focus on increasing the rate of information processing, rather than on increasing the density of transistors on a chip.

Although Moore's law predicted a rate for the decrease in cost per transistor, it is popularly viewed in terms of transistor size, which for two-dimensional (2D) chip arrays translates into an areal size or "footprint." During the last 75 years, as the footprint has decreased from micrometer to nanometer scales, issues with implementing new fabrication technologies have raised concerns several times about the "end of Moore's Law." Twenty years ago, a pessimistic outlook prevailed regarding the development of several difficult technologies for scaling to continue. In this context, one of the authors (M.S.L.) predicted that instead of slowing down, the scaling of metal-oxide-semiconductor field-effect transistors (MOSFETs) below the so-called 65-nm node, which was state-of-the-art in 2003, would continue unabated for at least a decade before the scaling limit was reached (*2*).

Scaling indeed continued from about 100 million transistors per chip in 2003 to as many as 100 billion transistors per chip today. One approach was to improve the on-off current ratio to allow practical operation and suppress leakage current to reduce wasted power. In 2003, strained silicon was introduced as channel material, and it increased the on-current by increasing the velocity of electrons (*3*), and in 2004, gate insulators with a high dielectric constant decreased the off-state gate-leakage current. In 2011, the FinFET, a nonplanar transistor structure that increases the electrostatic control of the energy barrier by the gate electrode (and thereby improves the on-off current ratio), was introduced in commercial ICs. Gate all-around transistors that further improve the electrostatic control of the gate are now in development (*4*). The size of transistors that can be fabricated is limited by patterning and etching. Patterning is done by a process known as photolithography, in which a photoreactive polymer creates a mask on the chip for etching steps. The minimum size of the pattern is determined by the wavelength of the light used. The recent emergence of extreme ultraviolet lithography (EUV) made it possible for Moore's law to continue beyond the 7-nm node (*5*).

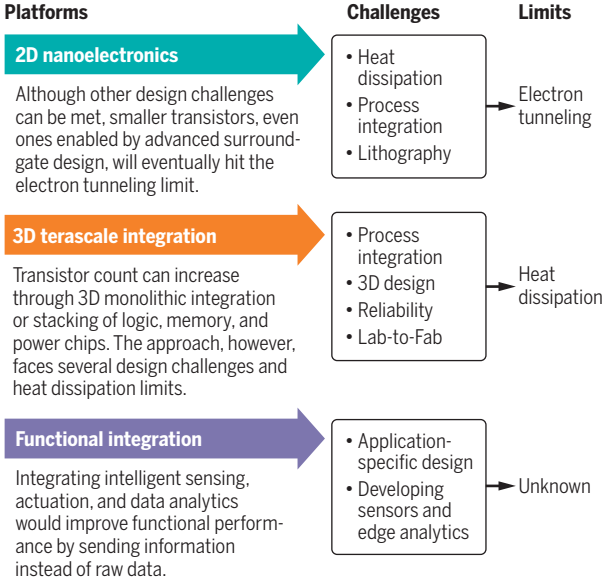
The number of transistors on a chip is still increasing, but the rate of scaling has slowed because smaller transistors do not function very well. Specifically, the length of the channel (the region between the source and drain electrode where the gate acts as a switch) is now ~10 nm. At shorter channel lengths, excessive quantum-mechanical tunneling degrades transistor action. Key performance metrics, such as on-current (which should be high for high-speed operation), off-current (which should be low to minimize standby power), and power supply voltage (which should be low to minimize the power consumed), all degrade simultaneously. Silicon MOSFETS are now about as small as they can get, and the 2D chips are about as large as they can be made, so new ways to advance performance must be found.

Performance is being enhanced by moving from general-purpose, "commodity chips" to ones that accelerate specific functions. For example, hardware acceleration offloads specific tasks to specialized chips such as graphics processing units or an application-specific IC. Companies such as Apple now design their own chips to meet their specific requirements, as will all of the major automobile manufacturers. Computing is the limiting factor for machine learning, and companies such as Google now design their own artificial intelligence (AI) accelerator chips. Custom chip designs can increase performance by orders of magnitude, but just as the cost of chip manufacturing facilities ("fabs") has multiplied (from ~$1 billion in 2000 to ~ $20 billion for a leading-edge fab), so has the cost of leading-edge design. The design of a leading-edge chip can cost $0.5 billion and require a team of ~1000 engineers. Lowering the cost of leading-edge, custom-chip design (possibly by using machine learning techniques) will be a key challenge for the next era of electronics.

Continued progress will also require advances in the underlying technology. Despite the sharp increase in the number of transistors on chips (both by decreasing

Three platforms forward

Two-dimensional (2D) nanoelectronics, three-dimensional (3D) terascale integration, and functional integration can all extend Moore's law, but all face substantial challenges and fundamental limits.



GRAPHIC: K. HOLOSKI/*SCIENCE*

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1971, USA.
Email: lundstro@purdue.edu

their size and increasing the 2D chip area), until recently one aspect of the design had remained largely unchanged. Individual chips are packaged and combined with other chips and other components (such as inductors) laterally on printed circuit boards. Sending signals on- and off-chip increases delays and power consumption.

An emerging design theme is to exploit the third (vertical) dimension to enable terascale integration (TSI), with trillions of transistors integrated into monolithic or stacked chips and terabits per second per millimeter communication speed for electrical or optical interconnections ("per millimeter" refers to the communication link distance between the chips). For example, a 3D NAND flash memory (which is based on NAND logic gates and retains its states with the power off) can have nearly 200 layers of devices and half a trillion memory transistors (*6*). Emerging logic transistors with new channel materials (such as transition metal dichalcogenides and indium oxide) that can be processed at low temperatures and embedded within the interconnect stacks offer further opportunities.

The third dimension also opens up the possibility of vertical heterogeneous integration of logic, memory, and power transistors. With "through-Si vias," metal wires that connect vertically from the chip, already-processed chips can be stacked to put them in close physical proximity to minimize signal delays and reduce power consumption (*7*). Vertically stacked logic and memory chips also enable new computing paradigms, such as "compute-in-memory." Monolithic 3D ICs would consist of layers of active devices, such as 2D logic transistors, magnetoresistive and resistive random-access memories, and ferroelectric FETs, along with the metal lines that interconnect them (*8*).

Recent packaging innovations, such as silicon-interposer and multi-die silicon bridges, inserted between the 3D chips and the substrate, create denser lateral interconnection and faster communication among the chips. Advanced packaging brings together logic, memory, power management, communications, and photonics through side-by-side integration. The proximity of integration now rivals that in stacked and monolithic 3D ICs (*8, 9*).

Monolithic 3D integration will require that the growth or deposition steps do not affect the already-processed layers. For example, the transistors embedded within the interconnect stack must be deposited at a low-enough temperature not to disturb the dopant profiles of the Si transistors underneath. The needed materials are often incompatible unless special processes are developed. Stacking already-processed 2D chips to achieve 3D systems has its own set of material and processing challenges, such as maintaining interconnect alignment over distances of ~1 to 5 μm. Heterogeneous integration of components such as Si high- and low-voltage logic and memory transistors, and compound semiconductor-based power and high-frequency transistors, presents another set of complex integration challenges.

Transistors generate heat when they operate, and removing the heat is a serious issue in electronics today (*7, 10*). Indeed, thermal cross-talk among logic, memory, power transistors, and inductors in a heterogeneous IC creates unprecedented design challenges. New ways to remove heat, perhaps mimicking the thermoregulation of organisms, and thermally aware design will be critical when trillions of transistors are placed in close proximity.

The reliability of electronic systems must be guaranteed for a minimum time, typically 10 years, but decades for some applications. Ensuring a failure rate between 1 and 10 parts per million for ICs with 100 billion transistors each requires predicting the reliability of quintillion (~10^{18}) transistors. In practice, reliability is determined through short-term accelerated testing of no more than a few thousand transistors. Thus, the reliability physics of the wear-out and catastrophic failure modes of these new systems need to be understood with unprecedented precision. When so many devices are interconnected and placed in close proximity, new phenomena will emerge, and these must be managed or exploited.

Future terascale systems will be fundamentally different from today's gigascale systems in that understanding the building blocks of a system do not inform how these blocks interact and lead to new phenomena (*11*). Chip design is already complex and expensive, but algorithms or tools to place devices for 3D design and routing the interconnections among them are not yet available. These design tools must model the complexity of the process and package integration, thermal cross-talk among 3D ICs, and operation-specific variability and reliability of the packaged system.

When new materials and processing techniques are developed in research, they must be translated to large-scale manufacturing. Translating advances achieved with research-grade equipment to large-scale manufacturing with different and more sophisticated state-of-the-art manufacturing equipment presents a serious "lab to fab" challenge. The research community will need access to advanced processing facilities, and short "conceive-conduct-analyze" experimental loops that maximize learning are needed.

Thermal issues will define the limits of 3D terascale integration, just as tunneling limits have stymied 2D scaling. This requirement need not herald an end of Moore's law. The goal of computing is not operations per second but information per second. In that regard, biology offers a guide. Human senses process information locally before forwarding it to the brain. Empowering the sensors at the edge that interfaces the analog world, supported by local memory and data processing (edge analytics), could prevent the data deluge from overwhelming the computer.

Electronics is at an inflection point. For 75 years, it has been possible to make transistors smaller, but that will not be the driving force for progress in the decades ahead. If Moore's law is understood to refer to the increasing number of transistors per integrated system (not necessarily per chip), then the end of Moore's law is not in sight (see the figure). The increasing number of transistors will not come by making them smaller, but by stacking them vertically or combining them laterally in sophisticated packages, and eventually in monolithic 3D chips and adding functionality.

Shifting from nanoelectronics (focused on reducing the transistor dimension) to terascale electronics (driven by increasing transistor count and related functionality) defines the paradigm shift and core research challenges of the future. It will require fundamental advances in materials, devices, processing, and the design and manufacture of the most complex systems humans have ever built. Someday the electrical tunneling and thermal bottleneck will define the limits of 3D integration. Until then, Moore's law will likely continue as researchers address the challenges of these extraordinarily complex electronic systems. ■

REFERENCES AND NOTES

1. G. E. Moore, Electronics (Basel) no. 8 (19 April 1965) (1965).
2. M. Lundstrom, *Science* **299**, 210 (2003).
3. K. Mistry *et al*., in *IEEE International Electron Devices Meeting*, pp. 247250 (2007).
4. D. Jang *et al*., *IEEE Trans. Electron Dev*. **64**, 2707 (2017).
5. WIRED, "The $150 million machine keeping Moore's law alive," 30 August 2021; https://www.wired.com/story/asml-extreme-ultraviolet-lithography-chips-moores-law/.
6. A. Goda, *Electronics* **10**, 3156 (2021).
7. R. W. Keyes, *Proc. IEEE* **60**, 225 (1972).
8. S. Iyer, *IEEE Trans. Compon. Packaging Manuf. Technol*. **6**, 973 (2016).
9. C. H. Douglas *et al*., in 2021 *IEEE International Electron Devices Meeting*, pp. 3–7.
10. M. A. Alam *et al*., *IEEE Trans. Electron Dev*. **66**, 4556 (2019).
11. P. W. Anderson, *Science* **177**, 393 (1972).

10.1126/science.ade2191