# CMOS Scaling Trends and Beyond

**Scaling transistors and following Moore's law have served the industry well for more than 50 years in providing integrated circuits that are denser, cheaper, higher performance, and lower power. This article describes trends in CMOS scaling over the past decade and discusses some of the new device options and technology directions being explored to continue scaling into the future.**

**Mark T. Bohr, Ian A. Young**
*Intel*

Gordon Moore famously predicted in his 1965 paper that the number of components per chip would continue to increase by a factor of two every year.[1] The goals of following Moore's law are to decrease the cost per component and reduce the power consumed per component. In 1975, Moore updated his earlier prediction by forecasting that components per chip would increase by a factor of two every two years, and that this would come from the combination of scaling component size and increasing chip area.[2] Back in 1965, the industry was producing chips using a minimum feature size of approximately 50 μm totaling about 50 components. Today's leading chips use a minimum feature size of approximately 10 nm and incorporate several billion transistors.

Robert Dennard and colleagues described in 1974 a scaling methodology for metal-oxide-semiconductor field-effect transistors (MOSFETs) that would deliver consistent improvements in transistor area, performance, and power reduction.[3] The methodology called for the scaling of transistor gate length, gate width, gate oxide thickness, and supply voltage all by the same scale factor, and increasing channel doping by the inverse of the same scale factor (see Figure 1). The result would be transistors with smaller area, higher drive current (higher performance), and lower parasitic capacitance (lower active power). This method for scaling MOSFET transistors is generally referred to as "classic" or "traditional" scaling and was very successfully used by the industry up until the 130-nm generation in the early 2000s.

For the past 20 years, we have been developing new generations of process technologies on a two-year cadence, and each generation scaled the minimum feature size by approximately 0.7 times to deliver an area scaling improvement of about 0.5 times (see Figure 2). Thus, we have been doubling transistor density every two years. But recent technology generations (such as 14 nm and 10 nm) have taken longer to develop than the normal two-year cadence, owing to increased process complexity and an increased number of photomasking steps. Nonetheless, Intel's 14-nm and 10-nm technologies have provided better-than-normal transistor density improvements that keep us on pace with increasing transistor density at a rate of doubling about every two years.


Published by the IEEE Computer Society
0272-1732/17/$33.00 © 2017 IEEE

## Transistor Innovations

As mentioned earlier, traditional MOSFET scaling worked well up until the 130-nm generation in the early 2000s. By that generation, the SiO2 gate oxide thickness had scaled to about 1.2 nm, and electron tunneling through such a thin dielectric was becoming a significant portion of total transistor leakage current. We had reached the limit for scaling transistors using traditional methods, and we needed to start introducing innovations in transistor materials and structure to continue scaling.

One of the first significant innovations was the introduction of strained silicon transistors on Intel's 90-nm technology in 2003.[4] This innovation used tensile stain in *n*-channel MOS (NMOS) transistor channels to increase electron mobility and compressive strain in *p*-channel MOS (PMOS) channels to increase hole mobility (see Figure 3). Tensile strain was induced by adding a high-stress film above the NMOS transistor. Compressive strain was induced by replacing the PMOS source-drain regions with epitaxial SiGe depositions. The resultant increases in electron and hole mobility provided increased transistor drive currents without having to further scale the SiO2 gate oxide thickness. This strained silicon technique has been adopted by all major semiconductor companies and continues to be used on the latest 10-nm technologies.

The need to improve the transistor gate dielectric to continue scaling could not be avoided, and Intel's 45-nm technology in 2007 first introduced high-κ metal gate transistors.[5] The traditional SiO2 gate oxide was replaced by a hafnium-based high-κ dielectric. The high-κ dielectric both reduced gate oxide leakage current and improved transistor drive current. The traditional doped-polysilicon gate electrode was replaced by metal electrodes with separate materials for NMOS and PMOS to provide optimal transistor threshold voltages. The combination of high-κ dielectric and metal gate electrodes (see Figure 4) was a revolutionary process change that provided significant improvements in transistor performance while also reducing transistor leakage current. High-κ metal gate transistors are now universally used on advanced logic technologies.

The next major transistor innovation was the introduction of FinFET (tri-gate) transistors on Intel's 22-nm technology in 2011.[6] Traditional planar MOSFETs had been able to scale transistor gate length down to about 32 nm to deliver good performance and density while also maintaining low off-state leakage. But scaling the gate length below 32 nm was problematic without sacrificing either performance or leakage. A solution was to convert from a planar transistor structure to a 3D FinFET structure in which the gate electrode had better electrostatic control of the transistor channel formed in a tall narrow silicon fin (see Figure 5). This improved electrostatic control provided scaled transistors with steeper sub-threshold slope (see Figure 6a). Steeper sub-threshold slope either provided transistors with lower off-state leakage or allowed threshold voltage to be reduced, which enabled improved performance at low operating voltage (see Figure 6b). Operating integrated circuits at a lower voltage is highly desired in order to reduce active power consumption. All advanced logic technologies now use FinFET transistors

| Device or circuit parameter | Scaling factor |
|---|---|
| Device dimension *tox, L, W* | 1/κ |
| Doping concentration *Na* | κ |
| Voltage *V* | 1/κ |
| Current *I* | 1/κ |
| Capacitance *εA/t* | 1/κ |
| Delay time/circuit *VC/I* | 1/κ |
| Power dissipation/circuit *VI* | 1/κ2 |
| Power density *VI/A* | 1 |

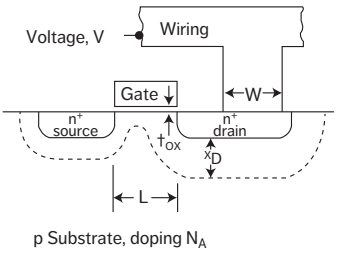


**Figure 1.** Traditional MOSFET scaling as described by Robert Dennard.

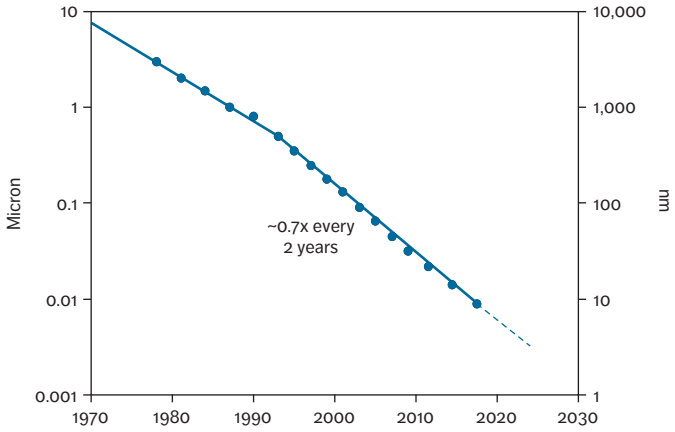


**Figure 2.** Minimum feature size scaling trend for Intel logic technologies.

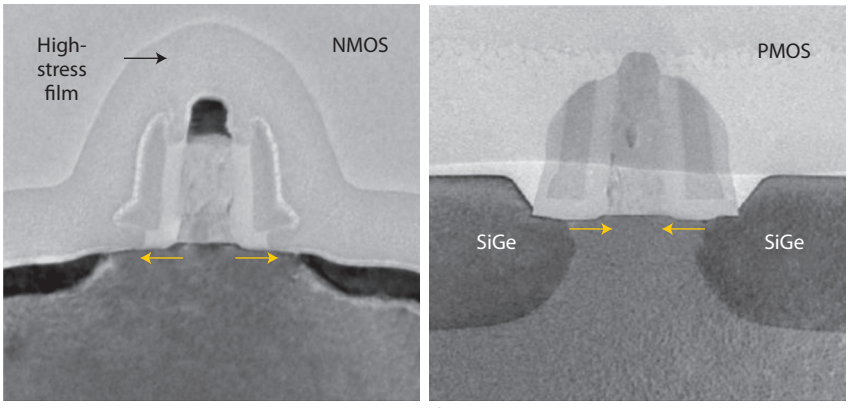


**Figure 3.** Channel strain techniques used on 90-nm generation transistors. (a) NMOS transistor using SiN cap layer; tensile channel strain. (b) PMOS transistor using SiGe source-drain; compressive channel strain.

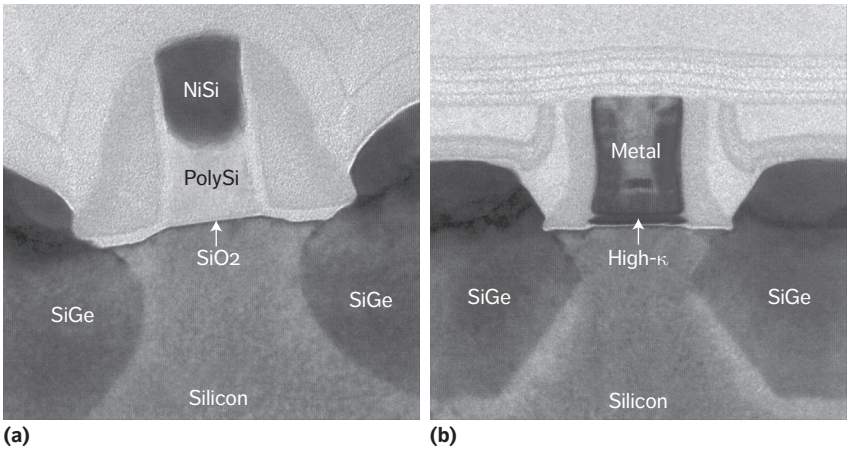


**Figure 4.** Comparison of transistor structures. (a) 65-nm generation transistor using $SiO_2$ dielectric; polysilicon gate electrode. (b) 45-nm generation transistor using hafnium-based dielectric; metal gate electrode.

for their good density and superior low-voltage performance compared to planar transistors. As Figure 7 shows, when traditional MOSFET scaling ran out of steam in the early 2000s, innovations such as strained silicon, high-κ metal gate, and FinFETs were needed, and we must now continually invent new transistor materials and structures to continue scaling.

## Recent Logic Technologies

Intel's 14-nm logic technology started volume production early in 2014. This was Intel's second-generation FinFET technology, and it used advanced features such as 70-nm transistor gate pitch, 42-nm fin pitch, 52-nm interconnect pitch, double patterning techniques, and a 6-T SRAM bitcell area of 0.0588 $\mu m^2$.[7] This technology took longer to develop and get ready for volume manufacturing due to the increased process complexity and mask count: about 2.5 years instead of the normal 2-year cadence. But this technology also provided better-than-normal area scaling. Instead of the 0.5 times area scaling that new technology generations normally provide, Intel's 14-nm technology provided about 0.37 times logic area scaling compared to the previous 22-nm technology (see Figure 8).

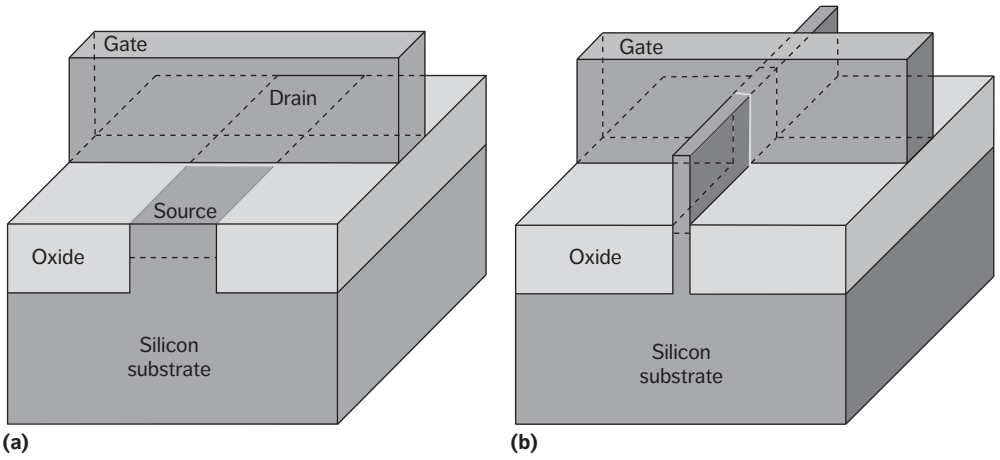


**Figure 5.** Comparison of transistor structures. (a) Planar transistor. (b) FinFET transistor.

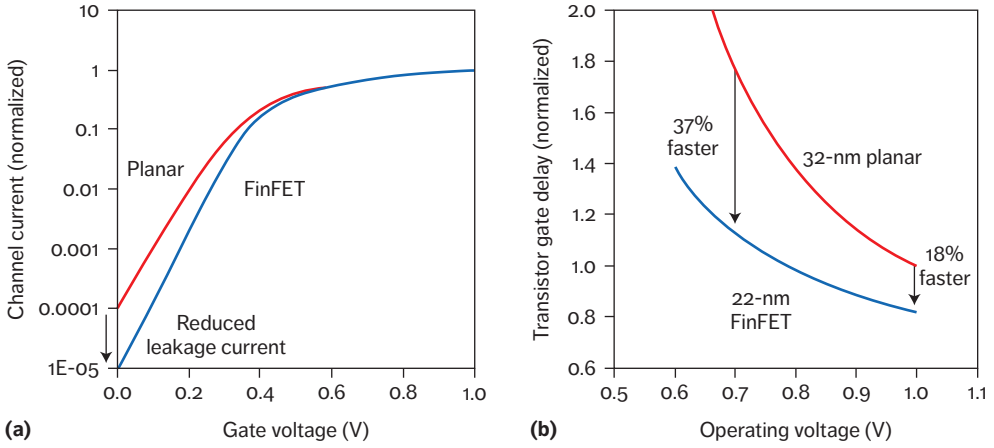


**Figure 6.** Comparison of planar versus FinFET transistor electrical characteristics. (a) Channel current versus gate voltage. (b) Transistor gate delay versus operating voltage.

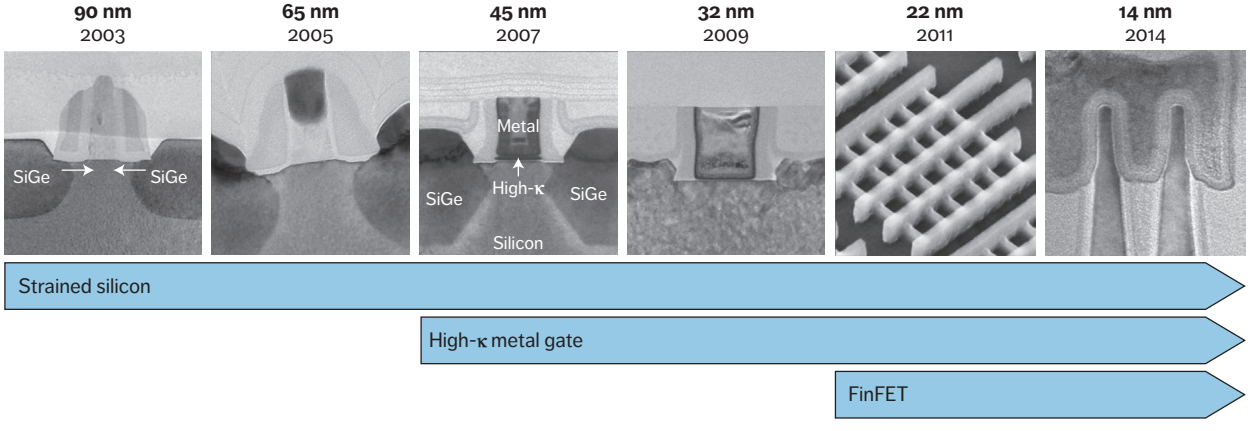


**Figure 7.** Six generations of Intel transistor innovations used to continue scaling.

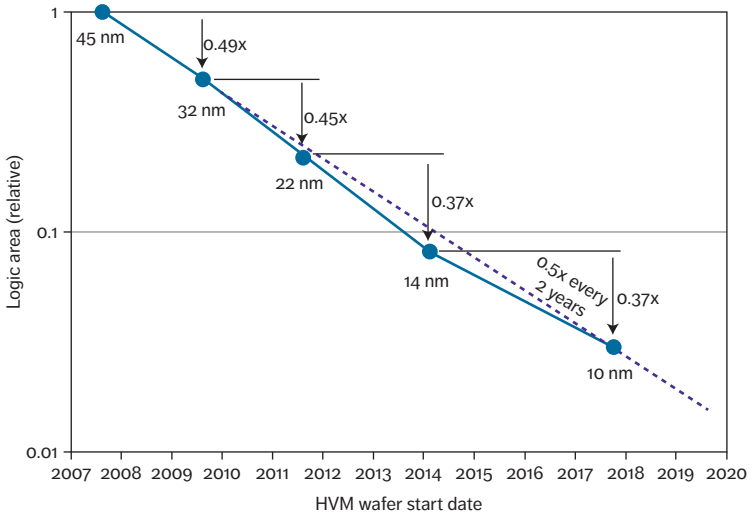


**Figure 8.** Intel's trend for scaling logic circuit area over the past five generations.

Intel's newest 10-nm logic technology is scheduled to start product shipments before the end of 2017. This 10-nm technology introduces some advanced process features such as 54-nm transistor gate pitch, 34-nm fin pitch, 36-nm interconnect pitch, quad patterning techniques, and a 6-T SRAM bitcell area of 0.0312 $\mu m^2$. This technology also introduces some important density-improvement techniques: single dummy gates adjacent to logic cells and the ability to make transistor gate connections directly over active gates. Again, this technology took more than two years to develop and get ready for volume manufacturing due to increased process complexity and mask count, but it also delivers better-than-normal area scaling. The innovative features on this technology deliver about 0.37 times logic area scaling compared to the previous 14-nm generation. As Figure 8 shows, Intel's 14-nm and 10-nm generations each took more than two years to develop, but they also took bigger steps in terms of scaling logic area. As a result, Intel logic technologies continue to deliver improved area scaling at the rate of about 0.5 times every two years.

It's apparent that after more than 50 years we're continuing to scale transistor area, but are we delivering the other promises of Moore's law and Dennard's scaling methodology: lower cost per transistor, higher performance, and lower active power? Figure 9a shows how Intel logic technologies have been scaling transistor area, and Figure 9b shows the trend of increasing wafer cost due to increased process complexity. Figure 9c shows how the cost per transistor continues to come down due to better-than-normal area scaling. Figure 10 shows Intel's trends for improving transistor performance (Figure 10a) and reducing dynamic capacitance to lower active power (Figure 10b). Figure 10c shows how performance improvement divided by active power consumption (performance per watt) continues to improve with each generation. Different products on a given technology can choose to tune the transistor or design to deliver better performance or lower power, depending on what the application values most. Figure 10 also shows the strategy of developing performance-enhanced versions of each generation (for example, 10+ and 10++) to deliver improved performance per watt and extend the life of these technologies.

## Future Device Options

MOSFET transistor researchers are exploring device structure and channel material changes to enable further generations of MOSFET scaling. The MOSFET implemented with stacks of multiple horizontal nanowires (see Figure 11b) is one option that, due to its superior electrostatics, could enable further gate-length scaling beyond what the FinFET (see Figure 11a) can achieve. MOSFETs with III-V semiconductor channel materials are a promising option for realizing a higher-mobility channel than silicon (see Figure 12). This higher mobility can be used either to provide higher drive current and higher performance or to allow the MOSFET to be operated at lower voltage for lower active power.[8]

Lowering the supply voltage of CMOS logic below about 0.5 V leads to a dilemma between logic having high performance and high static leakage current versus logic with lower performance and low leakage current. This is due to the choice of MOSFET threshold voltage and its electron "thermal tail" determined sub-threshold gate voltage swing of 60 mV/decade. One alternative transistor option that operates differently than a MOSFET (and as such could be classified as a *beyond-CMOS device*) is the Tunneling Field Effect Transistor (TFET).[9] The TFET can achieve subthreshold swing smaller than 60 mV/decade (that

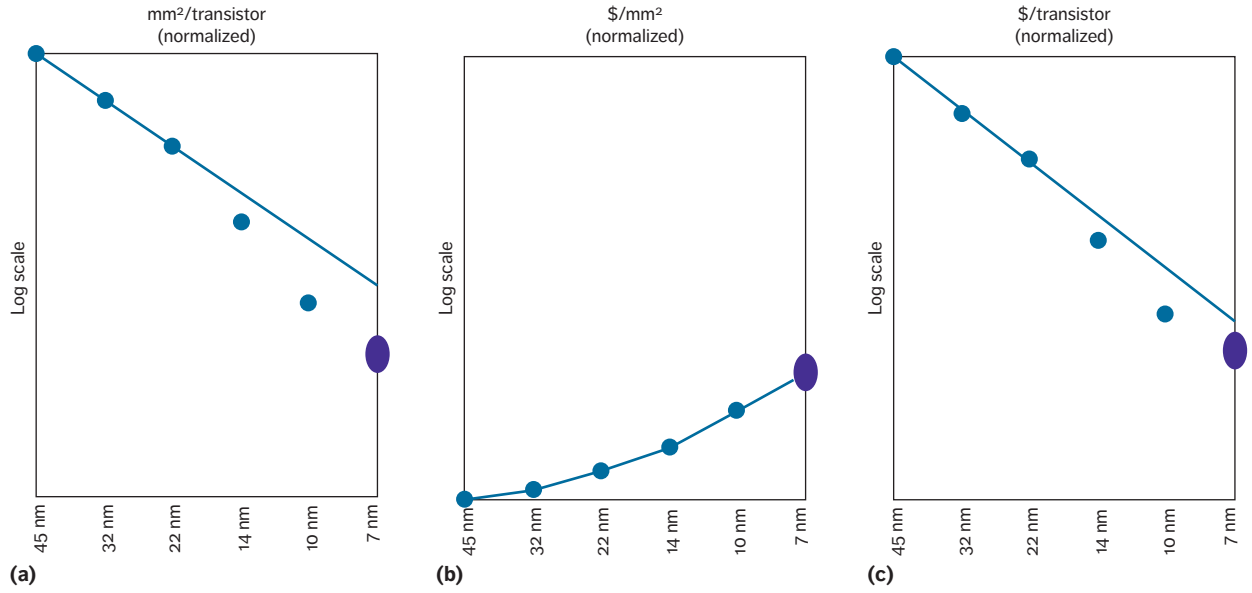


**Figure 9.** Trends for improving logic transistor area and cost per transistor. (a) Area per transistor. (b) Wafer cost. (c) Cost per transistor.

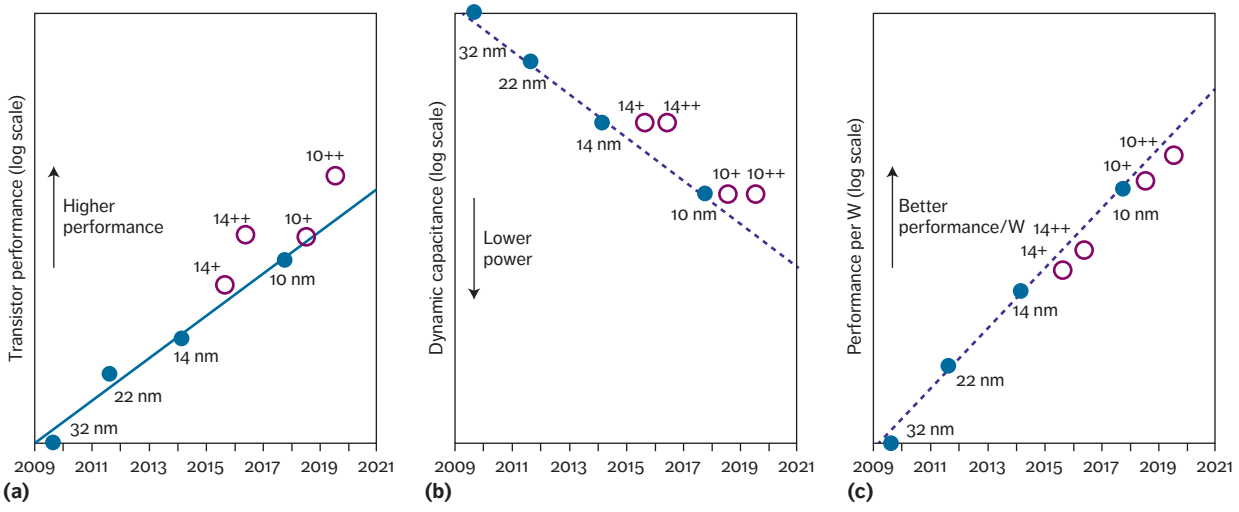


**Figure 10.** Trends for improving transistor performance and reducing active power. (a) Transistor performance. (b) Dynamic capacitance. (c) Performance per watt.

is, steeper current turn-on) and can therefore operate at a lower power supply voltage than a MOSFET. Figure 13 shows drain current versus gate voltage simulation results for nanowire TFETs implemented with different III-V semiconductor materials.

While the success of information technology progress in the past 50 years was based on Moore's law[1,2] scaling and mostly one underlying technology—CMOS transistors—present-day research efforts are exploring logic technologies going beyond CMOS,[10] with an objective to complement CMOS rather than to replace it. The goal of Beyond-CMOS research is to identify and enable an integrated circuit technology that will be more energy efficient than CMOS. If this happens, it will support the continuation of Moore's law.

Beyond-CMOS research efforts have been underway for 10 years, being funded in the US

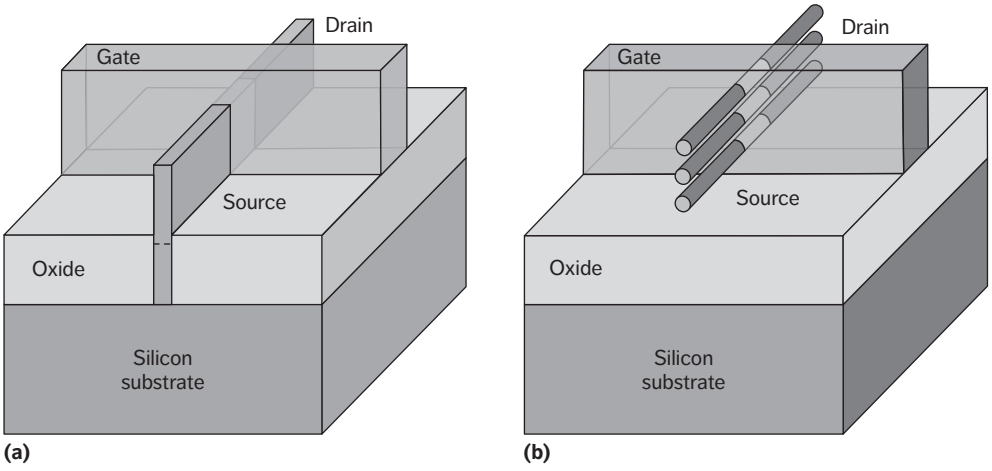


**Figure 11.** Comparison of transistor structures. (a) FinFET transistor. (b) Nanowire transistor.

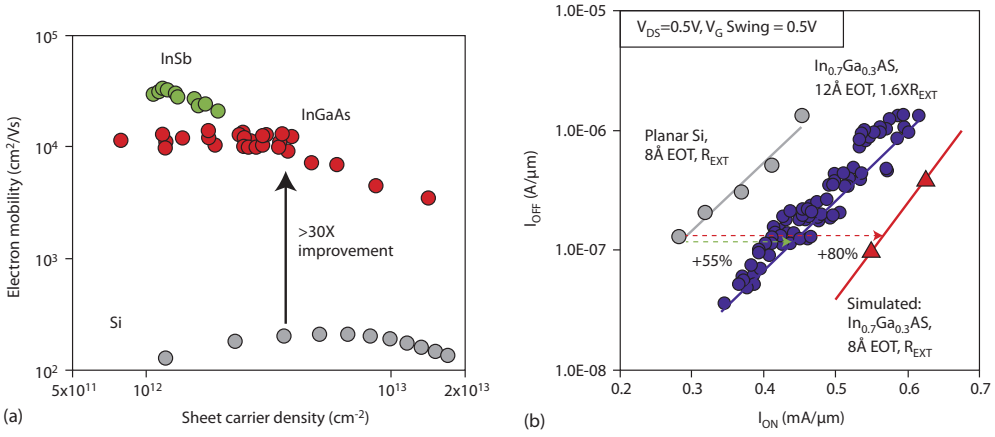


**Figure 12.** Comparison of III-V and silicon transistor electrical characteristics. (a) Electron mobility versus carrier density. (b) Off-current versus on-current.

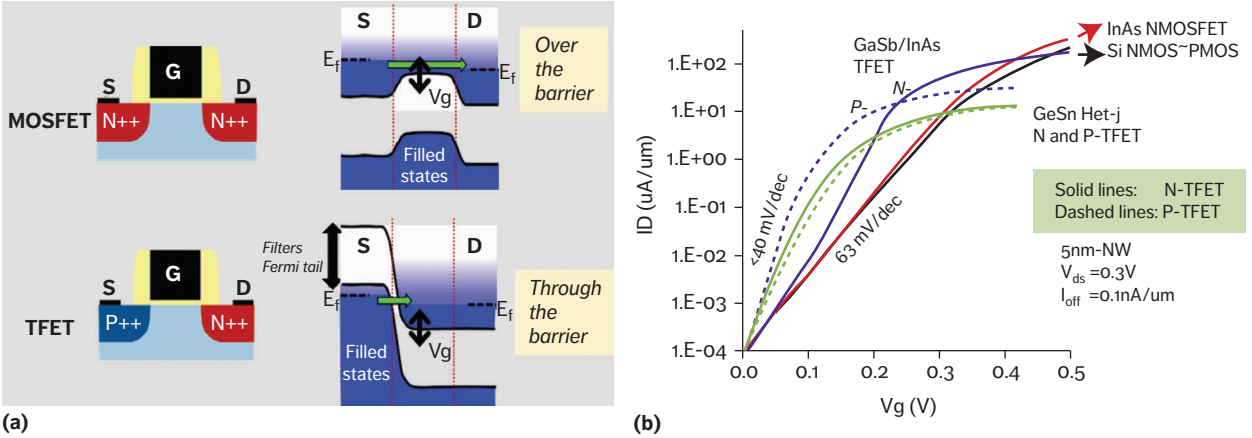


**Figure 13.** Comparison of Tunneling FET and MOSFET transistors. (a) Transistor structures and channel current modulation techniques. (b) Drive current versus gate voltage electrical characteristics.

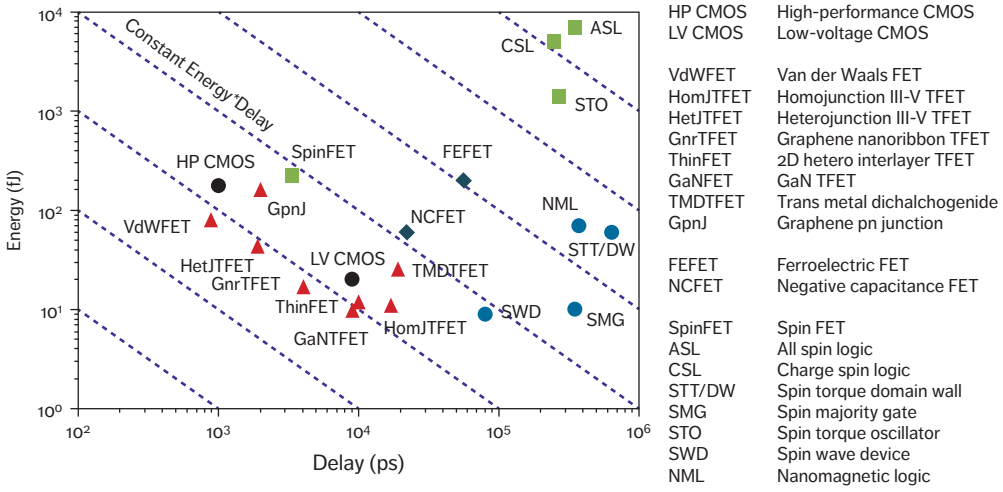


**Figure 14.** Simulated switching energy and delay for a 32-bit arithmetic logic unit circuit for CMOS and for various beyond-CMOS device options.

in large part via the Semiconductor Research Corporation (SRC).[11] The expectation of this industry–university research consortium 10 years ago was that this field would produce a computing technology that is better than CMOS for the majority of its applications. Reality showed that, among many impressive proposals and demonstrations, none of them beat CMOS. However, they do possess many valuable features, such as low-power operation and non-volatility. Thus, the current vision is that beyond-CMOS circuits will replace CMOS in some critically important computation or information processing applications. They would be monolithically integrated with CMOS on the same chip or packaged together in a multichip module.

Another expectation was that beyond-CMOS circuits would not require any MOSFETs as part of their operation and could maybe even eliminate any charge currents in the quest for energy efficiency. This did not come true: a thorough circuit analysis reveals that a MOSFET transistor is needed to supply power and for clocking and control of the logic circuit operation. However, this does not preclude pursuing the key direction of beyond-CMOS research, which is to discover and invent computation that can operate at significantly lower supply voltages than CMOS to enable dramatic improvements in energy efficiency.

To this end, beyond-CMOS benchmarking[12,13] (see Figure 14) was helpful in evaluating the potential of various materials and devices to implement computing technologies. It enabled the setting of expectations for power and performance and revealed some pathways for improvement. Experimental demonstrations have not yet achieved the theoretical modeling projections put forward in the benchmarking. One reason is that each computing technology requires solving numerous fabrication challenges.[14]

The various materials implementations of the TFET (see Figure 13) have shown that they have improved energy-delay product (and therefore power and performance) over the future CMOS technology node that they are benchmarked against (see Figure 14): the *International Technology Roadmap for Semiconductors* prediction in 2011 of the 2018 CMOS node. With a potential three-times improvement in energy-delay product over CMOS, this is starting to be an interesting device option, and it does not require a drastic change in circuit design for logic while it offers some additional circuit functionality.

The spintronic devices in Figure 14 operate with a wide range of switching energy and at slower switching speed compared to CMOS. The spintronic devices that match the best CMOS switching energy use magnetoelectric materials to do the switching of nanomagnets. Although they are slower than CMOS, they have the added benefit of being non-volatile. Non-volatility in the logic device has the potential to provide energy efficiency benefits by taking advantage of it in the computing microarchitecture.

A historic similarity for beyond-CMOS research is fitting—the disruption of bipolar transistors for computing logic by CMOS.[15] The latter had the advantage of lower power, but it was slower than bipolar and was much more difficult to manufacture. We believe that the same drive toward lower-power computing should compel technologists to solve implementation problems for beyond-CMOS computing. One should understand that a 100-times improvement in the energy-delay product (which is equivalent to more than four generations of historic Dennard-era CMOS scaling) will justify its integration for computer and information processing systems. As research into beyond-CMOS continues, it is going to be critical that researchers focus on the leading options and eliminate the less attractive ones. To do this will require all levels of benchmarking analysis covering materials, devices, circuits, and computing architectures.[16]

Transistor scaling, and in particular MOSFET scaling, has served our industry well for more than 50 years by providing new generations of integrated circuit technology that simultaneously provided improved density, higher performance, reduced power consumption, and lower cost per transistor. At times, transistor scaling was provided by the use of simple evolutionary techniques, but at other times more revolutionary technology changes were required, such as switching from bipolar to MOSFET transistors, and more recently by implementing high-κ metal gate and FinFET transistors. Furthermore, 14-nm and now 10-nm generations have continued to deliver the promises of Moore's law for improved density, performance, power, and cost.

Scaling of the MOSFET transistor will continue for future CMOS generations as far as researchers can see by exploiting the options in device structure and channel materials. Beyond-CMOS research into quantum nanoelectronics or nanomagnetics is aimed at inventing and developing another integrated circuit technology that offers improved power and performance. This will happen at the appropriate time when it can be integrated onto CMOS in a manufacturing process that offers lower cost per function and improved power and performance.

## References


1. G. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 1965, pp. 114–117.
2. G. Moore, "Progress in Digital Integrated Electronics," *IEEE Int'l Electron Devices Meeting Technical Digest*, 1975, pp. 11–13.
3. R. 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. T. Ghani et al., "A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors," *IEEE Int'l Electron Devices Meeting Technical Digest*, 2003, pp. 978–980.
5. K. Mistry et al., "A 45nm Logic Technology with High-k + Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," *IEEE Int'l Electron Devices Meeting Technical Digest*, 2007, pp. 247–250.
6. A.C. Auth et al., "A 22nm High Performance and Low-Power CMOS Technology Featuring Fully-Depleted Tri-gate Transistors, Self-Aligned Contacts and High Density MIM Capacitors," *Proc. Symp. VLSI Technology*, 2012, pp. 131–132.
7. S. Natarajan et al., "A 14nm Logic Technology Featuring 2nd Generation FinFET Transistors, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588um2 SRAM Cell Size," *IEEE Int'l Electron Devices Meeting Technical Digest*, 2014, pp. 71–74.
8. R. Kim, U.E. Avci, and I.A. Young, "Comprehensive Performance Benchmarking of III-V and Si nMOSFETs (Gate Length = 13 nm) Considering Supply Voltage and OFF-Current," *IEEE Trans. Electron Devices*, vol. 62, no. 3, 2015, pp. 713–721.
9. U.E. Avci et al., "Energy Efficiency Comparison of Nanowire Heterojunction TFET and Si MOSFET at Lg = 13 nm, Including P-TFET and Variation Considerations," *IEEE Int'l Electron Devices Meeting Technical Digest*, 2013, pp. 33–36.
10. W.M. Holt, "1.1 Moore's Law: A Path Going Forward," *Proc. IEEE Int'l Solid-State Circuits Conf.*, 2016, pp. 8–13.
11. J.J. Welser et al., "The Quest for the Next Information Processing Technology,"

*J. Nanoparticle Research*, vol. 10, 2008, pp. 1–10.
12. D.E. Nikonov and I.A. Young, "Overview of Beyond-CMOS Devices and a Uniform Methodology for their Benchmarking," *Proc. IEEE*, vol. 101, no. 12, 2013, pp. 2498–2533.
13. D.E. Nikonov and I.A. Young, "Benchmarking of Beyond-CMOS Exploratory Devices for Logic Integrated Circuits," *IEEE J. Exploratory Solid-State Computational Devices and Circuits*, vol. 1, 2015, pp. 3–11.
14. I.P. Radu et al., "Spintronic Majority Gates," *IEEE Int'l Electron Devices Meeting Technical Digest*, 2015, p. 32.5.1–32.5.4.
15. S. Borkar, "Electronics Beyond Nano-Scale CMOS," *Proc. 43rd ACM/IEEE Design Automation Conf.*, 2006, pp. 807–808.
16. I.A. Young and D.E. Nikonov, "Principals and Trends in Quantum Nano-Electronics and Nano-Magnetics for Beyond CMOS Computing," to be published in *Proc. European Solid-State Device Research Conf.*, 2017.

**Mark T. Bohr** is an Intel Senior Fellow in the Logic Technology Development group at Intel. His research interests include scaling logic transistors, interconnects, and memory cells. Bohr received a master's degree in electrical engineering from the University of Illinois. Contact him at mark.bohr@intel.com.

**Ian A. Young** is an Intel Senior Fellow in the Components Research group at Intel. His research interests include novel embedded memory and quantum nanoelectronic and nanomagnetic devices for energy-efficient integrated circuits beyond-CMOS. Young received a PhD in electrical engineering from the University of California at Berkeley. Contact him at ian.young@intel.com.