

BEYOND MOORE'S LAW: THE INTERCONNECT ERA

Reversing early limitations on Moore's law, interconnectors have replaced transistors as the main determinants of chip performance. This "tyranny of interconnectors" will only escalate in the future, and thus the nanoelectronics that follow silicon must be interconnect-centric.

In the 20th century, mainstream electronics evolved from vacuum tubes and discrete copper wires to individual transistors and batch-fabricated, printed wiring boards to its current form, which integrates more than a billion transistors and copper interconnections in a single silicon chip. Until the past decade, designers benignly neglected the electrical performance of wires or metal interconnections, beyond a cursory accounting for their parasitic capacitance. They effectively addressed this problem simply by increasing transistor channel width to provide larger drive currents, and thus enable transistor-level circuit performance.

Unfortunately, this simple fix is no longer adequate for two salient reasons: both interconnect latency and energy dissipation now tend to dominate key metrics of transistor performance. For example, for current, state-of-the-art 100-nm technology, the latency of a 1-mm-long interconnect benchmark is approximately six times larger than that of a corresponding transistor. Moreover, the energy dissipation associated with a benchmark interconnect's bi-

nary transition is approximately five times larger than that of a corresponding transistor.

This "tyranny of interconnects" escalates rapidly for future generations of silicon technology. Consequently, in the near- and medium-term future, exponential increases in transistors per chip—as Moore's law eloquently projects—will necessarily emphasize advances in interconnect technology. These advances will be extremely diverse and will include new interconnect materials and processes, optimal reverse scaling, microarchitectures that shorten interconnects, 3D structures and I/O enhancements, interchip optical interconnects, and more powerful computer-aided design tools for chip layout and interconnect routing. Here, I discuss these improvements, which are likely to evolve for decades. I also discuss the logic behind my prediction that any revolutionary technology that supplants silicon must be interconnect-centric. First, however, let's review Moore's law and what its ultimate limits mean for the future of interconnect technology.

Moore's Law

In his now-famous appearance at the IEEE International Electron Devices Meeting in 1975, Gordon Moore officially recognized that the number of transistors per chip N had been doubling annually for more than a decade.¹ Eight

years later, at the 1983 IEEE IEDM, I used a compact model of Moore's law to predict that, in the year 2000, the number of transistors per chip would be approximately one billion.² I based this prediction on an extrapolation of the expression $N = F^{-2} \times D^2 \times PE$, where N is the number of transistors per chip, F is the minimum feature size, D is the square root of chip area, and PE is the transistor packing efficiency, measured in number of transistors per minimum feature area, F^2 . The 1999 International Technology Roadmap for Semiconductors (ITRS) verified my prediction by citing the 1-Gbit DRAM as the generation at introduction in 1999.³ In 1995—again based on the expression $N = F^{-2} \times D^2 \times PE$ —I projected a one-trillion-transistor chip before 2020.⁴ The 2001 ITRS projects the 64-Gbit DRAM using 32-22 nanometer technology as the generation at introduction in the 2013–2016 time period.⁵

For several decades, a key question for semiconductor technologists has been: What are the ultimate limits on Moore's law or the number of transistors per chip? In pursuing an unambiguous response to this question, I first postulated a five-level hierarchy of limits in 1983,² and refined it in 1995⁴ and again in 2000.⁶ The five levels of this hierarchy are

- Fundamental
- Material
- Device
- Circuit
- System

Fundamental limits are independent of the properties of any particular material, device structure, circuit configuration, or system architecture. Consequently, we can reasonably assume that a fundamental limit is essentially absolute and cannot be surpassed. Researchers have rigorously derived the minimum energy that must be transferred in a binary logic circuit's switching transition—the canonical computing operation—to have the value $E_s = (ln2)kT$, where k is Boltzmann's constant and T is absolute temperature.⁶ The minimum energy that must be transferred in a single interconnect's binary transition has precisely the same value. The remarkable fact that we derive the same results from two distinctly different physical models, one based on transistor physics and the other on interconnect communication theory, serves to confirm its validity.⁶ Clearly, this limit on binary switching energy is fundamental since both Boltzmann's constant and absolute temperature

are independent of the properties of any particular material, device, or circuit.

Given that we know the fundamental limit on binary switching energy, how can we use it to derive the limit on a transistor-centric Moore's law? To find an answer, we must assume the most elementary model for a metal oxide semiconductor field effect transistor (MOSFET) that treats its input or gate electrode as a simple parallel plate capacitor. In addition, we must engage the binary switching energy limit ($E_s = (ln2)kT$), and take the associated charge transfer at its absolute minimum value (a single electron). We can then calculate the resulting transistor channel length to be approximately 10 nanometers for an equivalent gate dielectric thickness of 1.0 nanometers.⁶ Using both a rigorous numerical model⁷ and a compact physical model⁸ of a symmetric double gate MOSFET, we essentially confirm the 10-nanometer limit on transistor channel length.

Based on projections derived from both the hierarchy of limits on gigascale integration⁴ and the 2001 ITRS,⁵ a 10-nm minimum feature size could support a terascale chip—that is, a chip with a trillion transistors—between 2015 and 2020. The ultimate achievement of this estimate of Moore's law's limit is subject to several critical caveats, as discussed in the ITRS.⁵ For example, we must

- Demonstrate novel microfluidic methods to satisfy a terascale chip's cooling requirements
- Ensure that the chip's economic viability justifies the required multibillion-dollar financial investments

First, however, we must invent the necessary interconnect technology to effectively complement 10-nm transistors.

Interconnects: Near and Medium Term

The ITRS predicts that in 2012 industry will produce 35-nm generation technology.⁵ Expected features of the technology indicate that a 1-mm-long interconnect's latency will be 100 times larger, and its binary switching energy will be 30 times larger, than a corresponding transistor.⁹ To address this tyranny of interconnects, we have several options.

On-Chip Technologies

Reverse scaling, which calls for increasing interconnect cross-sectional dimensions to markedly reduce resistance per unit length—and therefore

latency—offers significant opportunity for improvement. For example, we can reduce the area of a 16-million-gate chip with eight interconnect levels by 33 percent by optimally scaling rather than simply doubling the pitch of each successive orthogonal pair of levels. The key to optimal reverse scaling is to derive a chip’s complete stochastic wiring distribution (that is, the number of interconnects versus interconnect length).⁹ Using repeaters can further enhance performance or reduce chip size.

To maximize the total bandwidth and minimize latency between two macrocells in a system-on-a-chip, we can calculate the optimal interconnect width that maximizes the quotient of bandwidth per unit of interconnect width and latency.¹⁰ We can also define an allowable design range for an integrated architecture’s interconnect width and height for a heterogeneous system-on-a-chip’s global signal, power/ground, and clock distribution networks. The key design constraints are the three wiring networks’ total area, the clock distribution network’s required bandwidth, and the allowable peak crosstalk of the signal and clock lines.^{9,11}

For several decades, it’s been common practice to use chip wiring patterns that are in the *x*-direction only on a single level and in the *y*-direction only on the adjacent level. Thus, engineers typically implement multilevel interconnection networks with “orthogonal pairs” of wiring levels. This restriction accommodates the capabilities of computer-aided design tools that facilitate chip layout and interconnect routing. Recently, vendors have introduced more powerful tools that enable both diagonal and orthogonal wiring directions.¹² Rigorous modeling reveals that pervasive diagonal routing can reduce the wiring-limited chip area by up to (an astounding) 70 percent, for a typical wiring efficiency factor of 0.4. However, for a software-impaired router wiring efficiency of 0.2, pervasive diagonal wiring is less effective than conventional orthogonal wiring. (The wiring efficiency factor is the ratio of the total chip area to the wiring area that a given interconnect level actually consumes.)

Three-dimensional structures composed of multiple transistor and wiring strata offer further opportunities to improve interconnect performance. Increasing the number of strata from one to four, for example, reduces the length of a distribution’s longest wires by 50 percent, with concurrent improvements of up to 75 percent in latency and 50 percent in interconnect energy dissipation.¹¹

I/O Technologies

For on-chip interconnect problems, novel chip I/O interconnect technologies offer new options. Wafer-level batch processing of both the chip package and *x-y-z* compliant I/O interconnects implemented with low cost, high-density “sea-of-leads” fabrication is a promising new technology.¹³ Researchers have demonstrated sea of leads I/O interconnect densities of 10,000 per cm².¹⁴ Potential advantages of sea-of-leads technology include enhanced I/O bandwidth; reduced on-chip area consumption of global power and signal distribution networks, decreased simultaneous switching noise, and improved isolation in mixed signal systems. The technology also satisfies 3D structures’ enormous I/O demand, is compatible with optical I/O interconnects, and has lower packaging and assembly costs than traditional I/O technologies.

For frequencies above about 10 GHz, optical clock distribution to gigascale chips can offer advantages over electrical distribution networks. The most notable advantage is in decreased power dissipation and jitter (that is, random variations in the clock pulse’s arrival time from cycle to cycle). In addition, for longer interconnects at the printed wiring-board level of packaging, the chip-to-chip optical interconnects promise to have better bandwidth and energy dissipation than electrical interconnects.¹⁵ You can implement optical clocking and I/O interconnects using on-chip vertical-cavity-surface emitting lasers (VCSELs) and quantum-well detectors. Promising approaches here include flip-chip bonding of laser and detector chips on a complementary metal-oxide silicon (CMOS) substrate, and, in the longer term, heteroepitaxial fabrication of compound semiconductor devices on silicon.^{15,16}

Interconnects: A Long-Term View

New interconnect technologies will extend Moore’s law by letting us effectively use future transistor-scaling advances, rather than watching them dissipate due to interconnect limitations. These interconnect innovations will likely constitute the key differences in future generations of silicon technology. This transition from transistor-centric to interconnect-centric silicon technology began in the late 1990s, when industry replaced aluminum interconnects with copper in advanced silicon chips.^{3,5} The duration of this transition period—which we can regard as a prelude to the interconnect era—is a matter of keen interest to the chip industry.

The Steel Analogy

We can broadly project the silicon technology era's future course by considering *analogical limits*.² It's particularly useful to compare steel, which was the industrial revolution's principal structural material, with silicon, which is the information revolution's principal electronic material. US steel production increased at an exponential rate from 1860 through 1900. It was the critical building material for the Industrial Revolution's most prominent manifestations: skyscrapers, bridges, railroads, and so on. From 1900 to about 1950, US steel production increased at a more modest rate. However, at mid-century, steel remained the industrial world's principal structural material. It maintains a strong position even today, despite being supplanted by materials such as aluminum and plastics in numerous applications.

In the second half of the 20th century, silicon was the information revolution's principal electronic material, and it's quite likely to maintain that position for the next two decades, as exponential growth of both productivity and performance continue. Indeed, given its outstanding features, it's quite feasible that silicon technology will dominate for decades beyond the 10-nm generation (circa 2020). The silicon-steel analogy suggests this prospect. After the 10-nm generation, however, Moore's law will no longer persist. After six decades of exponential increase rates, the number of transistors per chip is likely to saturate. Then what happens?

Scaling and the Latency Challenge

Scaling down transistor dimensions—and the resulting improvements in both transistor cost and performance—has been the principal mechanism for fulfilling Moore's law. Unfortunately, scaling down interconnect cross-sectional dimensions degrades performance, which has compelled the present interconnect-centric approach to silicon technology. The urgency of interconnect-centric design can only increase as scaling continues.

A simple reasoning process elucidates this assertion. For small wires, latency is the most challenging performance metric. Latency is given by the expression $\tau = RC$ where R and C are an interconnect's total resistance and capacitance. A wire's resistance is commonly expressed as $R = \rho(L/WH)$ where ρ is the metal's resistivity, and L , W , and H are the metal conductor's length, width, and height, respectively. Assuming that ρ remains constant, and W and H are scaled proportionately

for a wire of constant length, R increases quadratically as $1/WH$. Neglecting fringing, an isolated wire's capacitance is approximately $C = \epsilon(WL/T)$, where ϵ is the insulator's permittivity and T is its thickness. Assuming that ϵ remains constant, and W and T are scaled proportionately for a wire of constant length, C remains constant. Consequently, $\tau = RC$ increases quadratically as $1/HT$.

However, there is more to the problem. As wire cross-sectional dimensions W and H continue to scale downward and circuit speed continues to increase, several factors exacerbate the interconnect latency problem.⁹

First, surface scattering imposes rapid increases in effective resistivity, ρ , because wire cross-sectional dimensions become smaller than the bulk copper electron's mean free path length. Basically, the thin, relatively high resistivity liners—which must surround a copper interconnect to prevent copper atoms' migration into the silicon—become thicker than the copper interconnect itself. This effectively reduces the copper's cross-sectional area and thus increases wire resistance and hence latency.

Second, power dissipation causes temperature increases in the wires; therefore, resistivity increases. Finally, high-speed operation creates a greater current density near the wires' periphery than in their central region. This so-called "skin effect" further increases wire resistance and hence latency; combining this effect with surface scattering causes nonlinear behavior, significantly increasing wire resistance and latency.

Given that interconnect latency tends to increase rapidly as scaling continues, what antidotes to this problem are conceivable? The most potent antidote would be to discover an interconnect nanotechnology that provided high-temperature, superconductive materials with resistivity $\rho \rightarrow 0$. For $\rho \rightarrow 0$, we can no longer calculate an interconnect's latency using the approximation $\tau = RC$. When the RC product is extremely small, two mechanisms determine interconnect latency. For relatively short interconnects, latency is the time a driver transistor requires to charge its load capacitance according to the relationship, $t_d = C_t V/I$, where C_t is the total transistor and interconnect capacitance, V is the interconnect voltage swing, and I is the transistor drive current.⁹ This expression accurately describes binary transition time t_d from about 1960 to 1990. During that period, transistor delay commonly dominated interconnect delay because large cross-sectional dimensions resulted in small interconnect resistance.

For longer interconnects with $\rho \rightarrow 0$, an electromagnetic wave's sheer time-of-flight fundamentally constrains interconnect latency.⁹ An approximate time-of-flight expression is $T_0F = (\epsilon_r)^{1/2} L/c_0$, where ϵ_r and L are the relative insulator permittivity and interconnect length, respectively, and c_0 is light's velocity in free space (a fundamental limit). A brief calculation reveals interconnects' inescapable tyranny even for this extreme case of superconductive behavior or $\rho \rightarrow 0$. A simple model for the switching time of a 10-nm channel length transistor is $t_d = L_{ch}/v_{th}$ where L_{ch} is channel length and $v_{th} = 10^7$ cm/sec is the channel's average carrier velocity, which we assume is an electron's thermal velocity at room temperature. Therefore, for a 10-nm generation transistor and an average channel carrier velocity equal to the thermal velocity, transistor switching time delay $t_d = 0.1 \times 10^{-12}$ sec. = 0.1 picosecond. For an ideal interconnect with $\rho = 0$ and $\epsilon_r = 1$, the interconnect length traveled by an electromagnetic wavefront in 0.1 ps is $L = T_0F c_0/(\epsilon_r)^{1/2} = 30$ micrometers (μm). Thus, an ideal superconductive interconnect with a vacuum insulator whose length exceeds 30 μm will have latency exceeding that of a 10-nm transistor! Moreover, the interconnect's switching energy transfer will be much larger than that of a minimum-size 10-nm transistor.

As semiconductor transistors replaced vacuum tubes in the 20th century, radically new nanoelectronics will supplant silicon technology in this century. This new technology will likely use "transistors" that approach, if not surpass, the 0.1 ps latency of 10-nm generation silicon transistors. Consequently, if we optimistically assume that the interconnects of this post-Moore's Law nanotechnology will be superconductive, their latency will exceed that of the transistors for interconnect lengths greater than 30 μm , while long, on-chip interconnect lengths will be 1,000 times greater at 30 mm. Consequently, mainstream electronics will have an interconnect era beyond Moore's law.

References

1. G.E. Moore, "Progress in Digital Integrated Electronics," *IEEE Int'l Electron Device Meeting Tech. Digest*, IEEE Press, 1975, pp. 11–13.
2. J.D. Meindl, "Theoretical, Practical and Analogical Limits in ULSI,"

IEEE Int'l Electron Device Meeting Tech. Digest, IEEE Press, 1983, pp. 8–13.

3. *International Technology Roadmap for Semiconductors, Executive Summary*, ITRS, 1999, p. 25.
4. J.D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proc. IEEE*, vol. 83, no. 4, 1995, pp. 619–635.
5. *International Technology Roadmap for Semiconductors, Executive Summary*, ITRS, 2001, p. 38.
6. J.D. Meindl and J.A. Davis, "The Fundamental Limit on Binary Switching Energy for Terascale Integration," *IEEE J. Solid State Circuits*, vol. 35, no. 10, 2000, pp. 1515–1516.
7. D.J. Frank et al., "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE*, vol. 89, no. 3, 2001, pp. 259–288.
8. Q. Chen et al., "A Comprehensive Analytical Subthreshold Swing (S) Model for Double Gate MOSFETs," *IEEE Trans. Electronic Devices*, vol. 49, no. 6, 2002, pp. 1086–1090.
9. J.D. Meindl et al., "Interconnect Opportunities for Gigascale Integration," *IBM J. Res. & Dev.*, vol. 46, nos. 2–3, 2002, pp. 245–262.
10. A. Naeemi and J.D. Meindl, "Optimal Global Interconnecting Devices for GSI," *IEEE Int'l Electron Device Meeting Tech. Digest*, IEEE Press, 2002.
11. J.D. Meindl, "The Evolution of Monolithic and Polyolithic Interconnect Technology," *Digest of Papers of Symposium on VLSI Circuits*, June 2002, pp. 2–5.
12. M. Igarashi et al., "A Diagonal-Interconnect Architecture and Its Application to RISC Core Design," *Digest of Papers of IEEE Int'l Solid-State Circuits Conf.*, IEEE Press, 2002, pp. 210–211.
13. A. Naeemi et al., "Sea of Leads: A Disruptive Paradigm for a System-on-a-Chip," *Digest of Papers of IEEE Int'l Solid-State Circuits Conf.*, IEEE Press, 2001, pp. 280–281.
14. M. Bakir et al., "Sea of Leads Ultra-High Density Compliant Wafer Level Packaging Technology," *Proc. IEEE Electronic Components and Technical Conf.*, IEEE Press, 2002, pp. 1087–1094.
15. D. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," *Proc. IEEE*, vol. 88, no. 6, 2000, pp. 728–749.
16. E.A. Fitzgerald and L.C. Kimerling, "Silicon-Based Microphotonics and Integrated Optoelectronics," *MRS Bulletin* 234, Apr. 1998.

James D. Meindl is director of the Joseph M. Pettit Microelectronics Research Center and the Joseph M. Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology, where he also directs the Interconnect Focus Center, a multiuniversity research effort. His research interests focus on physical limits on gigascale integration. He received his BS, MS, and PhD in electrical engineering from Carnegie Mellon University. He is a fellow of the IEEE and the American Association for the Advancement of Science, and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. Contact him at the Microelectronics Research Ctr., Rm. 123, Georgia Inst. of Tech., 791 Atlantic Dr. NW, Atlanta, GA 30332-0269; james.meindl@mirc.gatech.edu.