

STANFORD UNIVERSITY

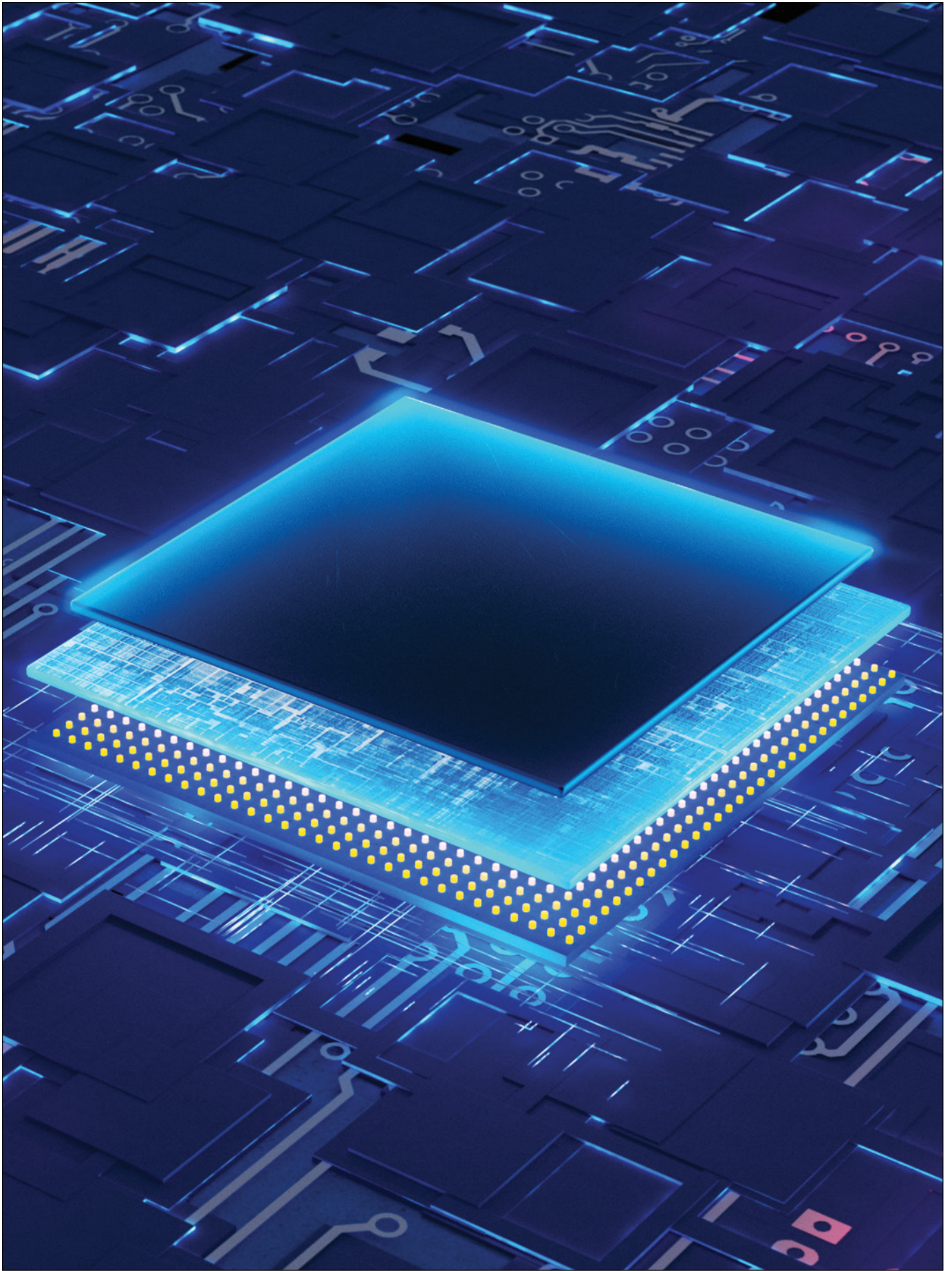
THE STANFORD EMERGING TECHNOLOGY REVIEW 2025

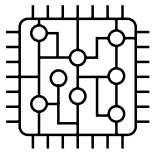
A Report on Ten Key Technologies and Their Policy Implications

CO-CHAIRS Condoleezza Rice, John B. Taylor, Jennifer Widom, and Amy Zegart

DIRECTOR AND EDITOR IN CHIEF Herbert S. Lin | **MANAGING EDITOR** Martin Giles







SEMICONDUCTORS

KEY TAKEAWAYS

- The growing demand for artificial intelligence and machine learning is driving innovations in chip fabrication that are essential for enhancing computational power and managing energy efficiency.
- Advances in memory technologies and high-bandwidth interconnects, including photonic links, are critical for meeting the increasing data needs of modern applications.
- Even if quantum computing advancements are realized, the United States will still need comprehensive innovation across the technology stack to continue to scale the power of information technology.

Overview

Semiconductors, often in the form of microchips, are crucial components used in everyday physical devices, from smartphones and toasters to cars and lawn mowers. Chips also control heating and cooling systems, elevators, and fire alarms in modern buildings. Traffic lights are controlled by chips. On farms, tractors and irrigation systems are controlled by chips. Modern militaries could not function without chips in their weapons, navigation devices, and cockpit life-support systems in fighter jets. The list goes on and on—in every aspect of modern life, chips are essential.

All chips are involved in the handling of information. Different types of them are specialized for different tasks. Some are processor chips that ingest data, perform computations on the data, and output the results of those computations. Memory chips store information and are used with processors. Still other chips act as interfaces between digital computations

and the physical world. In all these cases, some amount of energy is needed to represent each bit of information inside a chip. The magic of chips is that it takes several orders of magnitude less energy to represent information inside one than it takes to do so outside it (e.g., in wires leading to and from the chip). This means that, in a multi-chip system, much more energy and chip space are required for data moving between chips than for data that remains on a chip; this is one of the driving forces to integrate more functions on single chips.

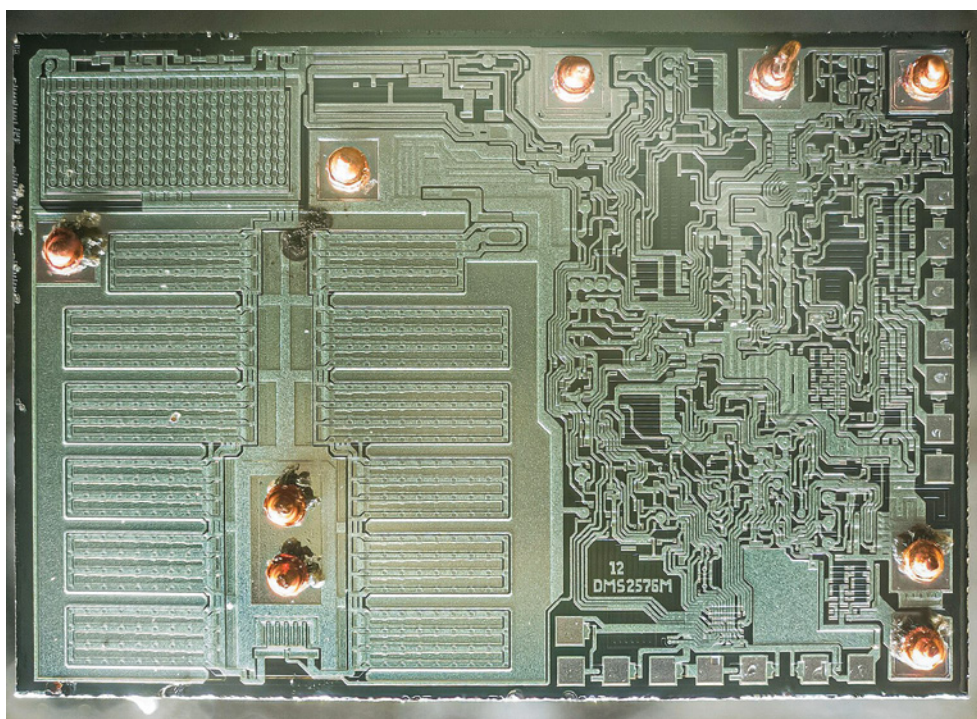
As chip fabrication technologies improve, it takes less energy and chip space to represent a given bit of information; hence, processing those bits becomes more energy efficient. This phenomenon is what has enabled the semiconductor industry to pack more processing power on chips over time—it enables

designers to create chips that do more complex processing (see figure 8.1), although the cost of designing them also increases with their complexity.

Recently, however, the energy costs associated with the hardware that holds information on a chip have been falling more slowly, and the cost of manufacturing per unit area has increased. This means that the cost and energy advantages of scaling have nearly stopped. As a result, researchers have been investigating other ways to improve computer technology and to deal with the problem of high design costs.

Since the best technologies for performing different chip functions are themselves different, systems still need to use different chips for those functions. Finding new ways to manage the inefficiency of

FIGURE 8.1 Increased energy efficiency has allowed designers to create more complex chips



Source: Wikimedia Commons. Used under CC BY-SA 4

FIGURE 8.2 Chip fabrication requires large factories that can produce chips at scale



Source: SkyWater Technology

information movement in and among chips, along with the issue of high design costs, is a central focus of research on semiconductors. Further improvement will take the form of innovation in design, materials, and integration methods.

Two aspects of chips are important for the purposes of this report. They must be designed and then fabricated (i.e., manufactured), and each function calls for different skill sets. Chip design is primarily an intellectual task that requires tools and teams able to create and test systems containing billions of components. Fabrication is primarily a physical effort that requires large factories, or fab facilities, that can produce chips by the millions and billions—and can cost billions of dollars and take several years to build from the ground up (see figure 8.2).

Fabrication integrates many complex processes to produce chips. Each one requires substantial expertise to master and operate, and the integration of all of them requires still further expertise. For these reasons, modern fabrication plants are operated by workforces with a substantial number of people trained in engineering.

Fabrication also entails a significant degree of process engineering to continue to improve process technology and to achieve stringent manufacturing standards. For example, the “clean rooms” in which chips are made require air that is one thousand times more particle-free than the air in a hospital operating room.¹

Because chip design and chip fabrication are so different in character, only a few companies, such as Intel, do both. Many businesses specialize in design, including Qualcomm, Broadcom, Apple, and Nvidia. Such companies are called “fabless” in recognition of the fact that they do the design work and outsource fabrication to others—a strategy based on the theory that the former activity has higher profit margins compared to the latter.

Today, “others” usually means one company: Taiwan Semiconductor Manufacturing Company (TSMC), which is by far the world’s largest contract chip-manufacturing company. In 2024, TSMC controlled over 60 percent of the world’s contract semiconductor manufacturing and 90 percent of the world’s advanced chip manufacturing.² Samsung, in South

Korea, is a distant second, with around 13 percent of the world's chip manufacturing.³ United Microelectronics Corporation, also based in Taiwan, ranks third at about 6 percent.⁴

By contrast, US chip-manufacturing capacity has lost significant ground. Fabrication plants in America accounted for 37 percent of global production in 1990, but their share dropped to just 12 percent by 2021.⁵ Industry concentration, low US production capacity, and geopolitical concerns about China's intentions toward Taiwan mean the global supply chain for chips will remain fragile for the foreseeable future, despite the passage of the Creating Helpful Incentives to Produce Semiconductors (CHIPS) and Science Act of 2022 (discussed in more detail later in this chapter).

Key Developments

Moore's Law, Past and Future

For over half a century, information technology has been driven by improvements in the chip fabrication process. In 1965, Intel cofounder Gordon Moore observed that the cost of fabricating a transistor was dropping exponentially with time—an observation that has come to be known as Moore's law. It's not a law of physics but rather a statement about the optimal rate at which economic value can be extracted from improvements in the chip fabrication process.

Although Moore's law is often stated as the number of transistors on a chip doubling every few years, historically the cost of making a chip was mostly independent of the components on it. This has meant that every few years, a chip whose size and cost remains approximately the same will have twice the number of transistors on it.

Moore's law scaling (i.e., the exponential increasing of the number of transistors on a chip) meant that each year one could build last year's devices for less

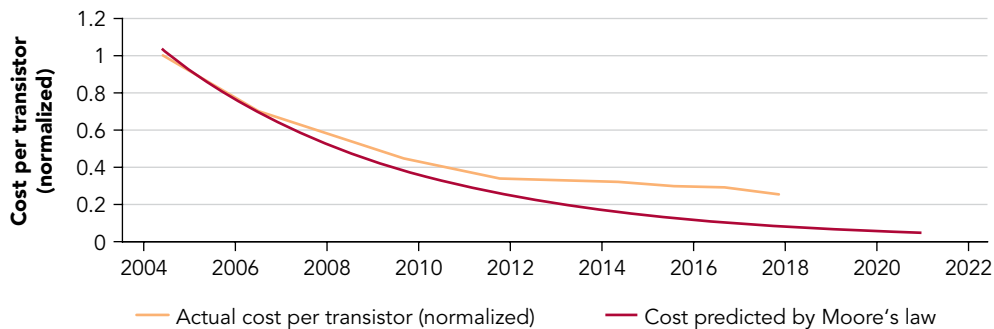
money than before or build a more powerful system for the same cost. This scaling has been so consistent that it is widely believed that the cost of computing will always decrease with time. This expectation is so pervasive that in almost all fields of work, people are developing more complex algorithms to achieve better results while relying on Moore's law to rescue them from the consequences of that additional complexity.

But the future will not look like the past. As the complexity of chips increases, the traditional cost-benefit relationships associated with Moore's law scaling are diminishing, leading to rising costs in chip manufacturing. As figure 8.3 shows, the actual cost of a chip per transistor (represented by the solid red line) was tracking the cost predicted by Moore's law (the dashed blue line) relatively well from 2004 to 2012.⁶ However, the actual cost per transistor started to level off around 2012—and it has not kept up with Moore's law predictions since then.⁷

Against this backdrop, the past year has witnessed significant advancements and challenges in the semiconductor industry. For example, increasing demand for computing power driven by artificial intelligence (AI) and machine learning (ML) applications (as discussed in chapter 1 on artificial intelligence) has led to a surge in the development of, and demand for, advanced graphical processing units (GPUs), creating a strain on both production and energy resources. Advanced GPU systems, such as Nvidia's GB200 NVL72 system, are pushing the boundaries of what is possible in terms of performance and power consumption. This surge in demand underscores the importance of finding innovative solutions to meet computational needs without compromising on energy efficiency or physical space.

The rapid expansion of AI applications is also driving demand for more specialized hardware. Traditional processors are no longer sufficient for the intensive computational tasks required by modern AI algorithms. As a result, there has been a significant investment in developing GPUs and other specialized

FIGURE 8.3 Cost per transistor over time



Source: Adapted from Steve Mollenkopf, "Our Future Is Mobile: Accelerating Innovation After Moore's Law," presentation at the Electronics Resurgence Initiative Summit, Detroit, MI, July 15–17, 2019

processors designed specifically for AI workloads. This shift is reshaping the semiconductor industry, emphasizing the need for high-performance, energy-efficient computing solutions.

Chiplets and 2.5-D Integration

Chiplets and 2.5-D (2.5-dimensional) integration represent a significant shift from traditional monolithic chip design. Chiplets are smaller functional blocks of silicon that can be combined to create complete systems, offering greater flexibility and customization. By enabling the use of different manufacturing technologies for various components, they facilitate the integration of high-density memory with high-performance processing units, resulting in improved system bandwidth and performance.

Central to 2.5-D integration, which puts chiplets next to one another in an integrated circuit, is the interposer—a specialized substrate facilitating communication between the chiplets. This technology enables more energy-efficient data transfer compared to traditional circuit board wiring. By allowing optimized memory, compute, and communication chips to reside side by side, it enhances system

performance while reducing the need for full integration on a single chip.

From an economic perspective, the chiplet approach offers significant advantages. Semiconductor companies can create a set of building-block chiplets that can be combined in various ways, allowing for a wide range of products with different performance characteristics. This strategy allows companies to better tailor their products to specific application domains and more effectively monetize their silicon investments, ultimately leading to increased product diversity and market responsiveness.

Given the changing economics of scaling, the use of chiplets reduces costs overall and allows for more customized solutions. Semiconductor company AMD's approach of keeping components that transfer data to and from devices in older technology nodes while advancing core computing resources with the latest processes exemplifies this strategy. Moreover, this modular strategy facilitates the integration of emerging technologies such as photonics (discussed in greater detail later in this chapter), which can significantly enhance communication speeds and bandwidth within and between chips.

Three-Dimensional Heterogeneous Integration

The research community is working on an even more advanced technique: three-dimensional (3-D) heterogeneous integration. Unlike 2.5-D integration, where different chiplets are placed on a common substrate, 3-D heterogeneous integration is a semiconductor-manufacturing technique that involves the vertical stacking of different electronic components, such as processors and memory. The heterogeneous aspect means that these stacked components can be made from different materials and technologies optimized for their specific functions. For example, a processor made from one type of material can be stacked with memory made from another, each using the most suitable technology for its purpose. This approach has the potential to improve performance and efficiency by reducing the distance data travels between components, making devices faster and more compact—albeit at the cost of a more complex fabrication process and a harder heat-dissipation task for the resulting chips.

Need for High Bandwidth

A second approach to enhancing performance focuses on improving bandwidth and interconnect technologies. AI-training models must handle vast amounts of data, and high-speed interconnects such as those employed in Nvidia's H100 systems play a critical role in facilitating the rapid transfer of data between compute units and memory.

The physical limitations of traditional electrical interconnects are one of the primary barriers to improving bandwidth. As data rates increase, so do power consumption and heat generation, which can limit the performance and reliability of semiconductor devices. For example, Nvidia's GPU modules achieve bandwidths on the order of several terabytes per second. The company's NVLink enables the connection of up to 256 GPUs, forming a supercomputer pod with immense computational power and bandwidth. Each GPU in these setups is approximately

ten times more powerful than a consumer one, dissipating a kilowatt of power each. This results in a pod that dissipates 10 kilowatts (kW), and a full structure of 256 GPUs dissipates over a third of a megawatt. (For comparison, a 2,500-square-foot house might require a furnace that generates about 20 kW of heat.) Thus, thermal management stands out as a critical issue. Effective thermal-management solutions, such as advanced cooling techniques and materials with high thermal conductivity, will be essential to maintaining performance and reliability in high-performance computing systems.

Photonic Links and Components

It is becoming difficult to scale electrical data-transmission links and their associated bandwidth. Innovations such as silicon photonics are emerging as potential solutions, offering the promise of higher bandwidth and lower-power consumption by using light to transmit data. Photonics are the optical analogue of electronics—the latter use electrons for signaling and carrying information while photonics use photons (light) for the same purposes.

Compared to traditional electrical interconnects, photonic links have the potential to reduce energy consumption and increase bandwidth in data centers and long-distance data transmissions.⁸ Furthermore, they can handle different wavelengths on a single optical fiber simultaneously, thereby enhancing data-carrying capacity and making photonics an attractive solution for high-performance computing and data center applications. By replacing electrical interconnects with optical ones, data centers can reduce the amount of energy required for transmitting data, leading to lower operational costs and a smaller environmental footprint. These advantages of photonics have always been drivers of research in this area, but the recent rise in the demand for power-hungry AI-enabled applications has created even more impetus for such research.

Integrating photonic components with silicon-based technologies is challenging as a result of material

incompatibilities; for example, efficient light-emitting materials like III–V semiconductors do not integrate well with silicon. (A III–V semiconductor is made by combining boron, aluminum, gallium, or indium with nitrogen, phosphorus, arsenic, or antimony.) While useful for light detectors, silicon is ineffective for light emission, complicating the scalable integration of these technologies at the chip or circuit board level. Overcoming these challenges is crucial for realizing the full potential of photonic links in large-scale, low-energy applications.

Memory Technology Developments

Memory technology continues to evolve, with innovations in both stacking and new materials. Techniques such as stacking multiple layers of flash memory are pushing the boundaries of what is possible, enabling higher density and better performance. These advancements are crucial for supporting the growing data needs of modern applications, from AI to big data analytics.

Dynamic random-access memory (DRAM) and flash memory technologies have both seen significant advancements in recent years, but the associated increase in manufacturing cost means there has been only modest improvement in cost per bit. The development of 3-D structures, such as vertical DRAM transistors, has allowed for continued scaling of memory density by overcoming the physical limitations of traditional planar structures and has enabled the production of memory devices with higher capacity and improved performance.

As memory technologies scale, maintaining performance and reliability becomes increasingly challenging. For DRAM, issues such as leakage currents and quantum effects limit the scalability of capacitors and transistors. To address these challenges, researchers have developed advanced manufacturing techniques for the creation of complex 3-D structures that enable increased storage density while maintaining the required electrical characteristics.

Similarly, NAND flash memory—the most common type of such memory—has faced scaling challenges due to the limitations of traditional planar cell architectures. The development of 3-D NAND, which involves stacking multiple layers of memory cells vertically, has enabled continued increases in storage density. However, this approach requires sophisticated manufacturing processes to ensure the reliability and performance of the resulting memory devices.

Emerging memory technologies, such as magnetoresistive random-access memory⁹ and phase-change memory,¹⁰ are also gaining traction in some applications. These technologies offer unique advantages in terms of speed, endurance, and energy efficiency, making them attractive alternatives to traditional embedded nonvolatile memory solutions. (Nonvolatile memory retains its contents even when power is turned off.)

Further innovations in memory technology are critical for enabling the continued growth of data-intensive applications. From AI-training models to cloud computing and big data analytics, modern applications require vast amounts of memory to store and process data efficiently.

Quantum Computing Advancements

Quantum computing remains a field of intense research and development, with significant progress made in both the number and quality of quantum bits, or qubits. Recent innovations in error correction and the potential for practical quantum computing could revolutionize specific applications,¹¹ although commercial viability remains years away. The promise of quantum computing lies in its potential to perform certain complex calculations at unprecedented speeds, with possible relevance for applications such as cryptography, materials science, and complex system simulations.

Quantum computing offers a fundamentally different approach to computation compared to classical

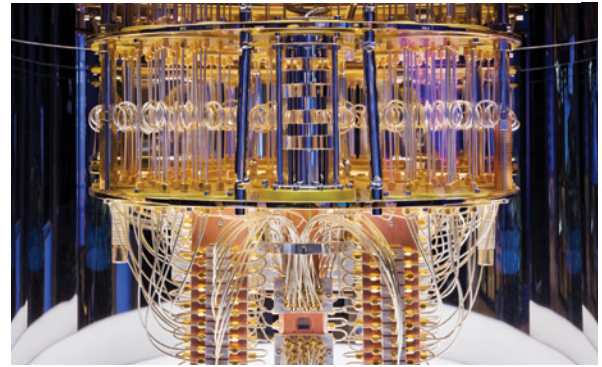
computing. Classical computers use individual bits as the smallest unit of data, with each being 0 or 1. In contrast, the qubits in quantum computers, like the one shown in figure 8.4, can be in multiple states simultaneously due to the principles of quantum mechanics. This property, known as superposition, allows quantum computers to process a vast number of possibilities at once, a phenomenon called quantum parallelism.

However, quantum computing also presents significant challenges. Using it to implement an application is hard because reading the result of a quantum computation generally yields a result corresponding to only one of the many results possible when such a computation is performed on a superposition state. Realizing the advantage of a quantum computation means reading the correct result, which in turn requires designing computational algorithms so that the correct result is the one most likely to show up. Moreover, quantum computing operations are analog rather than digital in nature and thus can be disrupted by “noise” in the environment, such as vibrations and fluctuations in temperature.

Recent advancements in algorithms and error-correcting codes that can compensate for such errors have reduced the overhead associated with error correction. Still, many further advancements in error correction will be necessary before quantum computing can be more widely applied. Creating a useful quantum computer necessitates innovation across the entire quantum software and hardware stack: algorithms, compilers, control electronics, error correction, and quantum hardware.

Various technologies are under consideration for the physical construction of qubits, including trapped ion, superconducting, cold atom, photonic, crystal defect, quantum dot, and topological technologies.¹² The most advanced quantum computing machines currently use trapped ion or superconducting qubits. However, neither technology has a clear path to scaling up to larger machines, prompting ongoing research into other approaches.

FIGURE 8.4 A close-up view of a quantum computer



Source: IBM. Used under CC BY-ND 2.0

Recent work has improved the fidelity of a modest number (around thirty) of qubits and has reduced the overhead of quantum error correction. But far larger numbers of high-quality qubits (two or three orders of magnitude more) will be needed for quantum computers to become more broadly useful. A large number of companies and research labs have been pushing forward with work on this area. Progress on the quantum algorithm front is harder to track. While many groups are working on finding practical applications for early quantum computers, no such applications have yet been publicized.

Over the Horizon

The Impact of Technology

The semiconductor industry is poised for significant advancements in coming years, driven by the growing demands of AI, especially ML, and high-performance computing. The introduction of new technologies, such as 2.5-D integration, chiplets, and photonic interconnects, is expected to play a crucial role in meeting these demands. These innovations will enhance performance, increase bandwidth, and

improve energy efficiency, addressing the limitations of traditional semiconductor designs.

Quantum computing remains an important area of development, with progress in error correction and qubit quality being made. However, its timeline remains uncertain. Even if it becomes successful, it will likely be useful for only a limited class of applications and won't replace today's semiconductor technology. Quantum computers will complement, rather than supplant, classical semiconductors, addressing specific complex problems in fields such as cryptography, materials science, and complex simulations.

Emerging memory technologies and advanced manufacturing techniques are also critical for the industry's growth. Innovations in 3-D memory stacking and integration with processors will improve data-transfer speeds and reduce latency, meeting the increasing data requirements of modern applications. The development of advanced materials and transistor architectures will further push the boundaries of semiconductor capabilities, enabling continued miniaturization and enhanced performance.

Finally, as Moore's law reaches its limits, future improvements in computing will rely more on optimizing algorithms, hardware, and technologies for specific applications rather than on general technology scaling. This requires innovation across the entire technology stack, from materials to design methods. However, the industry faces a paradox: The need

for radical innovation conflicts with the high costs and long timelines of chip development, which can exceed \$100 million and take over two years.

To address this, the industry must make system-design exploration easier, cheaper, and faster. Researchers are working to ensure that specific design changes to a chip do not require redesign of the entire chip. Solutions include enabling software designers to test custom accelerators without deep hardware knowledge and developing tools for application developers to make small hardware extensions to base platforms. This approach, described in more detail in the inaugural *Stanford Emerging Technology Review* (2023), depends on the involvement of major technology firms, who would need to participate in an app store-like model for hardware customization, balancing open innovation with profit motives.

CHALLENGES OF INNOVATION AND IMPLEMENTATION

A critical challenge facing the US semiconductor industry is its significant talent shortage, particularly in hardware design and manufacturing. For example, the Semiconductor Industry Association projects the number of jobs in the sector in the United States will grow by nearly 115,000 by 2030, to a total of approximately 460,000.¹³ Moreover, it estimates that roughly 67,000, or 58 percent, of these new jobs risk going unfilled at current degree-completion rates. Looking at just the new jobs that are technical

A critical challenge facing the US semiconductor industry is its significant talent shortage, particularly in hardware design and manufacturing.

in nature, the percentage at risk of going unfilled is higher, at 80 percent. Almost two-thirds of the unfilled jobs would require at least a bachelor's degree in engineering.¹⁴

The pipeline of college graduates interested in semiconductors is also troubling. Student interest in hardware has diminished as graduating students flock to software companies.¹⁵ Several factors appear to play a role, including the perception of higher salaries in software development and a lack of awareness about the diverse and exciting opportunities in hardware.

Since appropriately trained people are the only real source of new ideas, this trend does not bode well for the industry. Addressing this issue requires more and even closer collaboration among educational institutions, industry, and government to develop programs that attract and train the next generation of semiconductor engineers and researchers.

Policy, Legal, and Regulatory Issues

Supply chain resilience Building a resilient and diversified supply chain is critical for mitigating geopolitical risks. Investing in domestic manufacturing capabilities and promoting regional cooperation will enhance supply chain security and ensure a steady supply of essential components.

Geopolitical risks The extreme concentration of semiconductor manufacturing in Taiwan poses a significant risk to the global supply chain. Political tensions, trade disputes, and potential conflict in the region can disrupt the supply of critical components, impacting industries and economies worldwide. Diversifying supply chains and investing in domestic manufacturing capabilities are essential strategies for building resilience against geopolitical risks. Initial steps toward this were taken in the passage of the CHIPS and Science Act of 2022, which earmarked \$52.7 billion for semiconductor manufacturing, research, and workforce development,

plus significant tax credits for private investment in the field. Full implementation of the act has not yet occurred, partly because not enough time has elapsed and partly because the appropriations it called for have not been fully funded.

NOTES

1. Intel Corporation, "Inside an Intel Chip Fab: One of the Cleanest Conference Rooms on Earth," March 28, 2018, <https://www.intc.com/news-events/press-releases/detail/165/inside-an-intel-chip-fab-one-of-the-cleanest-conference>.
2. Jeremy Bowman, "This 1 Number May Ensure TSMC's Market Dominance," *The Motley Fool*, August 17, 2024, <https://www.nasdaq.com/articles/1-number-may-ensure-tsmcs-market-dominance>.
3. Counterpoint, "Global Semiconductor Foundry Revenue Share: Q1 2024," June 12, 2024, <https://www.counterpointresearch.com/insights/global-semiconductor-foundry-market-share/>.
4. Counterpoint, "Global Semiconductor Foundry Revenue Share."
5. Semiconductor Industry Association, *2021 State of the U.S. Semiconductor Industry*, 2021, <https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf>.
6. Electronics Resurgence Initiative 2.0, "2019 Summit Agenda, Videos, and Slides," Defense Advanced Research Projects Agency, accessed August 30, 2023, <https://eri-summit.darpa.mil/2019-archive-keynote-slides>.
7. In fact, transistor costs are usually plotted in a semi-log graph, where the log of the cost is plotted against time. In these plots the exponential decline in transistor costs becomes a straight line. The fact that the scale of the y-axis is linear is a clear indication that Moore's law is over.
8. David A. B. Miller, "Attojoule Optoelectronics for Low-Energy Information Processing and Communications," *Journal of Light-wave Technology* 35, no. 4 (February 2017): 34696, <https://doi.org/10.1109/JLT.2017.2647779>.
9. Paolo Cappaletti and Jon Slaughter, "Embedded Memory Solutions: Charge Storage Based, Resistive and Magnetic," in Andrea Redaelli and Fabio Pellizzer, eds., *Semiconductor Memories and Systems* (Cambridge, MA: Woodhead Publishing, 2022): 159–215, <https://doi.org/10.1016/B978-0-12-820758-1.00007-8>.
10. Cappaletti and Slaughter, "Embedded Memory Solutions."
11. Ziqian Li, Tanay Roy, David Rodríguez Pérez, et al., "Autonomous Error Correction of a Single Logical Qubit Using Two Transmons," *Nature Communications* 15, no. 1681 (February 2024), <https://doi.org/10.1038/s41467-024-45858-z>.
12. An overview of these technologies can be found in Eunmi Chae, Joonhee Choi, and Junki Kim, "An Elementary Review on Basic Principles and Developments of Qubits for Quantum Computing," *Nano Convergence* 11, no. 11 (March 2024), <https://doi.org/10.1186/s40580-024-00418-5>.
13. Dan Martin and Dan Rosso, "Chipping Away: Assessing and Addressing the Labor Market Gap Facing the U.S. Semiconductor Industry," Semiconductor Industry Association and Oxford

Economics, July 25, 2023, <https://www.semiconductors.org/chipping-away-assessing-and-addressing-the-labor-market-gap-facing-the-u-s-semiconductor-industry/>.

14. Martin and Rosso, "Chipping Away."

15. Tom Dillinger, "A Crisis in Engineering Education—Where Are the Microelectronic Engineers?," SemiWiki, July 3, 2022, <https://semiwiki.com/events/314964-a-crisis-in-engineering-education-where-are-the-microelectronics-engineers>.

STANFORD EXPERT CONTRIBUTORS

Dr. Mark A. Horowitz

SETR Faculty Council, Fortinet Founders Chair of the Department of Electrical Engineering, Yahoo! Founders Professor in the School of Engineering, and Professor of Computer Science

Dr. David Miller

W. M. Keck Foundation Professor of Electrical Engineering and Professor, by courtesy, of Applied Physics

Dr. David Schuster

Professor of Applied Physics

Dr. Asir Intisar Khan

SETR Fellow and Visiting Postdoctoral Scholar in Electrical Engineering