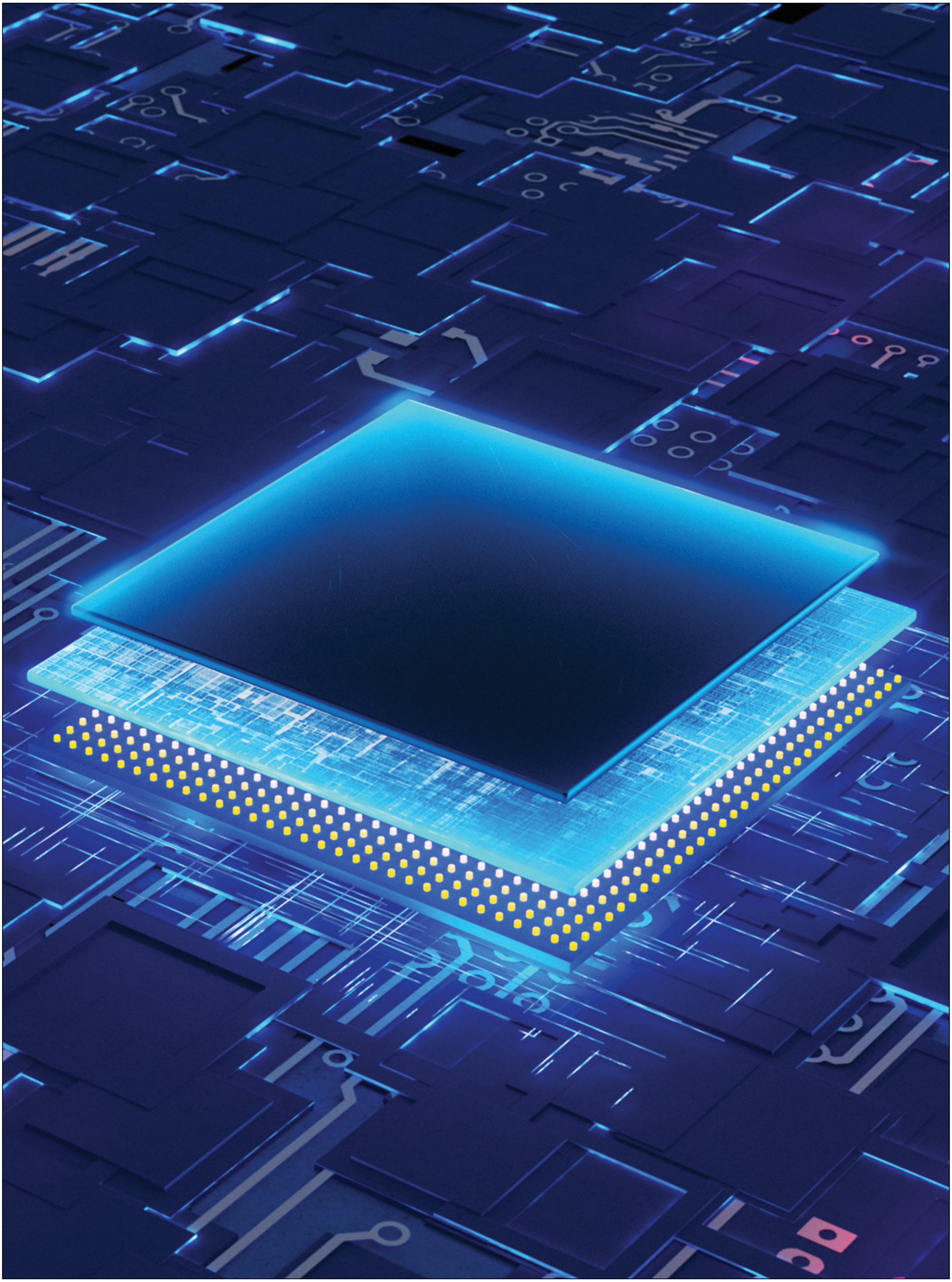
STANFORD UNIVERSITY

THE STANFORD EMERGING TECHNOLOGY REVIEW 2026

A Report on Ten Key Technologies and Their Policy Implications

CO-CHAIRS Condoleezza Rice, Jennifer Widom, and Amy Zegart

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



09

SEMICONDUCTORS

KEY TAKEAWAYS

- The growing demand for artificial intelligence (AI) and machine learning is driving innovations in chip fabrication, along with advances in memory technologies and high-bandwidth interconnects such as photonic links, all of which are essential for enhancing computational power, managing energy efficiency, and meeting the increasing data needs of modern applications.

- Semiconductor manufacturing is the most precise manufacturing process that exists. It is used to advance work in energy and biotechnology in addition to information technology and AI.

- Strategic technology containment efforts directed against China help constrain Chinese capabilities in the short term. However, they are likely to drive China into a technology posture that is considerably more decoupled from the West and hence less vulnerable to Western pressure in the future.

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 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.

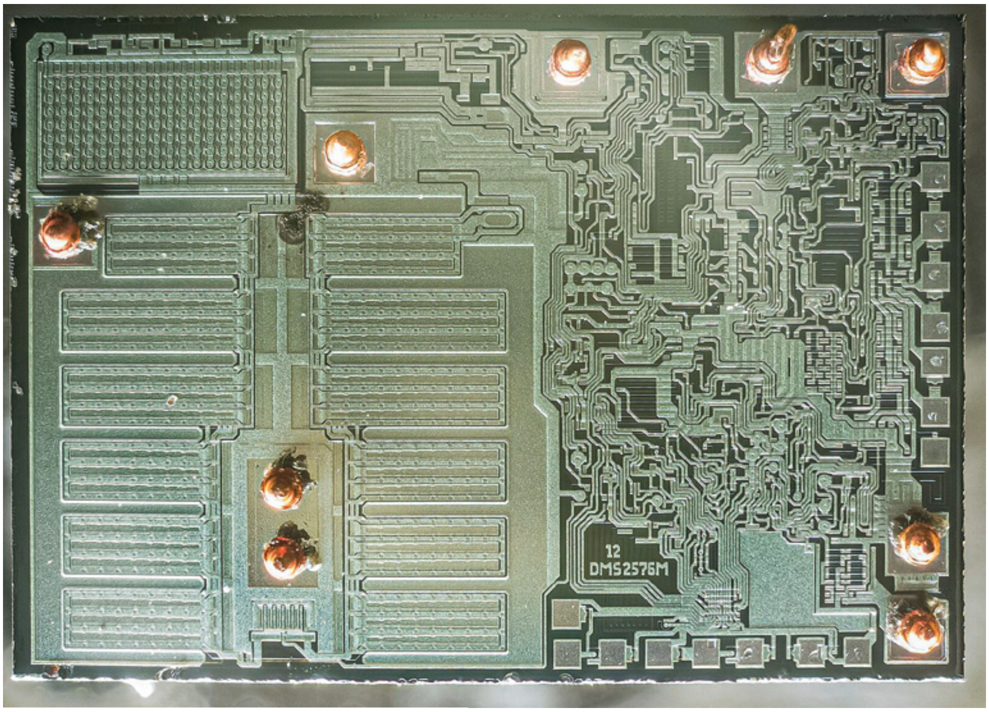
Most chips are involved in the handling of information. Different types of them are specialized for different tasks. Some are processor chips that ingest data, perform computations on the data, and output the results of those computations. Memory chips store information and are used with processors. 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 a single chip.

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 9.1). However, 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.

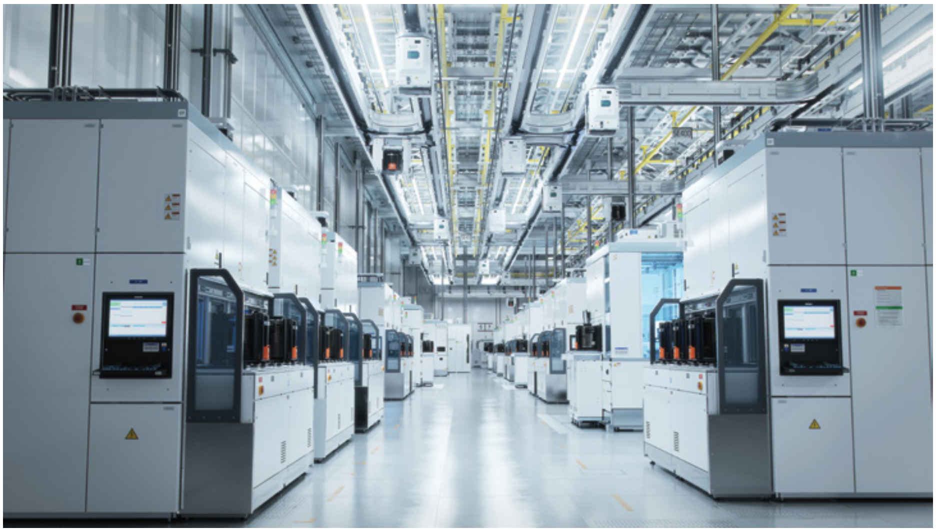
Since the best technologies for performing different chip functions are themselves different, systems still need to use different chips for those functions. Finding new ways to manage the inefficiency of information movement in and among chips, along with the issue of high design costs, is a central focus of

FIGURE 9.1 Improving fabrication has enabled the creation of more complex chips



Source: Wikimedia Commons, CC BY-SA 4.0

FIGURE 9.2 Chip fabrication requires large factories that can produce chips at scale



Source: IM Imagery/Shutterstock.com

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 9.2).

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 a 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. However, Intel is in trouble, and some technology analysts and former Intel directors think it should split its design and fabrication groups.[2] 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 than the latter.

Today, the "others" being outsourced to usually refer to one company: Taiwan Semiconductor

Manufacturing Company (TSMC), 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.[3] Samsung, in South Korea, is a distant second, with around 13 percent of the world's chip manufacturing.[4] United Microelectronics Corporation, also based in Taiwan, ranks third at about 6 percent.[5]

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.[6] 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).

The strength of semiconductor manufacturing affects more than just information technology. It is also the most precise manufacturing method on the planet and is now driving innovations in areas ranging from neuroscience and synthetic biology to energy and lighting. While many of these applications don't need the most advanced processing technology, they do require access to semiconductor fabrication and fabrication expertise.

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.

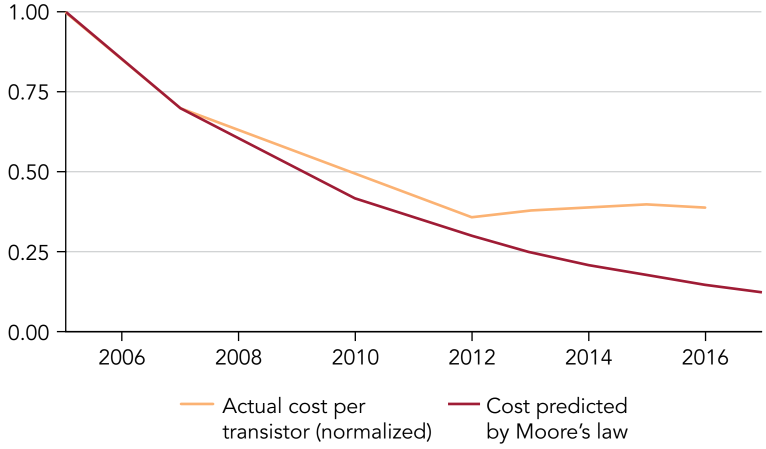
Although Moore's law is often stated as the number of transistors on a chip doubling every few years, historically what drove this scaling was that the cost of making a chip was mostly independent of the number of components on it. This has meant that every few years, a chip whose size and cost remains 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 could 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. It's an expectation 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 benefits associated with Moore's law scaling are diminishing, leading to rising costs in chip manufacturing. As figure 9.3 shows, the actual cost of a chip per transistor (represented by the solid orange line) was tracking the cost predicted by Moore's law (the solid red line) relatively well from 2004 to 2012.[7] 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.[8]

Historically, advances in technology have come from shrinking the size of the transistors and the wires that connected them, and the name of the technology was derived from the smallest feature in the design (e.g., a 130-nanometer [nm] chip was one in which the smallest feature was 130 nm across).

FIGURE 9.3 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, https://eri-summit.darpa.mil/docs/Mollenkopf_Steve_Qualcomm_Final.pdf

However, over a decade ago, it became increasingly difficult to boost circuit density (as measured by the number of circuits per square centimeter). Other approaches had to be found to do so, including having transistors use the vertical dimension to decrease their area.

Using the vertical dimension, along with other, more complex processing techniques (such as the use of new materials), enabled circuit density to continue to increase. But technology marketers continued to use a shrinking distance to characterize newer generations of more densely packed circuits, despite that distance no longer representing anything physical. In other words, the name became a marketing device (or, to put it more kindly, a generational technology label), even though it still sounded like it referred to a distance.

When the label had actual physical meaning, such as indicating the size of a feature, that number (e.g., 130 nm) could be used to make inferences about chip performance, such as the actual cost per compute or the energy needed for a computation. But once it became a marketing term, the connection between the label and the chip's performance was broken.

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). This has created a strain on both production and energy resources.

Traditional processors are not the best approach 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 processors designed specifically for AI workloads. This shift is reshaping the semiconductor industry, emphasizing the need for high-performance, energy-efficient computing solutions.

The increased density, provided by scaling and advanced packaging together, has enabled advanced GPU systems to connect a massive amount of compute in a small space, increasing performance. (Advanced packaging refers to the combination of circuitry from multiple semiconductor chips or dies into a single, compact electronic package, as in 2.5-dimensional [2.5-D] integration.) One example of this is Nvidia's GB200 NVL72 system. Unfortunately, the industry has been unable to shrink power and fabrication cost as rapidly as in the past, and thus these machines are both extremely expensive and very power hungry. They dissipate ten times the power of systems of a few years ago. They also require innovations in order to get the power into a system and to get the heat out. This surge in demand for high performance underscores the importance of finding innovative solutions to meet computational needs in the most energy-efficient and simple (reducing power and cost, respectively) way possible.

Addressing this challenge requires specialized hardware that is extremely efficient in computing the results needed by today's AI applications. This is the only known way to grow computation performance per dollar and per watt. Today, all computing devices used for AI applications, including GPUs, contain this type of specialized hardware. Nvidia reports its optimizations reduced the required energy to perform computations by a thousand times.[9] But even with these optimizations, current computing systems are dissipating a large amount of heat (over 100 kilowatts [kW] per rack). For comparison, a typical US household uses electric power at the rate of about 1 kW averaged throughout the year—and the power used in future systems is projected to keep increasing.

Chiplets and 2.5-Dimensional Integration

Integrating an enormous amount of compute and memory on a single piece of silicon is desirable for maximum energy efficiency. However, this poses two problems. First, some of today's most demanding applications require more computing resources than can be manufactured on a single piece of silicon. Second, the manufacturing processes for processors and memory are very different, so these two components can't be placed on the same piece of silicon.

One solution to these challenges is the use of chiplets and 2.5-D integration. Here, rather than forcing everything onto a single piece of silicon, multiple silicon pieces are connected together on an interposer (defined in more detail below) to create a much larger superchip.

This superchip combines processors and memory using chiplets and 2.5-D integration, which leverage different manufacturing technologies to optimize each component. Chiplets—functional blocks of silicon—can be combined in various ways, enabling vendors to tailor systems to customer needs. Central to 2.5-D integration is the interposer, a specialized substrate that connects chiplets and facilitates faster, more energy-efficient communication than traditional circuit board wiring. By allowing high-density memory, high-performance compute units, and communication chips to reside side by side, this approach boosts bandwidth, performance, and power efficiency while reducing the need for full integration on a single chip.

These superchips can contain both memory and processors. Representing a significant shift from traditional monolithic chip design, 2.5-D integration increases per-transistor cost because both the chips that the transistors are on and the substrate must be manufactured. Nevertheless, 2.5-D integration enables semiconductor companies to create a set of building-block chiplets that can be combined in various ways, allowing for the range of products with different performance characteristics mentioned earlier. 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. Exemplifying this strategy is the approach taken by semiconductor company AMD, which involves keeping components that transfer data to and from devices in older technology nodes while advancing core computing resources with the latest processes. 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.

High-Power Density

Moving the compute and memory elements closer together improves system performance. But all such systems generate heat—and combining more elements on a single chip increases the amount of heat that must be shed during system operation.

For example, in its NVL72 system, Nvidia packs seventy-two B200 2.5-D mega GPUs into a rack and uses its high-performance NVLinks technology to connect all the GPUs to each other, thus 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. Two of these are placed on a board that dissipates 2.7 kW, and the boards are placed in a rack that dissipates a total of 120 kW. Four to eight of these racks can be connected together with longer range links to create a super pod that dissipates 0.5 to 1 megawatt of heat. (For comparison, a 2,500-square-foot house might require a furnace that generates about 30 kW of heat.)

Thus, thermal management stands out as a critical issue. Historically, computing equipment was cooled by blowing cold air through a machine. But moving cold air is insufficient to remove this level of heat, and high-performance, compute-intensive machines must use liquid flowing through cooling plates to deal with it. The heat absorbed by the liquid must then be dissipated somewhere else, usually into the air using large air-conditioning units on the roof of a building. Effective thermal-management solutions, such as advanced cooling techniques and materials with high thermal conductivity, are essential to maintaining performance and reliability in high-performance computing systems.

Need for High Bandwidth

As the prior example showed, while 2.5-D integration helps provide local bandwidth, modern systems are large enough that they require many of these highly integrated superchips. Communicating information between these systems is therefore critical. AI-training models must handle vast amounts of data, and high-speed interconnects, such as those employed in Nvidia's B200 systems, play a critical role in facilitating the rapid transfer of data between compute units and memory.

Traditionally, this communication between the chips in a rack is done on electrical wires embedded in the boards that the chips connect to. As the communication rates have continued to rise, the physical limitations of traditional electrical interconnects are one of the primary barriers to improving bandwidth.

To overcome this bandwidth limit in communicating information from a chip on one board to a chip on another board, researchers are developing "flying cable" connectors. These connectors are placed directly on the top side of a superchip and enable a high-performance cable to be attached directly to the chip, while the other end is connected directly to the connector to the other board. This cable is built to have the best possible signal transmission properties. Researchers are experimenting with both electrical and optical cables, providing an ability to increase interface speeds above the 100 billion bit per second per wire rate used today.

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 (i.e., memory that retains its contents even when power is shut off) 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 three-dimensional (3-D) structures (e.g., vertical DRAM transistors) has allowed for continued scaling of memory density by overcoming the physical limitations of traditional planar transistors. Three-dimensional packaging has enabled the production of memory devices with higher capacity and improved performance, known as high-bandwidth memory (HBM).

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. Both DRAM and flash memory moved to 3-D structures decades ago and have had to use increasingly complex structures to scale the memory cell size. Chip stacking has also been used for many years to increase the number of memory bits shipped in a single package. What has changed recently is the growth of the HBM market. These memories require a more complex chip-stacking technology called through-silicon vias (TSV), which needs many wires to run vertically through the chips.

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.

The boom in AI computing has also affected the DRAM industry. The enormous compute power of today's specialized ML systems means enormous amounts of data per second, or data bandwidth, are required to keep them busy.[10] This need for high-bandwidth memory has created the growing market for HBM mentioned earlier. As a result, the South Korean firm Hynix, initially the only manufacturer of HBM, has grown into the largest DRAM manufacturer, overtaking Samsung, which led the market for many years.[11] China has also invested heavily in DRAM and flash memory production, with Chinese companies selling less advanced parts at very low (possibly below-cost) prices, forcing Samsung and Hynix out of that business.[12] Recently, these Chinese companies have been directed to move to the more advanced memory devices.[13]

Similarly, NAND flash memory—the most common type of flash memory—transitioned to a 3-D cell design in the mid 2010s and has been scaling density by scaling the number of layers in 3-D transistor stacks for the past decade. 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 memory[14] and phase-change memory,[15] are also gaining traction as an alternative to today's embedded nonvolatile technology. These technologies offer 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.

Over the Horizon

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 help to enhance performance, increase bandwidth, and improve energy efficiency, addressing the limitations of traditional semiconductor designs.

Emerging memory technologies and advanced manufacturing techniques are also critical for the industry's growth. Innovations in 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.

Three-Dimensional Heterogeneous Integration

As noted above, advanced chip designs sometimes use 3-D structures. Today, these designs are limited to a variety of niche applications, such as HBM and high-performance computing. These 3-D structures are the result of a fabrication technique known as 3-D heterogeneous integration. This is different from 2.5-D integration, where different chiplets are placed on a common substrate. Rather, true 3-D heterogeneous integration is a semiconductor manufacturing technique that involves the vertical stacking of different electronic components, such as processors and memory, with vertical interconnect between them. 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 with one type of fabrication can be stacked with memory made from another, with 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.

A variety of challenges need to be overcome for 3-D heterogenous integration to become more widely used. These include thermal management, mechanical stress and reliability, manufacturing complexity and cost, interconnect reliability, and design complexity. Many of these issues are also present with traditional 2-D and 2.5-D integration, but vertical stacking creates new failure modes that do not exist or are much less severe in traditional 2-D chips.

Photonic Links and Components

The distance that a high-performance electrical data-transmission link can span has been shrinking as its data-bandwidth has increased. Photonic (light) links are now used for longer-distance communications. Photonics is the optical analog of electronics—the latter use electrons for signaling and carrying information, while photonics use photons (light) for the same purposes. Innovations such as silicon photonics are emerging, making photonic links attractive for much shorter distances, including some chip-to-chip communications.

Silicon photonic links have the potential to reduce energy consumption and increase bandwidth in data centers and long-distance data transmissions that are not already photonic.[16] Furthermore, they can handle different wavelengths simultaneously on a single optical fiber. This enhances data-carrying capacity and makes 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. Such 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.

Applications-Specific Optimization

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* (*SETR* 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.

Policy Issues

Talent

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 total approximately 460,000.[17] 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 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.[18]

The pipeline of college graduates interested in semiconductors is also troubling. While student interest in hardware seems to be increasing, recent actions, including the voiding of up to $7.4 billion in CHIPS Act funding[19] and the cutting of government funding for research in general, will inevitably shrink the number of new graduates in this area.

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.

Strategic Technology Containment

The primary objective of actions taken under this rubric is to restrict China's access to high technology to preserve Western advantages in innovation, industry, and defense. For example, the United States has intensified export controls and revocation of export authorizations targeting Chinese semiconductor firms and key software and equipment for design and fabrication. These efforts aim to restrict China's ability to develop the most advanced semiconductors (currently 5 nm and 3 nm technologies). Restricted technologies include top-tier lithography machines, high-performance computing chips, and electronic design automation software.[20]

In December 2024, the US Department of Commerce released a set of export-control rules to further impair China's ability to produce advanced semiconductors, specifically targeting, among other things, certain semiconductor manufacturing equipment.[21] This equipment included lithography tools using extreme ultraviolet light (EUV), thus effectively blocking deliveries of ASML Holding's most advanced EUV systems to China.[22] ASML is a Dutch multinational corporation that develops and manufactures advanced photolithography machines used in semiconductor fabrication and the only company worldwide that produces EUV lithography systems for this purpose.

In addition, the US-led Clean Network program, launched in 2020 in the first Trump administration, sought to prevent Chinese telecommunications carriers and suppliers from accessing or influencing sensitive US telecommunications infrastructure and data networks.[23] Further, the Federal Communications Commission banned the US sale of certain communications equipment from Chinese technology firms such as Huawei and ZTE.[24]

Such actions have had a disruptive impact on China's semiconductor ecosystem,[25] at least in the short term. The inability of Chinese firms to access key tools for next-generation chip production has triggered supply chain delays, steep price increases, and direct setbacks, resulting in workforce reductions and postponed factory expansions. China's progress in producing state-of-the-art chips has been impeded, forcing it toward legacy technologies; its broader ambitions in AI and advanced computing have also been impacted.

In response, China has launched a comprehensive effort to achieve semiconductor self-sufficiency. This includes massive subsidies, accelerated investment, and reforms supporting indigenous chip design and manufacturing innovation. It has further supported research in the field (e.g., producing twice as many chip design papers than the United States) and made advances in materials like 2-D transistors and carbon nanotube chips. Finally, it has undertaken a variety of efforts to circumvent Western containment measures, including the use of smuggling and shell companies to purchase equipment and chips.

Thus, while Western technology containment efforts have effectively slowed immediate Chinese advances, they may have the unintended impact of decreasing Chinese dependence on Western technology.

To further discourage technology containment efforts, China also retains the option to retaliate against

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

nations pursuing them. For example, in December 2024, the country announced a total export ban to the United States on strategic critical minerals including gallium, germanium, and antimony, as well as industrial-strength diamonds and dense synthetic materials, citing national security concerns and responding directly to new US restrictions on semiconductor exports to China.[26] Such efforts are likely to continue into the future.

Geopolitical Risks and Supply Chain Resilience

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 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.

Investing in domestic manufacturing capabilities and promoting regional cooperation is intended to enhance supply chain security and ensure a steady supply of essential components. For example, TSMC is building facilities in Arizona and Japan, and Samsung is investing in Texas. Japan and India are both investing heavily to modernize and grow their own chipmaking industries, while Southeast Asian countries are focusing on assembly and testing.

Industrial Policy

The US government acquired a 10 percent equity stake in Intel by converting previously awarded but unused government grants proffered under the CHIPS Act into shares.[27] This move aims to support Intel's expansion of domestic chip manufacturing. The government's ownership is passive, with no board seat or governance rights, and includes a warrant for an additional 5 percent if Intel loses majority control of its foundry business.

This deal constitutes an unusual direct investment of the US government in a major private company, possibly signaling a move toward more active government involvement in strategic industries, and a number of analysts have raised concerns about it.[28] They suggest it raises the chance that Intel, or any company in a similar arrangement, would shape its own corporate decision making to align with government or political preferences, compromising its market-driven business priorities. Additionally, it risks distorting competition in markets that would otherwise be free of government stakes, creating potential conflicts between economic efficiency and political objectives. For example, it might lead to an undue bias in favor of meeting national security objectives over maintaining a competitive, innovation-driven product line.

NOTES

[1] "Inside an Intel Chip Fab: One of the Cleanest Conference Rooms on Earth," Intel Corporation, March 28, 2018, https://www.intc.com/news-events/press-releases/detail/165/inside-an-intel-chip-fab-one-of-the-cleanest-conference.

[2] Anton Shilov, "Former Intel Directors Believe Intel Must Split in Two to Survive," *Tom's Hardware*, October 26, 2024, https://www.tomshardware.com/tech-industry/former-intel-directors-believe-intel-must-split-in-two-to-survive; "Tech Analyst Explains Why Intel Should Be Split Apart and How to Do It," Investing.com, Yahoo Finance, March 5, 2025, https://finance.yahoo.com/news/tech-analyst-explains-why-intel-132943025.html.

[3] 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.

[4] "Global Semiconductor Foundry Revenue Share: Q1 2024," Counterpoint, June 12, 2024, https://www.counterpointresearch.com/insights/global-semiconductor-foundry-market-share/.

[5] "Global Semiconductor Foundry Revenue Share," Counterpoint.

[6] *2021 State of the U.S. Semiconductor Industry*, Semiconductor Industry Association, 2021, https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf.

7. “2019 Summit Agenda, Videos, and Slides,” Electronics Resurgence Initiative 2.0, Defense Advanced Research Projects Agency, accessed August 30, 2023, https://eri-summit.darpa.mil/2019-archive-keynote-slides.

8. 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.

9. Dion Harris, “Sustainable Strides: How AI and Accelerated Computing Are Driving Energy Efficiency,” *NVIDIA Blog*, July 22, 2024, https://blogs.nvidia.com/blog/accelerated-ai-energy-efficiency/.

10. See, for example, Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, and Kurt Keutzer, “AI and Memory Wall,” *IEEE Micro* 44, no. 3 (2024): 33–39, https://doi.org/10.1109/MM.2024.3373763.

11. Mark LaPedus, “SK Hynix Surpasses Samsung in DRAM Share,” *Semiecosystem*, Substack, April 11, 2025, https://marklapedus.substack.com/p/sk-hynix-surpasses-samsung-in-dram.

12. Nikkei Asia, “China Makes Inroads in DRAM Chips in Challenge to Samsung and Micron,” KrASIA, February 1, 2025, https://kr-asia.com/china-makes-inroads-in-dram-chips-in-challenge-to-samsung-and-micron.

13. Ray Wang, “Mapping China’s HBM Advances,” November 27, 2024, https://www.chinatalk.media/p/mapping-chinas-hbm-advancement.

14. 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* (Woodhead Publishing, 2022), 159–215, https://doi.org/10.1016/B978-0-12-820758-1.00007-8.

15. Cappaletti and Slaughter, “Embedded Memory Solutions.”

16. David A. B. Miller, “Attojoule Optoelectronics for Low-Energy Information Processing and Communications,” *Journal of Lightwave Technology* 35, no. 4 (February 2017): 34696, https://doi.org/10.1109/JLT.2017.2647779.

17. 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/.

18. Martin and Rosso, “Chipping Away.”

19. David Shepardson, “US Commerce Voids Biden’s $7.4 Billion Semiconductor Research Grant Deal,” Reuters, August 25, 2025, https://www.reuters.com/legal/government/us-commerce-voids-bidens-74-billion-semiconductor-research-grant-deal-2025-08-25/.

20. “Commerce Strengthens Export Controls to Restrict China’s Capability to Produce Advanced Semiconductors for Military Applications,” news release, Bureau of Industry and Security, US Department of Commerce, December 2, 2024, https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced-semiconductors-military.

21. “Commerce Strengthens Export Controls,” Bureau of Industry and Security.

22. “ASML Expects Impact of Updated Export Restrictions to Fall Within Outlook for 2025,” news release, ASML, December 2, 2024, https://www.asml.com/en/news/press-releases/2024/asml-expects-impact-of-updated-export-restrictions-to-fall-within-outlook-for-2025.

23. Meg Rithmire and Courtney Han, “The Clean Network and the Future of Global Technology Competition,” *Harvard Business Review*, April 2021, https://store.hbr.org/product/the-clean-network-and-the-future-of-global-technology-competition/721045.

24. Associated Press, “U.S. Bans the Sale and Import of Some Tech from Chinese Companies Huawei and ZTE,” NPR, November 26, 2022, https://www.npr.org/2022/11/26/1139258274/us-ban-tech-china-huawei-zte.

25. Sujai Shivakumar, Charles Wessner, and Thomas Howell, “Export Controls on U.S. Chip Technology to China,” Center for Strategic and International Studies, February 2024.

26. Krystal Bermudez, “China Retaliates Against U.S. Semiconductor Restrictions by Banning Critical Mineral Exports,” Foundation for the Defense of Democracies, December 4, 2024, https://www.fdd.org/analysis/policy_briefs/2024/12/04/china-retaliates-against-u-s-semiconductor-restrictions-by-banning-critical-mineral-exports/.

27. “Intel and Trump Administration Reach Historic Agreement to Accelerate American Technology and Manufacturing Leadership,” news release, Intel Corporation, August 22, 2025, https://www.intc.com/news-events/press-releases/detail/1748/intel-and-trump-administration-reach-historic-agreement-to.

28. See, for example, Ross Kerber, “Investors Worry Trump’s Intel Deal Kicks Off Era of US Industrial Policy,” Reuters, August 27, 2025, https://www.reuters.com/sustainability/boards-policy-regulation/investors-worry-trumps-intel-deal-kicks-off-era-us-industrial-policy-2025-08-27/; Alexander Bolton, “Donald Trump’s Government Buying Stake in Intel Prompts GOP Complaints,” *The Hill*, August 26, 2025, https://thehill.com/homenews/senate/5469740-gop-criticizes-trump-intel-deal/; David Shepardson and Arsheeya Bajwa, “Intel Warns US Stake Could Hurt International Sales, Future Grants,” China, Reuters, August 25, 2025, https://www.reuters.com/world/china/intel-warns-us-stake-could-hurt-international-sales-future-grants-2025-08-25/.

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. Asir Intisar Khan
SETR Fellow and Visiting Postdoctoral Scholar in Electrical Engineering

This publication reflects updates through December 2025

32 31 30 29 28 27 26 7 6 5 4 3 2 1

Designer: Howie Severson
Typesetter: Maureen Forys

Image credits: Linda A. Cicero/Stanford News and iStock.com/PTC-KICKCAT92 (cover); iStock.com/mofuku (p. 22); iStock.com/wacomka (p. 38); iStock.com/FeelPic (p. 56); iStock.com/JONGHO SHIN (p. 70); iStock.com/Chartchai San-saneeyashewin (p. 88); iStock.com/ArtemisDiana (p. 102); iStock.com/PhonlamaiPhoto (p. 116); iStock.com/imaginima (p. 142); iStock.com/Floriana (p. 156); iStock.com/dima_zel (p. 170); Tim Griffith (p. 225)