SEPTEMBER 2026

Beyond the Memory Cycle

AI, HBM, and the New Semiconductor Shortage

By Sujai Shivakumar, Charles Wessner, and Thomas Howell

Introduction

Semiconductor memory devices are essential components of virtually every modern software-driven electronic system, including computers, telecommunications equipment, AI and cloud data centers, consumer electronics, automobiles, aircraft, industrial automation systems, medical devices, and military platforms. As a foundational enabling technology, memory chips are indispensable to the functioning of the modern digital economy. The industry is now experiencing a global shortage that is expected to persist through 2027 and, by some industry estimates, into 2028, when significant new manufacturing capacity is projected to come online. This shortage is not simply another turn in the memory industry's familiar boom-and-bust cycle. It reflects a structural AI-driven demand shock layered onto a market still governed by extreme fixed costs, long construction lead times, and recurring incentives to overbuild once prices rise.

Memory markets have historically been characterized by pronounced cycles of shortage and overcapacity. The current shortage, however, reflects a structural increase in AI-driven demand alongside those familiar cyclical dynamics. The rapid and largely unforeseen expansion of AI data centers has driven an unprecedented surge in demand for high-bandwidth memory (HBM), a specialized memory technology **critical** for AI workloads. HBM is built from dynamic random-access memory (DRAM), the main form of high-speed volatile memory used by processors as temporary working memory in computers, servers, smartphones, and many other digital systems. In response, the **dominant** global memory manufacturers, or the Big Three—Samsung, SK Hynix, and Micron—have each made independent capacity and investment decisions to prioritize HBM and other AI-optimized memory products. This shift may change the structure of the memory industry by allowing leading producers to compete through differentiated technologies, advanced packaging, and long-term relationships with AI customers rather than primarily through price and scale in commodity DRAM. Yet the conditions that have historically made memory markets unstable—high fixed costs, long construction lead times, and incentives to operate fabrication facilities, or fabs, at high utilization—

remain. The current shortage therefore reflects both a structural shift toward AI-oriented memory production and the continuing dynamics of an unusually cyclical industry.

From the perspective of the United States and its allies, these developments raise concerns not only about aggregate supply but also about the geography and trustworthiness of future capacity. As the Big Three devote more resources to HBM and other AI-oriented products, tighter supplies of conventional DRAM and NAND flash memory could create openings for other producers, especially in China. Additional Chinese capacity could ease some global shortages in conventional memory, but it may also deepen dependence on suppliers that present trade, security, and policy challenges for the United States and its partners.

> *The current shortage therefore reflects both a structural shift toward AI-oriented memory production and the continuing dynamics of an unusually cyclical industry.*

The current shortage has become a supply chain security concern rather than simply a pricing problem because of where HBM manufacturing is located. While U.S. investments are expanding domestic DRAM fabrication and advanced packaging capabilities, commercial-scale HBM production remains **concentrated** in Asia, reflecting the colocation of advanced DRAM fabrication, stacking, packaging, and customer qualification ecosystems. To understand why these dynamics matter, it is useful first to distinguish among the major types of memory devices and the production steps that make HBM uniquely difficult to scale.

Memory Chip Taxonomy

Memory devices are known by a bewildering array of acronyms, but collectively they represent two basic types:

Volatile memory devices, which include random-access memory (RAM), provide working memory capability giving logic processors “random” rapid access to the data and instructions they are actively using. In effect, these devices provide temporary workspace, losing their information when electric power is removed. The most consequential applications of RAM devices are **found** in smartphones and other mobile devices; laptops, PCs, tablets, and workstations; automotive, industrial, and medical equipment; and consumer electronics such as TV systems and gaming consoles.

Within this category, RAM device types **include** “workhorse” dynamic RAM (DRAM) and other variants. Among these, HBM devices, while based on DRAM technology, differ fundamentally from conventional DRAM in that they feature multiple ultra-thin DRAM dies stacked vertically, up to 16 layers deep, and connected by thousands of copper through-silicon vias, a system that is then connected to a processor chip by a silicon interposer.

As a result, HBM is significantly more difficult to engineer and manufacture than conventional DRAM. It offers far higher bandwidth and greater energy efficiency and is in high **demand** for AI applications despite its significantly higher cost. Nvidia’s H100 GPU, one of the foundational pieces of hardware underpinning the generative AI boom, requires HBM to function at the performance levels **demanded** by large language model training and inference workloads. Importantly, HBM chips are essentially packages of stacked DRAM dies. DRAM fabs produce thin dies, but an HBM device exists only after advanced packaging. Therefore, the ability to produce large quantities of DRAM dies in a fab cannot be **translated** into HBM output without either internal or outsourced advanced packaging capability. Current efforts to expand HBM capacity thus feature the construction of additional advanced packaging facilities.

Nonvolatile memory devices retain their stored data when electric power is turned off, enabling tasks that require retention of information across power cycles and shutdowns. They can store vast amounts of data at low cost. These devices include NAND flash and NOR flash memories, the former being **used** for mass data storage and the latter for direct, rapid execution of code. Significant applications of nonvolatile devices are found in all systems and devices that require data storage, including cloud and AI data centers, USB drives, computers, smartphones, factory automation systems, motor vehicles, and a wide **range** of consumer electronic devices.

> *The same firms and facilities that produce conventional DRAM also provide the technological foundation for HBM, meaning that shifts in product mix can ripple across multiple end markets.*

These distinctions matter because the present shortage is not affecting all memory products in the same way. The same firms and facilities that produce conventional DRAM also provide the technological foundation for HBM, meaning that shifts in product mix can ripple across multiple end markets. The next question, therefore, is which firms control that production base and how their investment decisions shape supply across the memory market.

Concentration of Global Memory Production

The global semiconductor memory industry has evolved into a market dominated by the Big Three integrated device manufacturers, which internally design, fabricate, test, and package their own devices rather than outsourcing some process steps to foundries. The market research firm TrendForce **estimates** that Micron, Samsung, and SK Hynix accounted for 90.5 percent of global DRAM sales in the fourth quarter of 2025.

Samsung is the world's **largest** producer of DRAM and NAND memory, SK Hynix ranks second, and Micron, which more recently **expanded** into HBM, is the only remaining major producer headquartered in the United States. This concentration reflects decades of consolidation driven by the industry's extreme capital requirements and repeated downturns, which eliminated more than a dozen competitors. Large, well-capitalized firms were able to maintain investment and expand capacity during these downturns, while less-capitalized producers exited or were acquired. Samsung, for example, **drew** revenues from its diversified electronics businesses in the 1980s to invest countercyclically and gain market share, while Micron **maintained** and expanded its position through cost discipline, **investment** in process technology, and acquisitions, including its 2013 **acquisition** of the bankrupt

Japanese producer Elpida Memory. SK Hynix similarly emerged from **consolidation** in the Korean memory industry, including the 1999 merger of Hyundai Semiconductor and LG Semiconductor and SK Telecom's acquisition of Hynix in 2012. The resulting market structure illustrates how sustained **access** to capital and manufacturing scale have reinforced the dominance of a small number of producers.

The emergence of HBM has introduced a new competitive dimension without substantially reducing this concentration, shifting the basis of competition toward advanced memory design, stacking, and packaging. SK Hynix's early investment in HBM production and related technologies beginning in 2017 positioned it to supply Nvidia and other customers as demand for AI accelerators increased; it now leads the HBM market and in 2025 **surpassed** Samsung in annual operating profit for the first time. Samsung currently accounts for about 17 percent of global HBM sales, while Micron **holds** about 21 percent and competes partly through improvements in power efficiency. Although HBM has begun to reorder the competitive hierarchy among the three dominant firms, it has also increased the strategic significance of their concentrated control over advanced memory capacity. Micron's position as the sole major producer headquartered in the United States is therefore central to efforts to expand domestic DRAM and advanced memory manufacturing, while the broader market remains heavily dependent on the other major producers headquartered primarily in South Korea.

Other significant memory producers occupy narrower positions in DRAM, NAND, and specialty memory markets and could gain opportunities as the Big Three devote greater resources to HBM and other AI-oriented products. The most consequential emerging competitors are in China. Yangtze Memory Technologies (YMTC), based in Wuhan, focuses primarily on 3D NAND and **reportedly accounted** for around 13 percent of global NAND shipments in the third quarter of 2025. ChangXin Memory Technologies (CXMT), based in Hefei, produces commodity DRAM for computers, mobile devices, and consumer electronics and held an **estimated** 8 percent of the global DRAM market in mid-2025, ranking fourth behind the Big Three. Although neither company currently matches the scale and technological position of the leading producers across the memory market, their expansion could become increasingly important if the Big Three's shift toward HBM leaves additional demand for conventional DRAM and NAND to other suppliers.

Other established Asian producers occupy more specialized positions. Kioxia, **formerly** Toshiba Memory, is a major NAND producer, operating two joint-venture fabrication facilities with Western Digital in Japan. The company sees **opportunities** to supply high-density, solid-state storage for AI data centers as the Big Three devote greater investment to HBM. Taiwan's Nanya Technology **produces** DRAM and has been selected as a memory supply chain partner for Nvidia's Vera Rubin platform, while Winbond **specializes** in custom NAND, NOR flash, and specialty DRAM for industrial, automotive, communications, and consumer applications. These firms provide additional sources of memory supply, but their differing product specializations also underscore the limits of treating memory capacity as interchangeable across DRAM, NAND, HBM, and specialty devices.

> *The current shortage is, therefore, partly a consequence of industry structure: Concentrated supply amplifies the effects of a common shift in producer incentives.*

The significance of this concentration of production is not merely that three firms dominate global supply. It means that a handful of independent investment and product-allocation decisions can reshape availability across the entire memory market. When Samsung, SK Hynix, and Micron simultaneously find HBM more attractive than conventional DRAM, their choices collectively redirect a substantial share of the industry's capital, engineering resources, fabrication capacity, and advanced packaging capability toward AI customers. The current shortage is, therefore, partly a consequence of industry structure: Concentrated supply amplifies the effects of a common shift in producer incentives. Yet concentration alone does not make the market stable. To understand why the shortage may eventually give way to another glut, it is necessary to examine the memory industry's recurring cycle of underinvestment, price spikes, overbuilding, and collapse.

Relentless Cyclicality

The global semiconductor memory industry has long **followed** a brutal and predictable cycle: Prices rise, manufacturers overinvest, supply floods the market, prices collapse, and companies absorb significant losses before the cycle begins again. As a result, memory has remained one of the most reliably cyclical businesses in the technology sector. Only a handful of companies have managed to survive these cycles, underscoring the **risks** of investing in new capacity during periods of high prices. Manufacturers that expand aggressively when prices are elevated face the possibility that prices will collapse before new capacity begins generating significant returns.

Supercycles like this occur most readily in commodity markets because supply and demand cannot adjust smoothly. When demand rises suddenly, prices spike because new mines, wells, fabs, or processing capacity take years and enormous capital investment to build. High prices then trigger many producers to expand at once, but by the time new capacity arrives, demand may have cooled or inventories may have accumulated. Since commodity products are largely interchangeable, once supply exceeds demand, producers compete mainly on price, turning the boom into a bust. Markets for differentiated products usually have more room to compete through branding, features, software, or customer relationships; commodity markets have less margin for error.

> *Manufacturers that expand aggressively when prices are elevated face the possibility that prices will collapse before new capacity begins generating significant returns.*

An example of what can happen to memory chipmakers occurred in the years following the financial crisis of 2007-2008, when a market glut led to a cataclysmic and largely unforeseen collapse in DRAM prices. At the beginning of 2006, industry analysts **forecasted** that memory chip revenues would increase in 2007 and 2008, with market growth rates of 5 to 7 percent. STMicro's chief economist **projected** growth of 6.5 percent in 2007 and 8.7 percent in 2008. Even at the beginning of 2008, after the growth forecasted for 2007 had failed to materialize, an analyst at the consultancy firm iSupply **forecasted** modest market growth in 2008. Instead, DRAM spot prices fell nearly 50 percent **between** January and December 2008. DRAM vendors collectively **lost** $7 billion that year. Germany-based Qimonda (formerly Infineon), Europe's main **producer** of DRAMs, went bankrupt in 2009. Spanshion, a U.S. maker of flash memories, **filed** for bankruptcy. Prices continued to fall; the price of a **benchmark** chip, the DDR3 2GB DRAM, declined by about 85 percent in 2011. The world's third-largest DRAM producer, Japan's Elpida Memory—formed from the merged DRAM operations of NEC, Mitsubishi, and Hitachi—went bankrupt in 2012. Its president, Yukio Sakamoto, **commented** ruefully that "the price of one DRAM now hardly buys you one rice ball." Taiwan's proMOS Technologies effectively ceased operations as a DRAM producer in 2012 through a court-ordered liquidation process. The company was **dismantled** by its creditors, and its assets were sold off.

This history matters because the present shortage is unfolding in an industry whose response to scarcity has often been self-defeating. High prices encourage producers to invest aggressively, but the long lag between investment and usable capacity means that new supply often arrives after demand conditions have changed.

FROM COMMODITY SUPERCYCLES TO TECHNOLOGICAL DIFFERENTIATION

SK Hynix, currently the market leader in HBM, has invested heavily in a new packaging technology called mass reflow underfill molding, which has **given** it an edge in the chip stacking required to create HBM chips—reportedly overcoming the thermal and physical challenges that characterize the industry's previous standard technology, thermal compression with nonconductive film.

Micron is **reportedly** seeking to escape the memory sector's notorious boom-and-bust cycle through large-scale investments in HBM and the development of proprietary technologies that differentiate its devices from those of its competitors. It has **developed** hybrid bonding techniques for stacked chips that reduce thermal resistance at each connection point, reportedly achieving a 30 percent reduction in power consumption relative to competing designs.

Samsung Executive Vice President Sang Joon Hwang **said in February 2026** that "instead of taking the conventional path of using existing conventional designs, Samsung took the leap and adopted the most advanced nodes like the 1c DRAM and 4nm logic process for HBM4." He **continued**: "By leveraging our process competitiveness and design optimization, we are able to secure substantial performance headroom, enabling us to satisfy our customers' escalating demands for higher performance."

> *HBM creates greater opportunities for differentiation through stacking, packaging, power efficiency, process technology, and customer qualification.*

These strategies suggest that HBM may be changing the basis of competition in memory. Conventional DRAM has historically behaved largely as a commodity, exposing producers to severe price competition when capacity exceeds demand. HBM creates greater opportunities for differentiation through stacking, packaging, power efficiency, process technology, and customer qualification. The Big Three are therefore attempting not merely to capture an unusually profitable segment of the current market, but to reduce their exposure to the commodity dynamics that have historically made memory so cyclical. That strategic shift helps explain why the present shortage is different from earlier cycles, but it does not eliminate the underlying economics that make memory markets unstable.

An Atypical Oligopoly

In economic parlance, an industry in which three firms account for over 90 percent of sales is an oligopoly, in which producers can be expected to coordinate sales and investments in an attempt to **maintain** market stability. The Big Three were sanctioned for such activity between 1998 and 2002. At present, all three are collectively said to be urging buyers of memory chips to refrain from hoarding and double-ordering, practices which will **worsen** the current shortage in the near term, exacerbate price volatility, and virtually guarantee a nasty market collapse when the shortage abates and stockpiles are unloaded.

However, more often, the Big Three do not behave as a classic oligopoly. In part this reflects the fact that the market is characterized by extreme demand elasticity and equally extreme supply inelasticity, making it intrinsically unstable. In addition, all three firms have unique corporate cultures and a fiercely competitive élan that have enabled their survival and growth but **undermined** prospects for wider market stability. In a 2019 **oral history**, Micron CEO Sanjay Mehrotra emphasized the role of the company's corporate culture in enabling its survival, noting that the team, culture, family spirit, and total commitment fostered in Boise, Idaho, were essential to keeping the company together through periods of DRAM industry consolidation and boom-and-bust cycles. He **suggested** that this cohesion might have been less likely in Silicon Valley, where employees could have more easily "jump[ed] ship" for other opportunities."

The economic imperatives underlying this dynamic are painfully evident. Building a new memory fabrication plant **costs** approximately $15-20 billion and takes two to three years or more to complete, even under the best of circumstances. As a result, when demand spikes unexpectedly, prices rise because new capacity cannot be brought online quickly enough. Once new facilities become operational, however, they are most economical when run at full utilization rates. Consequently, even as prices begin to fall, producers typically continue operating at maximum output to minimize unit costs. Historically, when several manufacturers have brought substantial new capacity online at roughly the same time, the result has been oversupply, as firms simultaneously maximize production, flooding the market with memory chips and triggering sharp price declines. These downturns have frequently forced weaker competitors out of the market. Over the decades, this cycle has progressively consolidated the industry, reducing the number of major memory chipmakers from more than a dozen to today's Big Three.

The firms that survived and achieved leading market positions did so not through oligopolistic restraint but through large, high-risk investments in research and development (R&D) and capacity expansion, as well as periodic episodes of aggressive pricing.

- **Fierce intra-Korean competition:** Samsung has traditionally **pursued** a countercyclical investment policy, making massive investments in new capacity during market downturns. Samsung has been able to draw upon revenues **generated** by its diversified electronics businesses to pursue this strategy. Samsung and SK Hynix are long-standing rivals in what was originally a zero-sum, David-and-Goliath competition between the Samsung and Hyundai chaebols. In the latest chapter of this rivalry, SK Hynix's early and large investments in HBM and novel packaging technology have **enabled** it to seize HBM market leadership and leapfrog Samsung in profitability in 2025. Reflecting this success, SK Hynix **surpassed** Samsung in 2025 as the most desired employer among South Korean university students.

- **Massive capital investments:** Micron Technology is pursuing an aggressive capital investment program that could further destabilize the oligopoly. The program includes $200 billion in investments in as many as six new state-of-the-art fabs in the United States—in Boise, Idaho, and Clay, New York—as well as an **expansion** of an existing facility in Manassas, Virginia; a $24 billion fab in Singapore slated to start production in 2028; and the acquisition of a 12-inch DRAM fab from Taiwan's Powerchip for $1.8 billion. Of Micron's U.S. investments, $50 billion is being directed toward R&D, signifying a **determination** to differentiate its products from those of commodity memory chipmakers.
- **Chinese entrants:** Chinese memory chip firms CXMT and YMTC are **launching** their most aggressive expansion efforts to date. CXMT is building new fabs in Shanghai with capacity two to three times that of its flagship facility in Hefei. The new fabs will **produce** conventional DRAM for servers, motor vehicles, consumer electronics, and computers, while the company is also expanding production lines for HBM in Shanghai. Some observers warn that Chinese capacity expansion may lead to overcapacity, dumping, and trade friction, while others conclude that the risk is overblown and that the current buildup is geared toward **satisfying** burgeoning domestic demand.

> *The central question, then, is not whether AI has ended the memory cycle but whether it has redirected that cycle around a more strategically important bottleneck. The current shortage shows how quickly that redirection can occur.*

The result is a market that is concentrated but still intensely competitive, prone to both strategic overinvestment and sudden shortages. HBM gives producers stronger incentives to differentiate products and secure long-term, high-margin customers, potentially reducing their dependence on commodity DRAM. At the same time, the industry's basic supply economics have not changed; capacity remains expensive and slow to build, and today's scarcity is prompting another enormous wave of investment. The central question, then, is not whether AI has ended the memory cycle but whether it has redirected that cycle around a more strategically important bottleneck. The current shortage shows how quickly that redirection can occur.

Onset of the Current Shortage

The current shortage became evident in the latter part of 2024 and accelerated dramatically thereafter. Its significance lies in the speed with which the industry moved from oversupply to scarcity. As recently as early 2024, memory manufacturers were still **burdened** by excess inventories and collapsing prices. DRAM suppliers entered 2024 holding roughly 31 weeks of inventory, while mobile DRAM was **reportedly** being sold at almost any price as inventories continued to accumulate. The downturn was equally pronounced in NAND flash memory, where prices fell 39 percent between 2022 and 2023, **representing** a $25 billion year-over-year decline that persisted into early 2024. At the same time, the rapid adoption of generative AI was fundamentally reshaping demand. Following the success of ChatGPT, hyperscale cloud providers—including Amazon, Google, Microsoft, and Meta—and AI companies such as OpenAI and Anthropic began **procuring** increasingly large AI clusters requiring vast quantities of HBM and DRAM. By 2024, AI infrastructure spending had accelerated to an unprecedented scale. Meta, Microsoft, Amazon, and Alphabet collectively invested $217 billion in AI-capable data centers in 2024, rising to $380 billion in 2025 and projected to reach $650 billion in 2026—investment levels **described** as "rivaling the costliest human endeavors in history."

The shift from oversupply to scarcity became even more pronounced in October 2025, when OpenAI reportedly **signed** agreements to purchase up to 900,000 DRAM wafers per month for its planned $500 billion Stargate AI data center project. These agreements reportedly secured approximately 40 percent of global DRAM output, triggering stockpiling and intensifying shortages across the broader technology supply chain. In response, the Big Three memory manufacturers—Samsung, SK Hynix, and Micron—each made independent strategic decisions to rebalance manufacturing capacity away from conventional DRAM and NAND products and toward HBM and DDR5, the advanced memory technologies required for AI accelerators and data center servers.

These AI-optimized products command significantly higher margins because customer demand is exceptionally strong and qualification requirements are demanding. In the three years since the launch of ChatGPT, the major memory suppliers have therefore shifted manufacturing, research, and capital investment toward HBM and related server-memory products. That shift has tightened supply for conventional DRAM because HBM draws on many of the same fabrication assets while also requiring additional advanced packaging capacity.

Market dynamics mean that HBM used for AI chips by companies such as Nvidia and Advanced Micro Devices affects production capacity for commodity DRAM used in smartphones, PCs, automobiles, and other consumer electronics. As the AI supercycle continues, data centers are projected to consume approximately 70 percent of all memory devices produced worldwide in 2026, underscoring how AI is reshaping both the composition of memory production and the allocation of industry investment.

The surge in AI investment is reshaping the memory market in ways that extend well beyond a traditional supply cycle. This shift is **driven** by both technology and economics. HBM is manufactured using the same fabrication facilities and silicon wafers as conventional DRAM, meaning that producers typically reallocate existing cleanroom capacity rather than build new fabs. As a result, every increase in HBM production reduces the capacity available for commodity memory, and producing HBM is substantially more resource-intensive, requiring roughly four times the cleanroom **capacity** per gigabyte compared with standard DRAM. At the same time, demand has become increasingly concentrated among a handful of hyperscale cloud providers—including Meta, Google, Microsoft, and Amazon—which negotiate long-term supply agreements at premium prices, providing manufacturers with stable, high-margin demand while leaving consumer electronics manufacturers **exposed** to greater price volatility and supply uncertainty.

As semiconductor analyst Claus Aasholm **notes**, AI demand is highly price inelastic, creating powerful incentives for memory manufacturers to maximize HBM production. Industry leaders **increasingly** believe this represents a structural rather than cyclical transformation. In February 2026, Lam Research

CEO Tim Archer argued that AI-driven memory demand would "**overwhelm** all other sources of demand" through the end of the decade.

> *AI demand is highly price inelastic, creating powerful incentives for memory manufacturers to maximize HBM production. Industry leaders increasingly believe this represents a structural rather than cyclical transformation.*

Other analysts have concluded that the industry is experiencing not merely a temporary shortage but a strategic and potentially permanent reallocation of global silicon wafer **capacity** toward AI infrastructure. That reallocation explains why the shortage is already spreading from AI data centers into consumer, industrial, infrastructure, and automotive markets.

The supply-side reallocation described above is already visible in end markets. Products that depend on relatively mature and inexpensive memory are now exposed to the same capacity constraints created by high-margin AI demand.

Impact of the Memory Chip Shortage to Date

Given the ubiquity of memory chips in digital systems, the shortage is already affecting the global economy, impeding production across a wide array of products and raising costs. A recent CSIS *Critical Questions* describes this broader consumer-facing shortage as "RAMageddon," noting that AI data center demand for HBM is **tightening** supplies of conventional RAM used in cell phones, laptops, gaming consoles, and cars. Prices have increased dramatically. According to one source, Double Data Rate 4 (DDR4)—the fourth generation of synchronous DRAM, widely **used** in PCs, servers, and consumer electronics to provide high-speed, temporary data storage for processors—has increased in price by 700–800 percent since early 2025.

- **Smartphones:** One major impact is in the smartphone market. Large smartphone vendors are warning that the memory shortage **directly** affects how many phones they can make, particularly in the mid and low ranges. According to one recent estimate, the average selling price of smartphones will rise 14 percent in 2026, **reaching** an all-time high of $523. Manufacturers will no longer be able to make phones that cost under $100, and some phone producers are expected to go out of business.
- **Personal computers:** PC suppliers are implementing substantial price increases to pass through the higher cost of memory chips. According to **one source**, the price of systems configured with 128 gigabytes of memory has recently increased by $575 to $765. The price of AI laptops equipped with an Nvidia RTX PRO 500 Blackwell GPU configured with 24 gigabytes of memory has **increased** by $530.
- **Internet infrastructure:** Memory chips are also essential to the infrastructure of the internet, with routers, cable modems, Wi-Fi gateways, and other systems **heavily** reliant on DDR4 and

other types of memory devices. The shortage is placing pressure on providers with respect to equipment procurement and deployment timelines.

- **Automotive:** Modern automobiles contain up to 3,000 semiconductors, including memory chips that support advanced driver assistance, cockpit, and infotainment systems. DRAM prices for new automotive contracts are **forecast** to be 70-100 percent higher in 2026, with the impact felt mainly in high-end vehicles. Tesla has also warned that the DRAM shortage will affect its production in 2026, with CEO Elon Musk **warning** that the company must "hit the chip wall or make a fab." Musk said in January 2026 that he will build "TeraFab," a chip foundry **integrating** logic, memory, and packaging with an eventual output of 100 to 200 billion chips per year.
- **Factory automation:** Memory chips are critically important in factory automation systems, with RAM enabling human-machine interface systems, programmable logic controllers, supervisory control and data acquisition applications, and real-time utilization of data from sensors, cameras, and communications interfaces. Industry analysts forecast that the current chip shortage will **result** in "higher procurement costs, longer delivery times, and increased uncertainty in project planning."
- **Consumer electronics:** Products such as gaming systems, smart TVs, and **connected** home devices are reliant on memory devices. In this industry, "margins are already razor-thin. Smaller manufacturers will **likely** have no choice but to raise prices, potentially denting demand." Because of the memory chip **shortage**, "household items like televisions, Bluetooth speakers, set-top boxes and even 'smart' appliances like fridges could become extremely pricey."

These sector-level effects underscore why additional supply is essential, but they also illustrate the limits of near-term relief. Even when new fabs and packaging lines are announced, memory capacity cannot be brought online quickly, and much of the expansion now underway is aimed at AI-optimized products rather than conventional memory.

Limited Additions to Memory Capacity over the Near Term (2026–27)

All three leading memory chip firms are making massive investments in new capacity, but that capacity will arrive slowly and unevenly. New DRAM fabrication facilities typically take two to three years or more to build, qualify, and ramp. Expansion also faces constraints that include shortages of skilled production workers, regulatory approvals, supply chain uncertainty, and legal restrictions on the **sale** of chipmaking equipment to China-adjacent producers. As a result, the capacity response should be understood in three stages: limited near-term additions in 2026, more meaningful expansion in 2027, and a larger post-2028 wave that could either relieve scarcity or recreate the conditions for another glut.

> *Collectively, these investments will expand global memory production, but because most new capacity is targeted toward AI-optimized memory products, they are unlikely to fully alleviate shortages in conventional memory markets.*

- **New capacity:** New memory capacity has been coming online in 2026, but it is unlikely to alleviate the increasing supply constraints. Despite several major expansion projects, industry forecasts **suggest** that memory manufacturers will only meet around 60 percent of global demand in calendar year 2026, reflecting the extraordinary pace of AI-driven consumption.
  - In February 2026, Micron began **operations** at its $2.75 billion DRAM and NAND assembly and test facility in Gujarat, India.
  - Samsung is expected to begin mass production at its new P4 DRAM line in Pyeongtaek, South Korea, with a reported capacity of 80,000 300 mm wafers per month, although much of this output will be **dedicated** to HBM4 and advanced DDR memory rather than conventional DRAM.
  - SK Hynix is also bringing its M15X HBM facility in Cheongju, South Korea, online during the first half of 2026, initially producing 10,000 wafers per month before **expanding** several-fold in 2027.
  - In China, YMTC's Wuhan Phase 3 facility is expected to begin NAND production in late 2026 with an eventual capacity of 50,000 wafers per month, while **retaining** the option to allocate some production to DRAM in the future.
  - Meanwhile, Kioxia plans to repurpose an existing facility in Kitakami, Japan, to begin mass production of its 10th-generation NAND devices, primarily to serve growing demand from AI data centers.

Collectively, these investments will expand global memory production, but because most new capacity is targeted toward AI-optimized memory products, they are unlikely to fully **alleviate** shortages in conventional memory markets.

- **New fabs:** A significant expansion in global memory capacity is expected in 2027, led by major investments in the United States and South Korea.
  - The largest incremental increase is anticipated at Micron's new ID1 DRAM fab near Boise, Idaho, where the company will **bring** online a leading-edge, HBM-capable fabrication facility with approximately 600,000 square feet of cleanroom space. First wafer output is expected in the second half of 2027, and while Micron has not disclosed production volumes, the facility is comparable in scale to Samsung's and SK Hynix's largest fabs, suggesting the potential for hundreds of thousands of 300 mm wafers per month at full capacity.
  - SK Hynix is also advancing its $400 billion Yongin semiconductor cluster, with construction of its first fab, Y1, **scheduled** for completion in 2027 and **limited** production potentially beginning by year-end. In parallel, the company expects its new HBM packaging facility in Cheongju, South Korea, to **reach** full production by the end of 2027, further expanding its AI memory manufacturing capabilities.
  - In China, YMTC is projected to **double** production at its Phase 3 Wuhan facility from 50,000 wafers per month in 2026 to 100,000 wafers per month in 2027.

Although these investments will substantially expand global memory output, much of the new capacity is designed to support HBM and other AI-oriented memory products, meaning that supply constraints for conventional memory may persist even as overall production increases.

- **A future surge in capacity:** Beyond 2028, the memory industry is poised for another major wave of capacity expansion, assuming currently announced investment plans proceed as scheduled.
    - SK Hynix expects its Y1 fab in Yongin, South Korea, to ramp up to 150,000 wafers per month by the end of 2028, while also beginning construction of a second fab, Y2, in the same semiconductor cluster. The company is simultaneously expanding production at its M15X facility in Cheongju and refurbishing its M1 and M3 fabs to **manufacture** additional HBM and server DRAM.
    - Samsung has resumed construction of its P5 fab in Pyeongtaek, South Korea, after a six-month pause, targeting a 2028 start-up to produce HBM and sixth-generation 10 nm-class (1c) DRAM, with an eventual **capacity** of 100,000-120,000 wafers per month. The company is also **upgrading** existing production lines in Pyeongtaek to increase HBM output.
    - Micron likewise plans substantial expansion. Its first new Boise DRAM fab is expected to reach full production by the end of the decade, with an estimated capacity of 200,000 wafers per month, while a second Boise fab is **planned** with comparable capabilities. Looking further ahead, Micron's proposed $100 billion New York complex envisions four leading-edge fabs, each with approximately 600,000 square feet of cleanroom space, **coming** online in phases through 2045. Outside of the United States, Micron also expects its Taiwan facility—acquired through Powerchip—to **support** significant HBM and DRAM shipments beginning in 2028, while its new advanced wafer fab under construction in Singapore is projected to begin production the same year.

Taken together, these investments indicate that the industry's long-term expansion is overwhelmingly oriented toward AI-driven memory technologies, with much of the new capacity **dedicated** to HBM and advanced server DRAM rather than conventional consumer memory.

> *The investment needed to relieve today's shortage may eventually recreate the conditions for the industry's next glut.*

Because the capacity response is slow and increasingly concentrated in AI-oriented production, the United States faces a broader supply chain security challenge. Expanding domestic fabrication is necessary, but securing memory supply also requires advanced packaging, permitting reform, industrial-site readiness, workforce development, and coordination with allies. Moreover, this expansion creates a paradox. The investment needed to relieve today's shortage may eventually recreate the conditions for the industry's next glut. If AI demand continues growing at current rates, new capacity may be absorbed without a major correction. If demand growth slows, however, large simultaneous capacity additions by the Big Three and Chinese producers could once again produce excess supply and severe price competition. AI has changed the composition of memory demand, but it has not eliminated the industry's fundamental capacity risk. These supply chain and policy trade-offs frame the agenda for improving U.S. memory resilience.

Improving U.S. Supply Chain Security for Memory Chips

While reliable statistics are thin, it is **estimated** that as of 2023, only about 2 to 3 percent of the DRAM produced globally were made in facilities located in the United States. Most U.S.-based manufacturing of memory chips is attributable to Micron's fabs in Boise, Idaho, and Manassas, Virginia. Samsung's Austin, Texas, complex makes some embedded memory for use in systems-on-a-chip, but its DRAM and NAND products are made in Asia.

Micron fabs that are currently under **construction** or planned in Boise, Idaho, and Clay, New York, when completed and operational, will grow the U.S. share of global memory production from 2 percent to around 10 percent, according to Micron CEO Sanjay Mehrotra. These facilities could strengthen the United States' foundation for future HBM supply, but the near-term risk is not that AI firms in the United States will be cut off from HBM. The leading suppliers are based in the United States and allied countries, and major U.S. cloud and AI companies are likely to remain priority customers. The real concern is resilience: Commercial-scale HBM still depends on concentrated fabrication, stacking, packaging, yield-learning, and qualification capabilities, many of them in Asia.

SK Hynix is building a $3.9 billion packaging facility for HBM chips in West Lafayette, Indiana, and Micron is planning an advanced packaging facility in the United States at a site to be determined. Samsung is **undertaking** major investments in a new fab in Taylor, Texas, but that facility will be dedicated to logic chips.

The need to secure federal and state regulatory approvals **delayed** the groundbreaking for Micron's first fab in Clay, New York, for 19 months—from mid-2024 to January 2026. In October 2024, Congress and the Biden administration enacted the Building Chips in America Act, which exempts certain semiconductor construction projects that are receiving funding from the CHIPS and Science Act from full National Environmental Policy Act review. The purpose of the act is to **accelerate** these environmental reviews by narrowing when full reviews are required and streamlining approvals for projects that are already under construction or at risk of delay.

Paralleling federal efforts, some state governments are creating preapproved, shovel-ready industrial sites specifically dedicated to chip fabs and other advanced manufacturing operations. Ohio, North Carolina, and Pennsylvania have funded sites with utilities and basic environmental approvals, obviating the need for years of **permitting** and infrastructure buildout.

The U.S. State Department has made resolution of the memory chip shortage a core goal of its new Pax Silica, a **coalition** of 14 like-minded countries formed to coordinate activities with respect to semiconductors, AI hardware, and critical minerals (including Japan, South Korea, India, the Philippines, Singapore, and a number of European and Middle Eastern countries). The Trump administration is looking to **address** the memory chip crunch by leveraging the supply chain coalition, particularly with Asian allies such as South Korea. The United States and the Philippines are reportedly

examining whether a memory chip facility could be established as part of a 4,000-acre industrial hub being set up by the two countries on the Philippine island of Luzon.

Taken together, these measures point toward a practical policy agenda: strengthen domestic and allied capacity where it can add real supply, reduce avoidable delays in permitting and site preparation, expand advanced packaging and workforce capabilities, and avoid responses that would worsen scarcity or raise costs without improving resilience. The remaining policy question is how federal support, congressional oversight, and private investment should be calibrated to expand supply without simply subsidizing the next overcapacity cycle.

Government Support

Micron is **receiving** roughly $6.2 billion in direct CHIPS and Science Act funding for leading-edge DRAM production in Clay, New York, and Boise, Idaho, with an additional proposed award of up to $275 million to modernize and expand legacy DRAM production at its Manassas, Virginia, facility. Micron is also **eligible** for federal loans and investment tax credits, and in 2026, it raised its planned U.S. investment commitment to more than $250 billion through 2035. These investments will expand U.S.-based DRAM fabrication and may support future HBM-capable production, but HBM manufacturing today still depends on advanced stacking and packaging capabilities that are not yet available at commercial scale in the United States. New York has committed $5.5 billion in **incentives** to Micron's facility in Clay. Internationally, Micron is receiving $3.6 billion in support from Japan for R&D and capital investments at its Hiroshima memory facility, and the Indian central government and the state of Gujarat are **providing** roughly 70 percent of the funding for Micron's $2.75 billion assembly and test facility in Sanand, Gujarat.

In 2024, the government of South Korea **announced** a $19 billion program to support domestic semiconductor producers, most notably Samsung and SK Hynix. The assistance includes tax incentives and low-interest loans from government financial institutions. Samsung is also receiving $6.5 billion from the U.S. government in CHIPS and Science Act **funding** to support the construction of two logic fabs in Taylor, Texas, which will also receive $250 million in incentives from the state of Texas; however, memory fabs are not part of either agreement. SK Hynix is receiving nearly $1 billion in U.S. grants and loans to **support** its new HBM advanced packaging facility in Indiana, along with substantial state, local, utility, and university incentives tied to training, infrastructure, land, and supply chain development. Both firms will also benefit from the CHIPS and Science Act investment tax credit.

YMTC was formed in China in 2016 with a **planned** investment of $24 billion, including large stakes from the provincial government of Hubei and the National Integrated Circuit Industry Investment Fund, also called the Big Fund. In 2023, YMTC **reportedly** received a $7 billion capital infusion from the Big Fund and Hubei-backed entities.

The Japanese government is **providing** the Kioxia-Western Digital NAND joint venture with over $1.6 billion in financial support.

Emerging Public Policy Issues

The memory chip shortage has erupted so abruptly that few public policy proposals have emerged, much less secured broad support. Commerce Secretary Howard Lutnick has threatened 100 percent tariffs on major overseas memory chip producers, **notably** Samsung and SK Hynix, unless they significantly expand their U.S.-based production. Critics argue that such a **measure** could increase chip prices and worsen the global shortage. In any event, existing plans to increase U.S. production will take time, making further reductions in excessive regulatory barriers a more promising near-term path.

In March 2026, the American Enterprise Institute's Center for Technology, Science, and Energy **called** for a formal inquiry into the effects of the memory shortage on U.S. manufacturers, pointing to the National Economic Council's ability to gather industry input and congressional oversight of semiconductor policy and consumer protection. A formal **inquiry** into the effects of the memory shortage on U.S. manufacturers across the automotive, defense, medical, and consumer electronics sectors is recommended as a sensible and timely first step.

> *Financial incentives for U.S.-based production are one way to encourage additional supply, but they should be designed to ensure genuine additionality.*

Z2Data, a supply chain management platform, **suggests** establishing a national strategic stockpile of memory chips as a hedge against shortages, "giving sectors like automotive, aerospace and defense, and telecommunications a crucial buffer when protracted disruptions strike." Such a proposal would need to be narrow and carefully designed. Memory products are not interchangeable, and a poorly specified stockpile could accumulate the wrong devices for future demand.

Private companies rarely store enough microchips because they only bear part of the cost when a shortage hits the wider economy. However, a government-run backup supply is difficult to manage. Buying chips during a shortage only drives prices higher and makes the immediate crisis worse. On top of that, stored chips quickly become obsolete as newer, faster models are invented. Because of these financial and technical risks, a government stockpile only makes sense for emergency or military use–not as a broad fix for commercial chip shortages.

Financial incentives for U.S.-based production are one way to encourage additional supply, but they should be designed to ensure genuine additionality. Expanding the semiconductor investment tax credit in duration and scale should be considered for domestic production, while recognizing that memory manufacturing remains inherently global in its production sites, incentives, and interactions among competing–and sometimes complementary–industrial policies. Encouraging data center investors to invest in expanded memory supply, as some are beginning to do in energy, may also help accelerate the long-term growth of this key component of the modern economy. The conclusion draws these strands together by distinguishing between the immediate shortage, the structural shift toward AI memory, and the longer-term risk of overcapacity.

Conclusion

The memory shortage points to five broader conclusions:

1. **AI leadership now depends on memory as well as logic.** Advanced logic remains essential, but AI infrastructure also requires large and reliable supplies of HBM, conventional DRAM, and advanced packaging capacity.
2. **The shortage is structural, not merely cyclical.** AI demand has redirected scarce fabrication and packaging capacity toward HBM and other AI-optimized memory products, even as memory markets remain vulnerable to boom-and-bust dynamics.
3. **New capacity will take time and may not relieve all markets equally.** Much of the announced expansion is aimed at HBM and server memory, leaving users of conventional memory exposed to continued price pressure and supply uncertainty.
4. **Chinese overcapacity could become the next memory-market risk.** Expanding Chinese DRAM and NAND production may ease some near-term pressure, but it could also produce future surplus, dumping, trade friction, and deeper dependence on Chinese suppliers for commoditized but strategically important memory devices.
5. **Policy should focus on resilience rather than market management.** The most useful response is to reduce avoidable bottlenecks, expand domestic and allied capacity where it adds real supply, strengthen advanced packaging, accelerate site readiness, and coordinate with partners.

These conclusions suggest a practical policy stance is needed. The goal should not be to manage memory markets directly, but to reduce avoidable bottlenecks and strengthen resilience. Congress and the executive branch should begin with a focused inquiry into the shortage's effects on critical sectors; accelerate permitting and site readiness for strategically important semiconductor facilities; sustain and refine incentives for domestic memory fabrication and advanced packaging where they produce genuinely additional capacity; and coordinate with allies to expand geographically diversified supply. A strategic stockpile may have value for narrow defense or emergency applications, but it cannot substitute for expanded, flexible, and commercially viable production.

> *The goal should not be to manage memory markets directly, but to reduce avoidable bottlenecks and strengthen resilience.*

The most important longer-term risk is that today's shortage gives way to tomorrow's overcapacity–especially if Chinese producers expand rapidly in conventional DRAM and NAND. As the Big Three devote more capacity to HBM and AI-optimized memory, lower-cost Chinese suppliers may gain a larger role in serving demand for basic DRAM used in automobiles, telecommunications equipment, industrial systems, medical devices, and defense applications. If this expansion follows the industry's familiar pattern of overinvestment, it could produce another cycle of surplus capacity, dumping, and price collapse, while giving China greater influence over the supply of commoditized but strategically important memory devices. ■

Sujai Shivakumar *is the director of Renewing American Innovation at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***Charles Wessner*** *is a non-resident senior adviser with Renewing American Innovation at CSIS.* ***Thomas Howell*** *is an international trade attorney serving as a consultant to CSIS Renewing American Innovation.*

This report is made possible by general support to CSIS. No direct sponsorship contributed to this report.

This report is produced by the Center for Strategic and International Studies (CSIS), a private, tax-exempt institution focusing on international public policy issues. Its research is nonpartisan and nonproprietary. CSIS does not take specific policy positions. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the author(s).