GLOBAL X INSIGHTS

Memory Is the New Bottleneck in AI Semiconductors

Ido Caspi
icaspi@globalxetfs.com

Date: March 23rd, 2026
Topic: AI Semiconductors, Artificial Intelligence

The multi-trillion-dollar AI buildout is straining the semiconductor value chain in ways the market initially underappreciated. GPUs have dominated the narrative, but as model sizes grow and data movement becomes the limiting factor, the bottleneck is shifting from compute to memory. Inside a hyperscale data center, the "memory wall" is increasingly the binding constraint on how fast AI applications can run and how far they can scale.[1]

That constraint is showing up most clearly in high-bandwidth memory (HBM), the critical companion to advanced GPUs. Without it, even the most powerful AI chip is throttled. After memory prices surged 246% year-over-year (YoY) in 2025, suppliers are now effectively sold out through 2026.[2] As a result, memory companies now sit at an opportunistic junction of structural demand growth, constrained supply, and unmatched pricing power.

This dynamic reinforces that AI hardware is broader than GPUs, and that value is likely to accrue across a wider set of semiconductor enablers. The Global X AI Semiconductor & Quantum ETF (CHPX) is designed to provide exposure to that evolving AI hardware ecosystem.

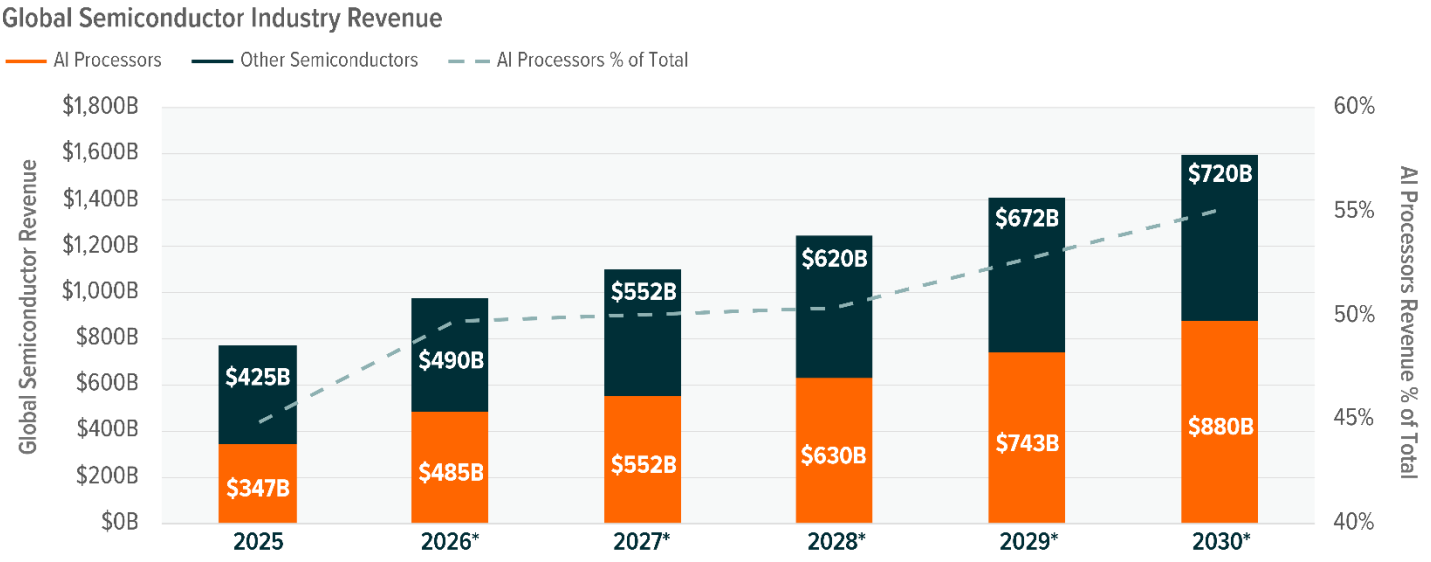
Key Takeaways

- AI semiconductors have seen explosive growth, fueled by record hyperscaler capital expenditure (capex) and a compounding buildout cycle.
- As large-scale inference and agentic AI initiatives ramp up, memory has emerged as the binding constraint
- Memory leaders like SK Hynix and Micron have benefited from sold-out capacity, long-term supply agreements, and rising margins, compounding the investment opportunity within the AI semiconductor ecosystem.

AI Semiconductor Spend is Accelerating and Broadening

The AI semiconductor industry's growth continues in force. Global chip sales rose 22% YoY in 2025 and are projected to grow another 26% in 2026 to approximately $975 billion. Generative AI chips are a primary engine of that expansion, with revenue expected to reach about $500 billion in 2026, or nearly half of total industry sales.[3]

GLOBAL SEMICONDUCTOR MARKET COULD REACH $1.6 TRILLION BY 2030, LED BY AI CHIPS



*Forecast

Source: Global X ETFs estimates with info derived from: Deloitte. (2026, February 5). 2026 Global Semiconductor Industry Outlook.; McKinsey. (2026, January 15). Hiding in plain sight: The underestimated size of the semiconductor industry.

That growth is being funded by hyperscaler capex, in what is shaping up to be one of the largest buildouts in corporate history. In 2026, hyperscalers could spend nearly $650 billion on capex, with a significant portion directed toward AI data centers.[4] Nvidia notes hyperscalers are deploying roughly 72,000 GPUs per week, marking the fastest product ramp for GPUs in the company's history.[5]

Two forces are extending the runway. First is the shift from AI model training to large-scale inference. Second is the industry's move toward a tighter, annual hardware upgrade cadence. Nvidia's Vera Rubin platform is the prime example. Slated for release in the second half of 2026, Rubin is expected to deliver roughly five times better inference performance than prior generations, with subsequent architectures poised to drive additional step-function gains.[6] Shipment trends reinforce the broader trend, as Nvidia shipped an estimated 5.2 million Blackwell graphics processing units (GPUs) in 2025, with volumes projected to rise to 5.7 million in 2026 as Rubin ramps.[7]

Record AI Spend Creates Opportunities Across the Chip Stack

For every dollar spent on AI processing chips, another $1.00–1.50 flows into the surrounding stack, including advanced packaging, power management, cooling systems, networking, and memory.[8] For example, one Nvidia GB200 NVL72 system requires 72 Blackwell GPUs and 36 Grace central processing units (CPUs), along with specialized networking infrastructure, chip-mounted liquid cooling systems, and up to 13.4 terabytes (TB) of high-bandwidth memory.[9]

And as these systems scale, the constraint is shifting from processing to connectivity and data access. High-speed networking is essential for linking GPUs and clusters, and spending is rising rapidly, with data center networking projected to grow from $20 billion in 2025 to $75 billion by 2030, a 30.2% annualized growth rate.[10]

Yet while networking enables system-level connectivity, it is memory bandwidth and capacity that increasingly determines how efficiently these systems perform at scale.

Memory Is Today's Critical Bottleneck

No matter how powerful the GPU or application-specific integrated circuit (ASIC), the quantity of data and the speed at which it can be moved determine a large language model's performance. These models operate on billions to trillions of parameters, creating bandwidth requirements that conventional memory architecture simply cannot meet. To solve this issue, high-bandwidth memory was designed as a specialized form of advanced dynamic random-access memory (DRAM) that delivers exceptionally fast data transfer to GPUs and AI accelerators. By stacking memory chips vertically and positioning them near the processor, HBM provides significantly higher bandwidth and energy efficiency than traditional memory.

The current standard, HBM3E, achieves approximately 1.2 terabytes per second (TB/s) of memory bandwidth per stack. The next iteration HBM4, which is now entering mass production and early shipments, will push that figure to over 2 TB/s while further reducing power consumption.[11] This roadmap provides several-fold higher bandwidth in a much smaller footprint, enabling the performance gains required for next-generation AI workloads.

As the industry solution to the memory wall, HBM has transformed from a volatile commodity component into a high-margin asset. HBM is entering a new paradigm characterized by sold-out capacity, unprecedented pricing power, and demand visibility extending years into

the future. HBM costs approximately five times more than standard memory, with prices generally secured via long-term contracts that insulate producers from spot market volatility.[12]

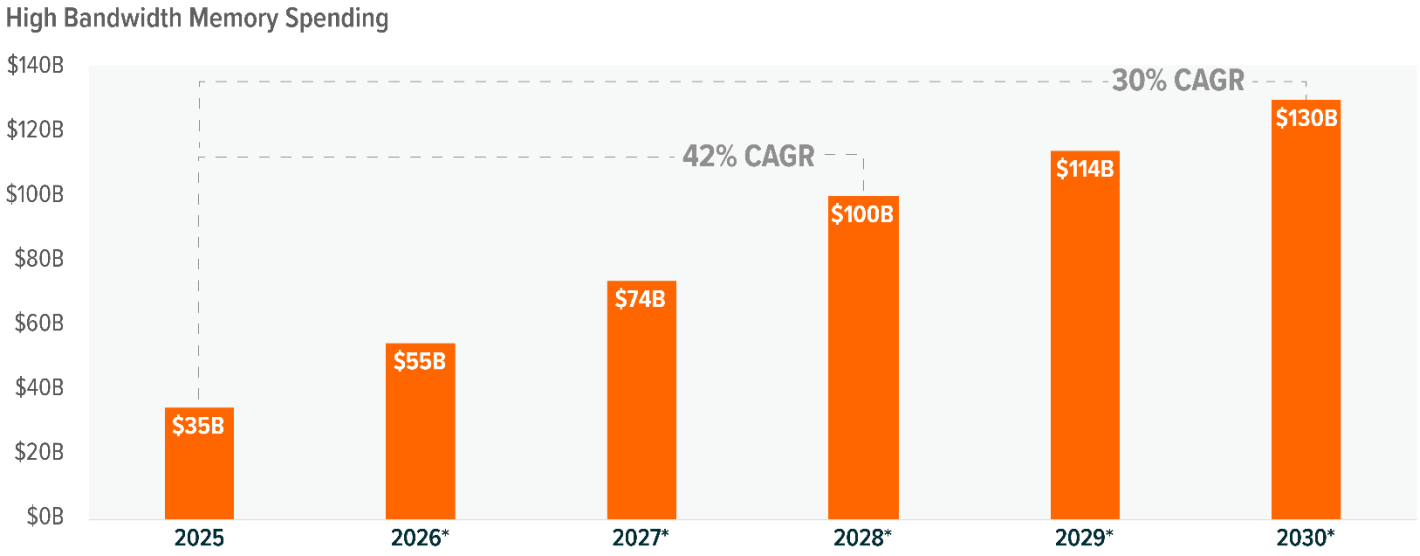
HBM Market Poised for Explosive Growth

As broader AI spending accelerates, the HBM market has witnessed rapid growth and it is expected to continue. Globally, HBM spend is projected to grow 58% YoY in 2026 to reach $54.6 billion and then compound at roughly 42% annually to approach $100 billion by 2028.[13,14] At this rate, the $100 billion milestone arrives two years ahead of prior industry forecasts and would surpass the size of the entire DRAM market as of 2024.[15]

This is also expected to boost total memory semiconductor revenues, which are projected to exceed $440 billion in 2026, a 30% YoY increase. By 2030, memory is expected to represent an increasingly dominant share of the semiconductor value chain as AI workloads intensify.[16]

Supporting this sustained growth trajectory for memory-based solutions are three main factors. First, the shift from AI training to inference will increase memory requirements as inference deploys models at scale. Second, the emergence of agentic AI systems capable of multi-step reasoning requires persistent memory contexts. Third, sovereign AI initiatives globally are creating demand pools outside traditional hyperscaler channels.

HIGH BANDWIDTH MEMORY SPENDING COULD REACH $130 BILLION BY END OF DECADE



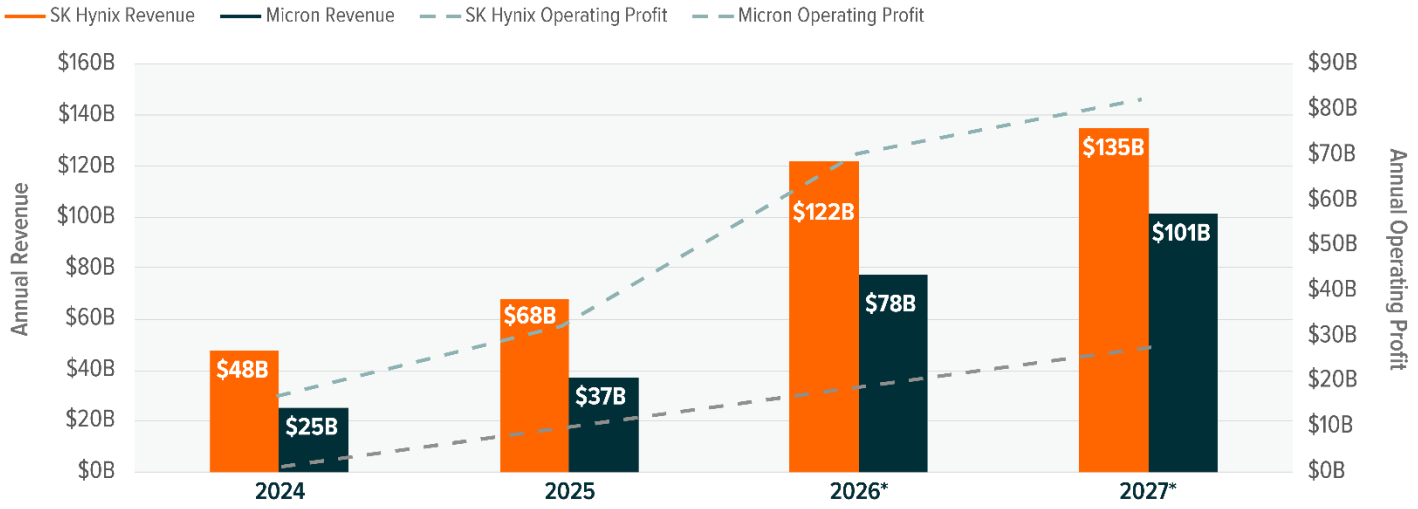
CAGR = Compound Annual Growth Rate. *Forecast.
Source: SK Hynix. (2026, January 5). 2026 Market Outlook – "Focus on the HBM-Led Memory Supercycle."; Bloomberg. (2025, January 13). High-Bandwidth Memory Chip Market Could Grow to $130 Billion by 2033.

Pure-Play Leaders in High-Bandwidth Memory

As the memory-bottleneck gets addressed, the pure-players stand to benefit. Currently, SK Hynix dominates HBM, having successfully parlayed its technical leadership into significant market share, but companies like Micron are catching up on the technology.[17]

REVENUE AND PROFITS HAVE SURGED FOR HBM LEADERS ON RELENTLESS AI MEMORY DEMAND

SK Hynix & Micron Annual Revenues and Operating Profit



Values in USD; SK Hynix values are converted from South Korean won (KRW) to USD. Not a guarantee of future results. *Forecast.
Source: SK Hynix. (2026, Jan 29). FY2025 Q4 Earnings; Factset. Company Financials and Analyst Estimates. Data accessed March 4, 2026.
Leaders are based on market share.

SK Hynix: The Current HBM Market Leader

SK Hynix is recognized as the first mover in HBM, backed by early product innovation and its manufacturing scale. They currently own 57% share of global HBM revenue and 62% of global shipment volume.[18] The company extended its lead by launching the first HBM3E in early 2024 and HBM4 in September 2025, delivering pin speeds above 10 gigabits per second (Gbps) and bandwidth exceeding 2 TB/s per stack, making aa 60% performance improvement and 40% power efficiency gain over HBM3E.[19,20]

This leadership is already showing up in company fundamentals, with 2025 revenues rising nearly 45% YoY to ~$68 billion and operating profit nearly doubling to ~$33 billion.[21]

A key driver has been SK Hynix's position as Nvidia's primary HBM supplier, having already secured more than two-thirds of Nvidia's HBM orders for the 2026 Rubin platform.[22] Management expects no slowdown in demand, supported by multi-quarter commitments and strong order visibility through 2026.[23] To meet demand projected to grow more than 30% annually through 2030, the company is ramping HBM4 production and plans to ramp DRAM capacity up to eightfold.[24]

Micron: The Challenger

As the only U.S. manufacturer of DRAM and NAND at scale, Micron occupies a unique position in the global memory supply chain. Over the past several years, Micron has closed what was once a meaningful technology gap with Samsung and SK Hynix by achieving manufacturing parity on leading-edge nodes while securing deep design-in partnerships with hyperscalers and AI hardware makers. These moves helped cement its status as the leading U.S.-based memory producer and a credible #2 player in HBM.[25]

Notably, Micron's entire calendar 2026 HBM supply is under price and volume agreements.[26] To reach its 22–23% HBM market share target, Micron raised fiscal 2026 capex to $20 billion from $18 billion previously to fund capacity expansion, which includes a dedicated HBM facility in Idaho backed by the CHIPS Act.[27]

Conclusion: Memory Is Shaping the AI Semiconductor Growth Trajectory

As AI hardware investments accelerate, memory is shifting from a volatile commodity into a strategic input that determines AI performance and scalability. As chip performance climbs and data movement becomes the constraint, HBM leaders are positioned to benefit from surging demand and structurally tight supply. Single-name exposure can amplify volatility, so exposure through a broader AI semiconductor ETF may be a cleaner way to express the theme.

Related ETFs

CHPX – Global X AI Semiconductor & Quantum ETF

Click the fund name above to view current performance and holdings. Holdings are subject to change. Current and future holdings are subject to risk.

Footnotes

1. The Scenariost. (2026, January 29). Deep Tech Startups & Venture Capital: An Analysis of 2025 | Chapter 3.
2. Introl. (2026, January 3). The AI Memory Supercycle: How HBM Became AI's Most Critical Bottleneck.
3. Deloitte. (2026, February 5). 2026 Global Semiconductor Industry Outlook.
4. Yahoo! Finance. (2026, February 6). Big Tech set to spend $650 billion in 2026 as AI investments soar.
5. DigiTimes Asia. (2025, May 29). How Nvidia's next-gen GPUs are fueling an inference supercycle.
6. The Economic Times. (2026, January 6). What's Nvidia's Rubin platform, and why it matters for AI.
7. TweakTown. (2025, June 30). NVIDIA expected to ship 5.2M Blackwell GPUs in 2025, 1.8M in 2026, and 5.7M Rubin GPUs in 2026.
8. Global X ETFs estimate, with info derived from: McKinsey. (2025, April 28). The cost of compute: A $7 trillion race to scale data centers.
9. Nvidia. (n.d). NVIDIA GB200 NVL72. Accessed on February 22, 2026.
10. Global X forecast, with information derived from: Bloomberg. (2024, March 21). Broadcom's AI Investor Day Takeaways.
11. Datacenter Technology. (2026, February 17). Mass-Produced HBM4 Boosts AI Datacenter Memory Bandwidth.
12. Tom's Hardware. (2024, May 6). Explosive HBM demand fueling an expected 20% increase in DDR5 memory pricing — demand for AI GPUs drives production cuts for standard PC memory.
13. SK Hynix. (2026, January 5). 2026 Market Outlook – "Focus on the HBM-Led Memory Supercycle".
14. CNBC. (2025, December 18). Micron stock pops 10% as AI memory demand soars: 'We are more than sold out'.
15. SK Hynix. (2026, January 5). 2026 Market Outlook – "Focus on the HBM-Led Memory Supercycle".
16. Ibid,
17. S&P Global. (2025, June 03). SK Hynix set to overtake Samsung as DRAM leader amid AI-driven memory boom.
18. SK Hynix. (2026, January 5). 2026 Market Outlook – "Focus on the HBM-Led Memory Supercycle".
19. Ibid.
20. Kynix. (2025, December 24). HBM3e vs HBM4: 2026 Specs, Performance & Supply Guide.
21. SK Hynix. (2026, January 28). SK Hynix Announces FY25 Financial Results.
22. SK Hynix. (2026, January 5). 2026 Market Outlook – "Focus on the HBM-Led Memory Supercycle".
23. Astute. (2025, December 17). SK Hynix ramps DRAM output eightfold but global memory scarcity pressures pricing and supply chains.
24. SK Hynix. (2026, January 5). 2026 Market Outlook – "Focus on the HBM-Led Memory Supercycle".
25. DigiTimesAsia. (2025, September 26). Micron rises to second in global HBM market as Samsung slips.
26. Micron. (2025, December 17). FQ1 2026 Financial Results.
27. Ibid.

Glossary

CPU: The CPU, or Central Processing Unit, is a general-purpose semiconductor chip that executes sequential instructions and manages everyday computing tasks.

GPU: A processor designed for massively parallel operations, originally for graphics rendering but now critical for AI and high-performance computing.

NAND: NAND stands for "Not AND," and it is a type of memory chip used to store data permanently in devices like smartphones, laptops, and data centers, even when the power is turned off.
