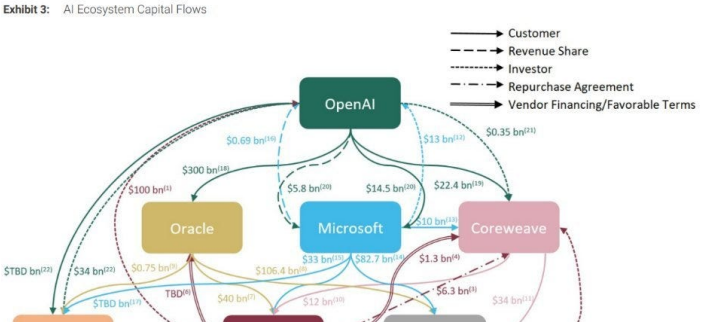
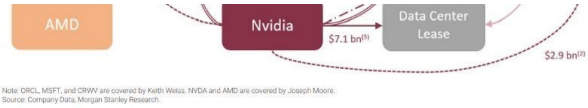
The Infinite AI Compute Loop: HBM Big Three + TSMC × NVIDIA × OpenAI Shaping the Next-Generation Industry Chain

Original Article By SemiVision Research (TSMC, Nvidia, SKHynix, Samsung , Micron , CXMT , Marvell , Broadcom , AWS , GUC , OpenAI)

SEMIVISION
OCT 13, 2025 • PAID

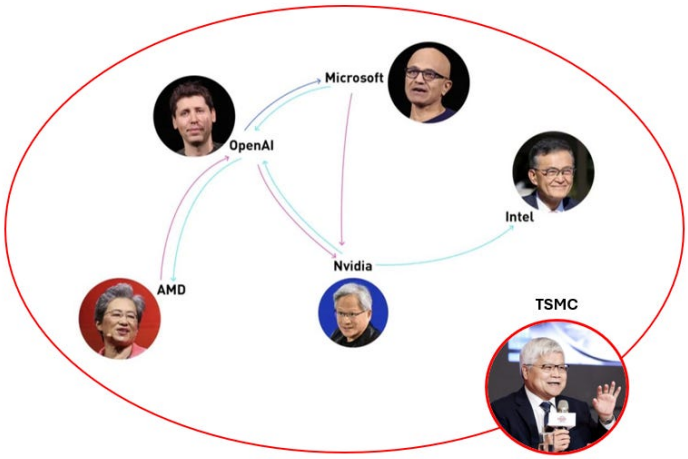
The development of AI has entered an unprecedented acceleration phase. From compute architectures and interconnect technologies to memory bottlenecks, the entire industry is rapidly revolving around a few key companies and technologies. NVIDIA's recent wave of massive investments has popularized what many in the industry now call the **"AI perpetual motion cycle."** In this cycle, AI chips drive compute demand; compute demand stimulates infrastructure investment; and infrastructure, in turn, fuels further expansion of AI chip applications.



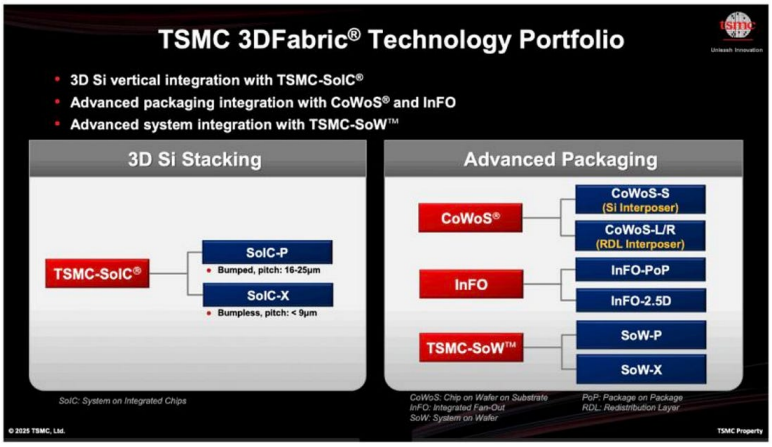


Note: ORCL, MSFT, and CRWV are covered by Keith Weiss. NVDA and AMD are covered by Joseph Moore.
Source: Company Data, Morgan Stanley Research.

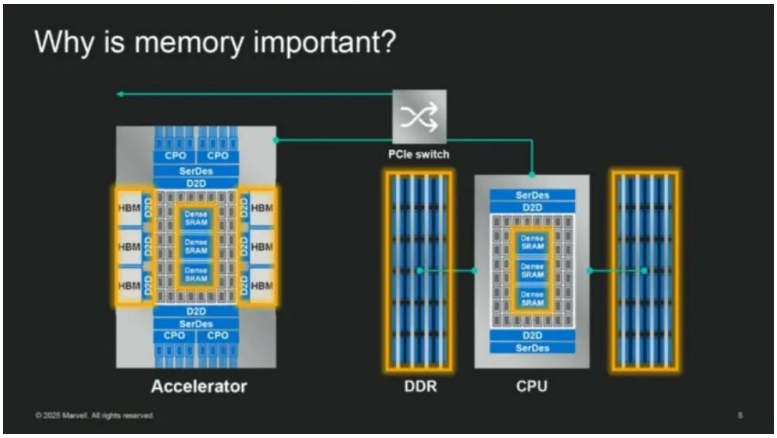
However, at the foundation of this entire cycle lies **TSMC's advanced process and packaging capabilities**.



Today, AI chips and high-bandwidth interconnects have fully entered the **"advanced process + advanced packaging"** era. Whether it's optical engines, sub-3 nm AI logic nodes, or various advanced packaging technologies such as **InFO, CoWoS, SoW, and CoPoS**, and even future SiC interposers (CoSoS), **TSMC remains the sole strategic hub** capable of supporting the entire stack — from design and manufacturing to packaging and system integration. At the same time, the industry has started using the term **CoWoP**, where the "P" refers to both **PCB** and the broader **platform** concept, emphasizing the depth and breadth of future system-level integration.



For an AI chip to truly achieve high performance, it's not just about the process node. It involves **ASIC design services, power and signal integrity (PI/SI)**, and high-speed transmission planning from **RDL copper interconnects to optical engines**, all of which drive fundamental changes in data center architecture. NVIDIA's cluster strategy is evolving from **Scale Up** and **Scale Out** toward **Scale Across**, accelerating the adoption of **Optical Circuit Switching (OCS)**. From **PIC (Photonic Integrated Circuits)** to **MZI (Mach-Zehnder Interferometer)** designs and next-generation transmission materials (e.g., **TFLN, BTO, SOH**) supporting speeds beyond 400 G, the entire high-speed optical interconnect ecosystem is being rapidly restructured.

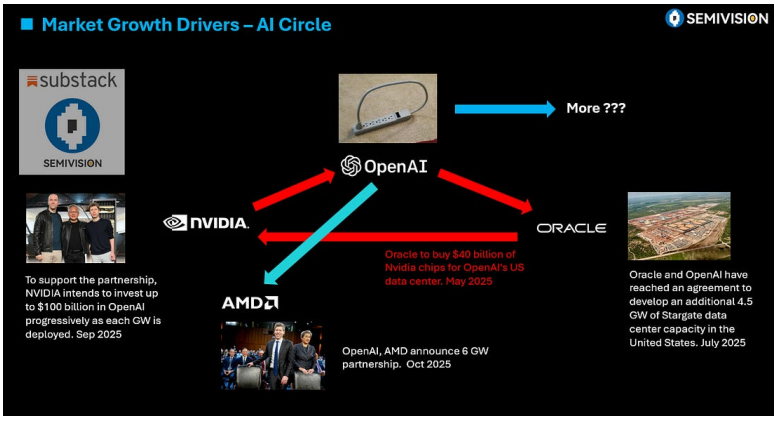


However, the ultimate performance boundary of AI systems is determined not only by compute and interconnect but also by **memory**. GPU performance has been advancing far faster than **HBM (High Bandwidth Memory)**, creating a widening **memory wall** that has become a central industry focus in 2025. **Marvell Technology** has proposed concrete solutions for memory architectures and optical interconnects at **HOT CHIPS 2025** and **OCP Global Summit 2025**, signaling a new wave of innovation to address this critical bottleneck.

The development of HBM goes far beyond memory process technology alone — it represents a **deeply coupled engineering system** with advanced packaging. From **TSV (Through-Silicon Vias)** and **TCB/HCB/HB bonding** to the gradual incorporation of the **Base Die** into TSMC's process roadmap, HBM has become inseparable from leading-edge integration. This means that in the future AI race, the **ability to master logic process + memory integration + optoelectronic packaging + interconnect architecture** will be the decisive factor across the value chain. And at the center of all these critical nodes stands **TSMC**.

This article begins by examining the **technological pillars behind the "AI perpetual motion machine"**, focusing on the evolution of HBM processing technologies, the **industry-wide challenge of the Memory Wall**, and the strategic implications of **optical interconnects and packaging platforms**

for future AI architectures.

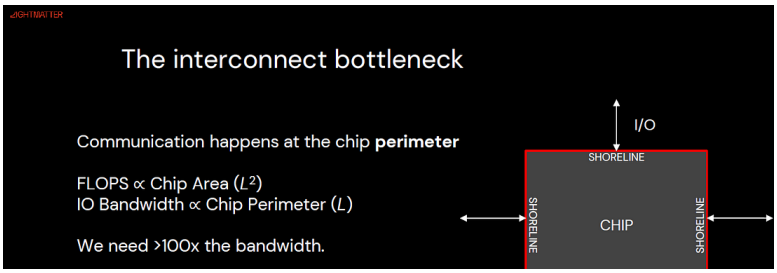


The vast interweaving of **capital and technological ecosystems** between **NVIDIA and OpenAI** — spanning hardware supply chains, cloud services, investment strategies, and startup networks — forms a powerful **financial engine** that drives the AI boom. Yet behind this flow of capital and compute lies a **hidden strategic hub**: **TSMC**.

From NVIDIA's GPUs to AMD's compute chips and large-scale cloud server processors, nearly all rely on TSMC's **advanced process nodes and 3DFabric packaging capabilities** for production. In other words, behind every arrow in the capital–technology diagram, the **true anchor of compute power** is TSMC's fabs and packaging platforms. Without TSMC, the ongoing AI capital explosion simply wouldn't exist.

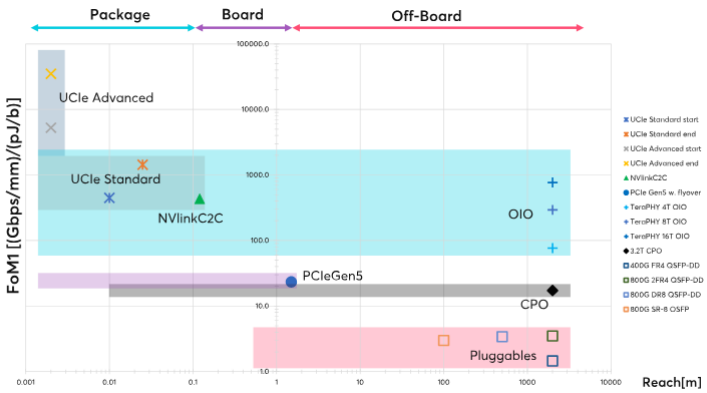
2026 Inflection Point: Systemic Reconstruction of AI Infrastructure

The year **2026 will mark a critical inflection point** in the structural transformation of the AI industry. Over the past few years, performance gains in large models have been driven by parameter scaling and GPU cluster expansion. But as compute power transitions from **Exa** to **Zetta scale**, exponential growth is hitting **physical and engineering limits**. The true bottlenecks are no longer GPUs themselves, but **memory bandwidth, packaging interconnects, thermal management, and power supply**. Compute is the engine, but bandwidth and thermal management are the drivetrain; when the drivetrain can no longer scale linearly, the entire value chain is redefined.



Energy and Thermal Become Strategic

1 GW-class data centers are driving the adoption of **liquid cooling and microfluidic (MCLP)** technologies. **High-thermal-conductivity diamond materials** are emerging as mainstream heat spreaders. Site selection now depends on both **grid capacity** and **water resources**. Meanwhile, **electrical interconnects are approaching physical limits** in bandwidth and power efficiency, making **CPO (Co-Packaged Optics)** and **OIO (Optical I/O)** the core of the next architectural shift.



Industry Realignment: From NVIDIA-Centric to Distributed Control

Cloud service providers (CSPs) such as **OpenAI, Google, Meta, AWS, and Huawei** are pursuing **"de-NVIDIA-ization"** strategies—developing their own **ASICs** and purchasing **HBM** directly. This disperses pricing power and turns **advanced packaging capacity into a strategic resource**. HBM is no longer "just DRAM," but a **deeply integrated system of logic, memory, and interposers**.

The **technical focus is shifting from stack height to bandwidth engineering**. Capabilities in **RDL (Redistribution Layer)** and **PDN (Power Delivery Network)** design are becoming key performance bottlenecks. In the short term, **CoWoS-R** supports scaling, while **CoWoS-L** and **SoIC** are expected to dominate mass production in the longer term.

Beyond Packaging: CXL and Photonics Reshape Memory Systems

The **CXL (Compute Express Link)** architecture is enabling **resource pooling** and **near-memory compute**. **CXL 2.0/3.x** combines **Switching**, **Pooling**, and **P2P topologies**, using **OS-level hot page migration and remote memory scheduling** to increase bandwidth utilization. However,

maintaining high performance ultimately **requires optical interconnect**.

Companies like **Ayar Labs, Celestial AI, Lightmatter, and Ranovus** are breaking through packaging and reticle limits with **silicon photonics**, achieving **nanosecond-level latency** and **Tb/s bandwidth**, enabling **distributed memory architectures**. **HBF/HFM** technologies serve as the **capacity layer**, supplementing memory pools for large-scale inference at lower cost.

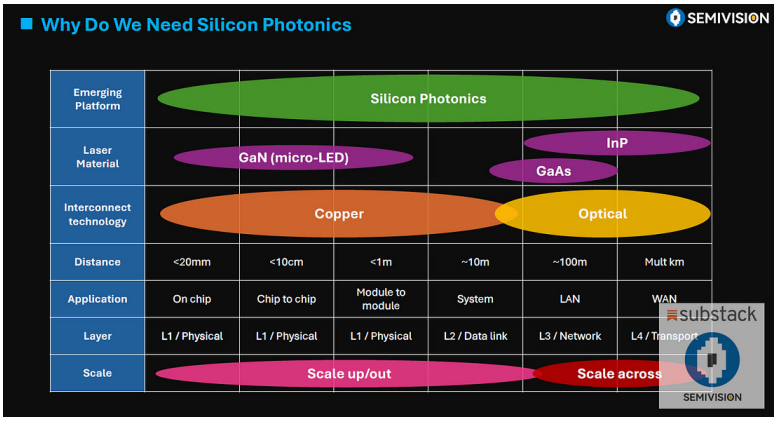
The Final Battleground: Packaging Space and System Design

The strategic competition ultimately returns to **advanced packaging and system co-design**. **Asymmetric HBM placement**, **multi-layer RDL**, **PDN partitioning**, and **optical module integration** are driving **STCO (System-Technology Co-Optimization)** to become the mainstream design methodology. Companies with strong **packaging integration capabilities** will command the new value chain.

- **Marvell and Broadcom** are pushing **custom HBM solutions** and **in-package optical interconnects** to enter emerging markets.
- **TSMC** is reinforcing its leadership through **HBM base-die technology** and **CoWoS/SoIC** platforms.
- **Korean–U.S. alliances** dominate high-end HBM capacity, while **China faces structural pressure**.

2026: The Year of Repricing Bandwidth, Power, and Packaging

By 2026, **electricity, water resources, advanced packaging capacity, and optical interconnect capability will all be repriced**. The winners will be those who can **convert bandwidth engineering into productivity and pricing leverage**. This is not merely a technological race—it is a **reorganization of power across the supply chain and infrastructure layers**.



Risks on the Horizon and Challenges remain:

- **AI scaling laws may face diminishing returns**.

- **CPO's thermal management and maintainability** issues are not yet fully solved.
- **CXL's NUMA software overhead** could slow commercialization.
- **HBF latency limits** its usefulness in training scenarios.

As **wires become antennas** and **photons replace copper**, the future power structure of AI infrastructure will belong to those who can **orchestrate bandwidth, thermal, and energy systems into a coherent whole**.

For Paid Members , SemiVision will discuss topics on

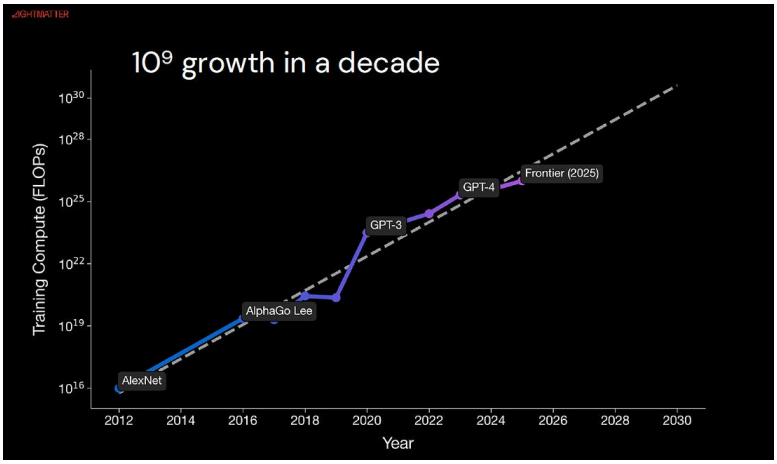
- The Breakpoint in AI Capability Curves: A Critical Nonlinear Leap by 2026
- The Evolution of HBM, HBM IP, and the Memory Wall Challenge
- The Concept of HBM IP
- TSMC's Strategic Position in the HBM Base Die Ecosystem
- Bandwidth Limits and Packaging Bottlenecks: From DRAM Scaling to RDL Electrical Control
- From CoWoS-R Validation to System-Level Packaging Integration
- The Memory Wall Problem
- Memory Allocation and Expansion under the CXL Architecture
- Memory Expansion and Allocation
- Solutions Combining Optical Chips and Memory
- Celestial AI Photonic Fabric
- Lightmatter Passage M1000
- Ayar Labs TeraPHY Optical I/O
- Ranovus (Odin) Multi-Wavelength Optical Platform
- Academic Research: Optical Multi-Stacked HBM
- Will HBM Be Replaced by Flash? — The Potential of HBF/HF
- SemiVision's Quick FAQ on HBM & Memory Technologies
- Q1: How has HBM evolved across generations, and what are the key features?
- Q2: What is HBM IP and why is it needed?
- Q3: What is the "Memory Wall" and why does it exist?
- Q4: How does CXL address memory allocation challenges?
- Q5: How can optical chiplets and photonics help overcome the Memory Wall?
- Q6: Will Flash replace HBM in the future?
- Technology Advancement for Performance Boost in TSMC's viewpoints

- HBM Allocation and Packaging Space Optimization Strategies
- Marvell : SRAM — Boosting Bandwidth and Efficiency in AI/XPU Devices
- Marvell: Custom HBM — Tailoring Memory to Match XPU Requirements
- Marvell’s Custom HBM Architecture
- Marvell : CXL — Expanding and Sharing Memory Through Compute Express Link
- Structera A: Near-Memory Accelerator
- Integrated Perspective: Marvell’s Memory Strategy
- Advantages and Limitations of ASICs in AI Inference
- Why are major tech giants investing in custom ASICs?
- ASIC WAR : The Race Toward Specialization and Systemization
- The Impact of the AI Chip Era on Memory and HBM
- Memory Requirements for AI Training and Inference
- Large-Scale Clusters and CSP In-House ASIC Development
- HBM: The Key Memory Technology Behind Generative AI
- 1. Basic Structure and Advantages of HBM
- 2.HBM Manufacturing and Competitive Landscape
- 3. HBM Supply-Demand Outlook and Market Size
- Why Cloud Service Providers Develop Their Own AI ASICs
- ASIC Growth Drives HBM Demand
- HBM Competitive Landscape and Supply Chain Evolution
- Cerebras :Near-memory computing and HBM alternatives
- Summary of Explosive Growth of the HBM Market
- Structural vs. Cyclical Cycles — Will AI Reshape the Traditional Memory Boom–Bust Pattern?
- HBM4e and Advanced Logic Process Competitiveness — Process Dependence and Strategic Collaboration in Next-Gen Memory
- Strategic Collaboration Models Between Foundries and Memory Vendors
- Edge AI and Mobile HBM Technology – The Outlook for High-Bandwidth Memory in Smartphones, XR, and Automotive Devices
- Breakthrough Conditions: Power, Packaging, and Cost
- Strategic Implications for the Edge AI Era
- Supply Chain Bottlenecks and Capacity Expansion — Key Factors in HBM / Advanced Memory Manufacturing
- Materials and Process Equipment Bottlenecks in Memory Device

- Strategies to overcome these bottlenecks
- Materials and Process Innovations – Impact of New Technologies on Yield, Cost, and Capacity
- 5 nm Base Die and Logic Packaging Innovations:
- Risk Management and Road-Mapping for New Process Technologies

AI Supercycle and the Restructuring of the HBM Supply Chain
From Nonlinear Capability Leaps to Power Shifts in the Memory Industry

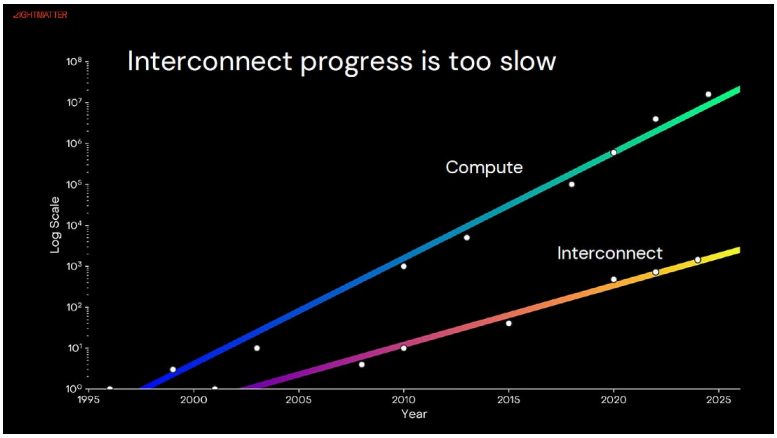
The Breakpoint in AI Capability Curves: A Critical Nonlinear Leap by 2026



The year 2026 is poised to become a major inflection point in the global AI development trajectory. A compute-driven “capability curve break” is emerging — not merely from scaling models, but from a systemic transformation spanning technology stacks, infrastructure, and capital markets. Looking back, deep learning’s rise in 2012 and the explosion of large language models (LLMs) in 2022 reshaped the entire tech and industrial landscape. Now, with leading model developers expected to increase training compute by roughly 10× by the end of 2025, a concentrated wave of breakthroughs is anticipated in early 2026, acting as a powerful catalyst across the ecosystem.

AI capability growth is not linear. Over the past decade, simultaneous increases in compute, data, and model parameters have followed scaling laws—as model size grows, error rates drop sublinearly. However, once total compute reaches the exaFLOPs–zettaFLOPs range, improvements may shift from a slow curve to step-function leaps. A 1 GW-scale data center can deliver compute equivalent to thousands of supercomputers combined, dramatically boosting language models’ abilities in reasoning chain length, memory, tool use, multimodal integration, and logical coherence. This represents a system-level qualitative leap, not just a quantitative scale-up.

Yet such nonlinear leaps are not guaranteed. The field has long debated the “scaling wall” hypothesis: beyond a certain threshold, performance gains may flatten or stall, entering diminishing returns. Pessimists argue algorithmic and architectural limits will cap model performance, with skyrocketing training costs and energy usage. Optimists counter that synthetic data and transfer learning can sustain training efficiency, while innovations like Mixture-of-Experts (MoE), retrieval-augmented generation, and multi-agent systems can push beyond saturation. In large-scale synthetic data environments, models show no signs of collapse, suggesting the scaling wall may be much farther away than expected.



Even if capability curves continue rising, the physical world bottlenecks must be confronted. As 1 GW data centers become standard, energy consumption, cooling, optical interconnect bandwidth, and land infrastructure will pose unprecedented challenges. A single mega training hub could consume as much electricity as a small country. To maintain stable operation, data centers must adopt microchannel liquid cooling, diamond heat spreaders, and direct liquid cold plates. Copper wiring can no longer meet bandwidth demands, making CPO (Co-Packaged Optics) and OIO (Optical I/O) inevitable. Moreover, site selection must shift closer to hydropower or nuclear energy sources to sustain such massive power levels. Infrastructure is no longer just an IT expense — it has become a national strategic asset.

Lightmatter: Transforming AI Infrastructure with the Power of 3D Photonics
SEMIVISION · AUGUST 29, 2025
Read full story

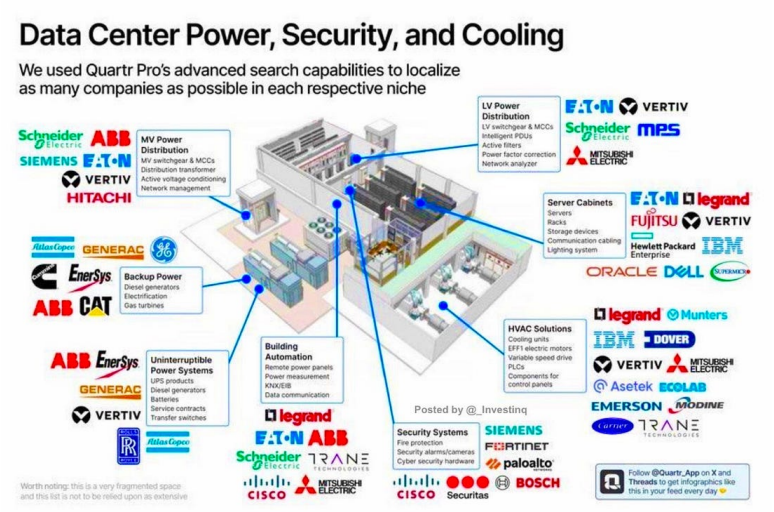
TSMC x Nvidia : Breaking the Thermal Wall: How Advanced Cooling Is Powering the Future of Computing
SEMIVISION · OCTOBER 5, 2025
Read full story

Market Repricing and Supply Chain Realignment Driven by the

Compute Surge

When AI capabilities make their **step-function leap in 2026**, a powerful **multi-industry, multi-asset market repricing wave** will follow.

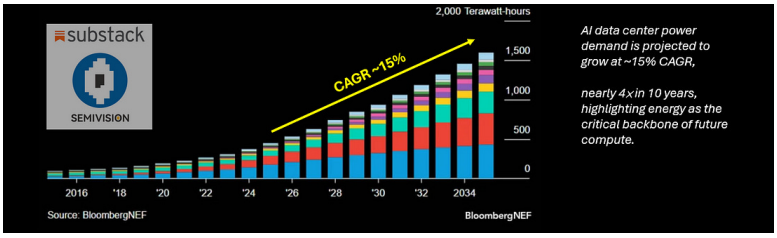
First, AI infrastructure suppliers will be direct beneficiaries. Companies providing **advanced cooling, optical interconnects, packaging, and power solutions** will no longer be treated as "backend costs," but as **strategic nodes** on the compute curve.



Second, supply chains will undergo further restructuring. Under geopolitical pressures, the **United States will strengthen its strategic autonomy** in memory, chips, packaging, and energy. Meanwhile, **Asia—particularly South Korea and Taiwan—will become indispensable hubs for technology and production capacity**.

Third, AI adoption will diverge among end users. Not every enterprise will benefit equally from capability jumps; only those that can **translate AI efficiency gains into real productivity or pricing power** will emerge as winners. This **capability-to-value conversion** will redraw profit pools across industries.

Fourth, **scarce assets** will be repriced. **Power infrastructure, critical minerals, water resources, ports, and regulated concessions**, due to their non-replicable nature, will see a sharp increase in relative value. At the same time, **human experiences, brand trust, and proprietary data**, which AI cannot easily replicate, will become central to long-term value reassessment.



The **technological rhythm of this transformation is already clear**.

- **Compute architectures** will shift from GPU dominance toward **multi-chip parallelism and the rise of ASICs**.
- **Memory technologies** will evolve from **HBM3E to HBM4E and HBM5**.
- **Data center interconnects** will transition from **PCIe to NVLink and CXL combined with Optical I/O**.
- **Packaging technologies** will advance from traditional **2.5D/FCBGA** to **CoWoS, SoIC, and fan-out optical packaging**.

These shifts will fundamentally **reshape supply chain divisions**, elevating the strategic importance of **chip design, advanced packaging, photonic modules, and material suppliers**, and creating a new technological and industrial landscape.

It is increasingly evident that **2026 will not be a year of gradual, linear growth**, but a **sharp inflection point in the capability curve**. As compute power surpasses the **zettaFLOPs threshold**, **technological leaps, infrastructure stress, and capital market repricing will happen simultaneously**.

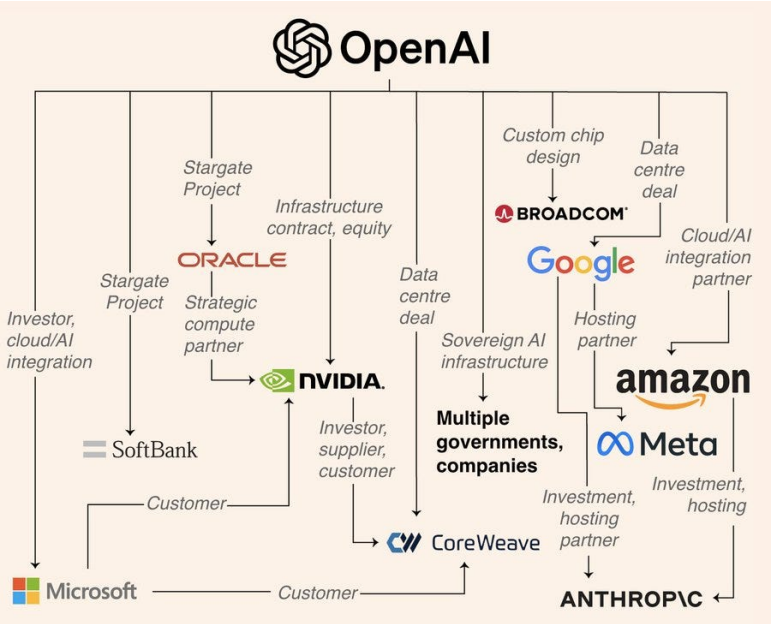
For the industry, this is a moment of **"accelerated compression"** — faster technology cycles, more intense capital competition, and infrastructure as a critical differentiator. In other words, **the future of AI will not just be about model evolution**, but about the **synchronous restructuring of the technology–energy–industry–capital–geopolitics chain**.

The landscape of the **AI chip industry is undergoing a profound transformation**. For years, the supply chain — from compute cores (GPUs) to HBM procurement — has been **almost entirely dominated by NVIDIA**, creating a highly concentrated structure. Today, however, technology giants including **OpenAI, Google, Meta, Amazon Web Services (AWS), and Huawei** are strategically pursuing a **"de-NVIDIA-fication"** approach, developing their own AI chips and **reshaping the power dynamics of the HBM supply chain**.

HBM (High Bandwidth Memory) is one of the **core components of AI chips**, offering **ultra-high bandwidth (> 1 TB/s)**, **high stacking (12–24 layers TSV)**, **significant power consumption (20–30 W per stack)**, and **complex co-packaging requirements**. The market is currently dominated

by **SK hynix, Samsung, and Micron**, with NVIDIA accounting for about **70% of total HBM procurement**. **HBM3 and HBM3E** are the mainstream specifications today, while **HBM4E** is expected to enter commercial deployment in **2026**.

OpenAI has recently **shifted its technology strategy**, moving away from full reliance on GPUs. It is now **collaborating with Broadcom** to develop **custom AI ASICs using TSMC's 3 nm process**, paired with **HBM4/ HBM4E** on **CoWoS-L or SoIC platforms**. This marks OpenAI's transition from being an **NVIDIA customer** to becoming a **direct HBM buyer**, a shift that represents not just a technical change, but a **fundamental redistribution of power within the supply chain**.



OpenAI CEO **Sam Altman recently traveled to South Korea**, signing **HBM supply LOIs** with Samsung Chairman Jay Y. Lee and SK Group Chairman Chey Tae-won, with a **planned procurement volume of 900,000 DRAM wafers per month**—equivalent to **roughly 75% of the combined capacity** of Korea's two major memory suppliers. This move indicates OpenAI's intention to **procure directly from memory manufacturers** rather than through NVIDIA, signaling an erosion of NVIDIA's dominance over the HBM market.

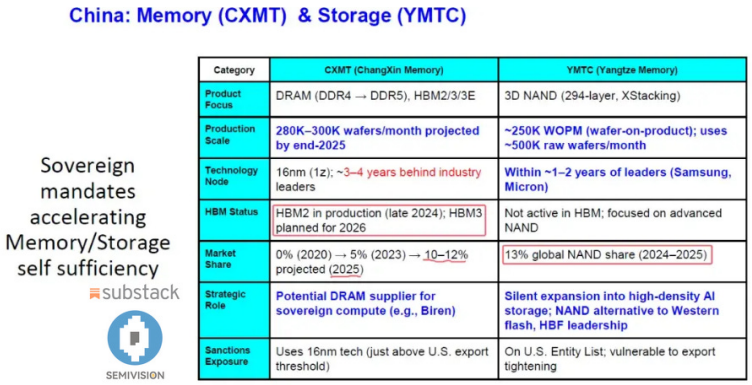
> **Samsung and SK join OpenAI's Stargate initiative to advance global AI infrastructure**

Historically, NVIDIA has **secured strong pricing leverage** through long-term pre-purchases, driving down unit HBM prices. In the future, as **OpenAI, Google, AWS, Meta, and Huawei** enter direct negotiations, **three major shifts** in HBM pricing and capacity allocation are expected:

1. **Bargaining power will decentralize**, reducing memory vendors' reliance on a single customer.
2. **ASP (Average Selling Price)** is expected to rise significantly, especially for **HBM4E and HBM5** product lines, which will undergo a pricing revaluation.
3. **Advanced packaging capacity** will quickly be locked up. Major OSATs like **TSMC, Amkor, ASE, and SPIL** will become critical strategic nodes in this new competitive landscape. Upstream packaging and testing resources will turn into **key strategic assets** fiercely contested by cloud giants.

HBM's technical evolution is also driving **parallel advancements in packaging and thermal management**. **CoWoS, SoIC, and H-Cube** technologies will see wider adoption, while **the high power dissipation of HBM** will push **microchannel liquid cooling** and **diamond heat spreaders** into the mainstream. At the same time, **CPO (Co-Packaged Optics)** and **OIO (Optical I/O)** will become essential solutions to bandwidth bottlenecks. In this tightly coupled compute–memory architecture, **power and thermal design are no longer secondary engineering details**—they are the **determinants of commercial viability**, and a new battleground for OSATs and materials companies.

This transformation carries **deep geopolitical implications**. OpenAI's collaboration with Korean memory suppliers strengthens the **U.S.–Korea tech alliance**, allowing the U.S. to **reduce its reliance on Chinese supply chains**, while Korea consolidates its role as a **strategic hub for AI semiconductors** through HBM and advanced packaging. Meanwhile, Chinese memory makers like **YMTC** and **CXMT** are being **further marginalized** in the high-end HBM market, unable to enter mainstream technology circles. This is not merely a **technological upgrade**, but a **supply chain restructuring and power shift**.



For the past five years, the AI chip market has been built on the **GPU-driven NVIDIA economy**. The next five years will mark a **multi-chip, multi-architecture, multi-supply-chain era**, where **competition and collaboration coexist**. OpenAI's in-house ASIC strategy is directly reshaping the HBM landscape. **Samsung and SK hynix**, with their capacity and technical strengths, are emerging as **new pricing power centers**; **TSMC and OSATs** will dominate **packaging and system integration**; and **NVIDIA's supply chain leverage will gradually diminish**.

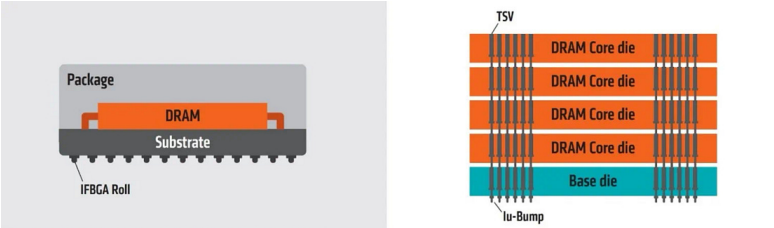
2026 will not only be the year of AI capability leaps—it will also mark the **beginning of synchronized restructuring across memory, packaging, energy, and industrial power structures**, redefining the entire **AI value chain**.

The Evolution of HBM, HBM IP, and the Memory Wall Challenge

1. The Development History of HBM

Origins and Fundamentals of HBM

High-Bandwidth Memory (HBM) is a stacked DRAM technology designed to deliver **extremely high memory bandwidth**, far surpassing that of traditional GDDR or DDR memory. Conventional memory improves bandwidth by increasing clock frequency or bus width, but these approaches are constrained by power consumption and signal integrity issues.



GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack

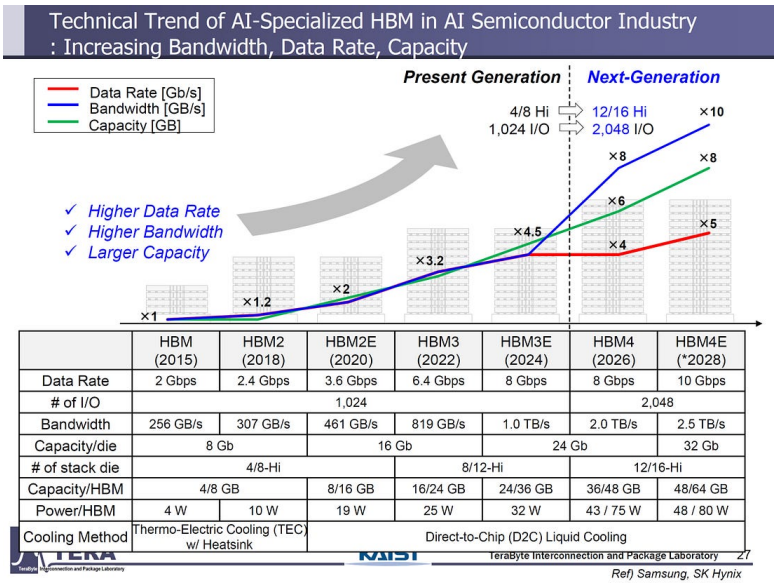
1.5V	Voltage	1.3V

HBM adopts a **3D-stacked architecture**: multiple DRAM dies are vertically stacked on top of a base logic die and interconnected using **thousands of Through-Silicon Vias (TSVs)** and **micro-bumps**. This stacked memory module is then connected to the processor via a **silicon interposer** (2.5D integration).

This **vertical stacking enables thousands of I/O connections per HBM stack**, providing bandwidth in the range of **hundreds of GB/s up to more than 1 TB/s per stack**. Because of the extremely dense interconnects, HBM must be placed in close proximity to the compute die, typically within a **CoWoS or SoIC advanced packaging platform**. Although HBM comes with higher manufacturing complexity and cost, it **significantly reduces power consumption and footprint** compared to traditional memory architectures.

Evolution of HBM Generations

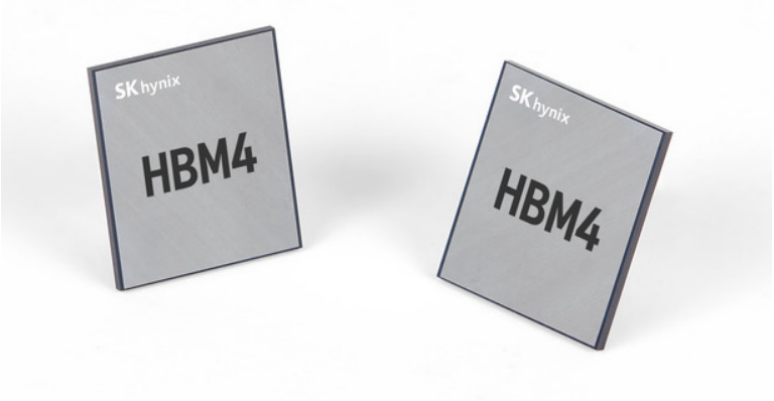
The table below summarizes the **standards, bandwidth, and key features of each HBM generation**.



	HBM (2015)	HBM2 (2018)	HBM2E (2020)	HBM3 (2022)	HBM3E (2024)	HBM4 (2026)	HBM4E (*2028)
Data Rate	2 Gbps	2.4 Gbps	3.6 Gbps	6.4 Gbps	8 Gbps	8 Gbps	10 Gbps
# of I/O	1,024					2,048	
Bandwidth	256 GB/s	307 GB/s	461 GB/s	819 GB/s	1.0 TB/s	2.0 TB/s	2.5 TB/s
Capacity/die	8 Gb		16 Gb		24 Gb		32 Gb
# of stack die	4/8-Hi			8/12-Hi		12/16-Hi	
Capacity/HBM	4/8 GB		8/16 GB	16/24 GB	24/36 GB	36/48 GB	48/64 GB
Power/HBM	4 W	10 W	19 W	25 W	32 W	43 / 75 W	48 / 80 W
Cooling Method	Thermo-Electric Cooling (TEC) w/ Heatsink		Direct-to-Chip (D2C) Liquid Cooling				

- In **2013**, JEDEC released the **first-generation HBM (JESD235)**, marking the beginning of standardized high-bandwidth stacked memory.
- Each subsequent generation has **increased bandwidth**, **expanded the number of channels**, and **raised stack height** to support more capacity.
- Starting with **HBM2**, pseudo-channel architecture was introduced to improve efficiency.

- **HBM4** represents a major leap, **doubling the number of channels to 32**, **lowering core voltage to 1.05 V**, and adding **Directed Refresh Management (DRFM)**, **WSO bus remapping**, and **decision feedback equalization** — enabling higher data rates and improved signal reliability for next-generation AI and HPC systems.



SK hynix completed the world's first HBM4 development in 2025 and is preparing for mass production. Looking ahead, HBM5/6 may adopt glass substrates and immersion cooling technologies.

> **SK hynix Completes World's First HBM4 Development and Readies Mass Production**

The future HBM roadmap shows four generations from HBM5 to HBM8 beyond 2029. Research from KAIST indicates that HBM5 could achieve bandwidths of up to 64 TB/s with 16-high stacking. HBM6/7/8 are expected to incorporate copper-to-copper bump-less hybrid bonding and double-sided interposers. HBM8 may even be integrated with **High-Bandwidth Flash (HBF)**, leveraging 3D NAND stacking to provide over 1 TB of capacity.

The Concept of HBM IP

HBM IP refers to the silicon intellectual property of HBM controllers and PHY circuits that can be adopted by SoC designers. This type of IP enables chip companies to quickly integrate HBM with AI accelerators or data center SoCs without developing the HBM interface in-house. Two examples illustrate this:

- **Synopsys HBM IP** – Synopsys provides a complete set of HBM controller, PHY, and verification IP. Its 6th-generation HBM4 IP achieves 12 Gb/s per pin and over 3 TB/s total interface bandwidth, with robust design verification and support inherited from previous generations.

High-Quality, High-Bandwidth Memory Interface Solution

- **Global Unichip (GUC) HBM IP** – GUC's HBM4 IP supports 12 Gbps/pin speeds, using proprietary interposer routing to reduce signal loss and power consumption, while integrating I/O monitoring within the controller to enhance reliability. The company has already taped out multiple 8.4/8.6 Gbps HBM3E products for data center customers.

GUC HBM Controller and PHY IP

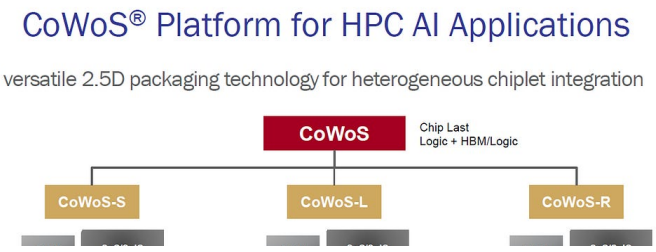
TSMC Process Nodes	Speed	Readiness
7nm HBM3	7.2 Gbps	V (6nm compatible)
5nm HBM3	8.4 Gbps	V (4nm compatible)
3nm HBM3	8.6 Gbps	V

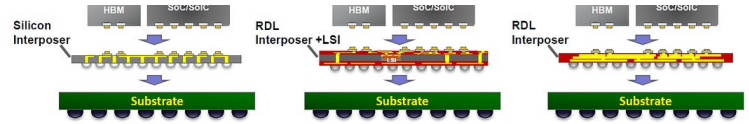
GUC D2D (GLink-2.5D & UCIe) IP

TSMC Process Nodes	Bandwidth (full duplex)	Readiness
7nm GLink 1.0	0.7 Tbps/mm	V (6nm compatible)
5nm GLink 2.3LL	2.5 Tbps/mm	V (4nm compatible)
3nm GLink 2.3LL	2.5 Tbps/mm	V
3nm UCIe 1.0	5.1 Tbps/mm	4Q23

A key signal emerged at TSMC's recent Tech Forum: TSMC is officially entering the design and manufacturing of HBM base dies, collaborating deeply with the three major DRAM vendors — Samsung, SK hynix, and Micron. This marks a strategic shift: HBM is no longer merely a DRAM process competition, but has entered a new era of **memory–logic co-design and integration**.

Currently, there are two main approaches to increasing HBM bandwidth. The first is **continued DRAM process scaling**, which boosts the data transfer rate of each individual chip. The second is **increasing the number of stacked layers (Hi stacks)** to expand total bandwidth. However, as signal transmission speeds approach their physical limits, **Power Integrity (PI)** and **Signal Integrity (SI)** at the packaging level have become critical factors for maintaining system stability. To simultaneously achieve high speed and stable operation, the **design of the Redistribution Layer (RDL)** has become the architectural centerpiece.





Source: TSMC

To ensure signal integrity and power stability in high-speed packaging, the number of RDL layers must be continuously increased, creating more sophisticated ground shielding and power distribution networks. This is why **TSMC's CoWoS-R platform** has become a crucial testbed for HBM packaging verification. CoWoS-R utilizes an organic interposer combined with multilayer RDL structures, enabling full-scale electrical, thermal, and mechanical validation under realistic large-area, high-density conditions.

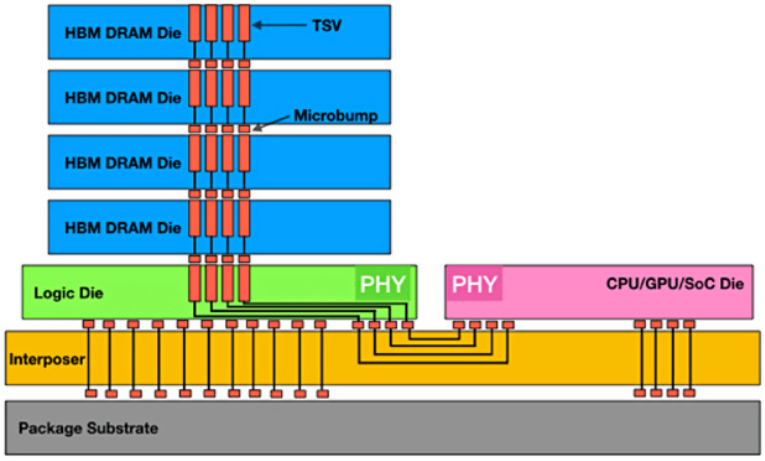
In other words, TSMC is redefining the technological foundation of HBM through its **co-optimization of advanced packaging and process technologies**. The focus has shifted from simply scaling DRAM stack height and bandwidth toward an era of **"System Bandwidth Engineering,"** where **electrical design at the packaging level** plays the central role in sustaining future performance scaling.

TSMC's Strategic Position in the HBM Base Die Ecosystem

As HBM4 and HBM3e enter the ultra-high-bandwidth era, the focus of memory design has shifted from DRAM stacking and process scaling toward **"co-integration of memory and logic."** In this transformation, **TSMC plays a pivotal role as both a foundry and advanced packaging platform** within the HBM ecosystem. At its Technology Forum and OIP (Open Innovation Platform) events, TSMC clearly stated that it will collaborate with the three major DRAM vendors — **Samsung, SK hynix, and Micron** — to co-develop the next generation of HBM base dies.

This **base die (logic die)**, located at the bottom of the HBM stack, is no longer just a simple control or test logic chip. It is evolving into a **high-performance logic layer** that integrates **high-density I/O PHY, clock management, error correction (ECC), and power control circuits**. Because these functions require advanced logic processes, DRAM vendors are increasingly relying on external foundries for manufacturing. According to TrendForce, starting with HBM4, the base die will move to FinFET nodes,

and for HBM4e may even adopt 3nm-class technologies. TSMC has therefore introduced a **two-tier foundry strategy**: using **12FFC+ (12nm class)** for mainstream products, and **N5 (5nm class)** for top-tier HBM4 customers, providing flexibility between performance and cost.



This tiered process strategy positions TSMC not merely as a manufacturing provider but as a **co-architect of the overall HBM packaging structure**. By working closely with DRAM vendors on **base die design, signal routing, and packaging validation**, TSMC is effectively transforming HBM from a traditional memory component into a **heterogeneous stacked device with integrated logic functionality**. This integration is essential for HBM to continue advancing toward **higher bandwidth and lower power consumption** in the coming generations.

Bandwidth Limits and Packaging Bottlenecks: From DRAM Scaling to RDL Electrical Control

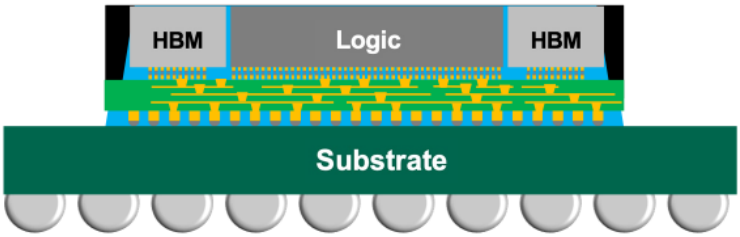
To increase HBM bandwidth, there are currently two main approaches:

1. **Enhancing DRAM process scaling**, such as shrinking cell capacitors and bitline dimensions or increasing I/O operating frequencies.
2. **Increasing the number of stacked layers (Hi count)** to boost overall data parallelism through denser TSV vertical channels.

While both methods can improve bandwidth, each faces fundamental physical and engineering limits.

As DRAM processes shrink to the 1b/1c nm nodes, resistance and capacitance effects become more pronounced, making signal delay and noise control increasingly difficult. Meanwhile, as stack heights reach 12-Hi and even 16-Hi, TSV channels lengthen, impedance rises, and coupling interference and thermal density issues worsen. As a result, even if DRAM device performance improves, the **overall bandwidth is constrained** if signal integrity (SI) and power integrity (PI) cannot be maintained within the package.

This is why **TSMC and its memory partners are now focusing on packaging-level challenges — specifically, the electrical and structural design of the Redistribution Layer (RDL)**. The RDL connects the base die, DRAM dies, and SoC, acting as the "vascular system" for high-speed data channels. To maintain stable transmission at >10 Gbps per pin, the RDL must address **impedance matching, crosstalk suppression, power noise management, and thermo-mechanical stress control** simultaneously.

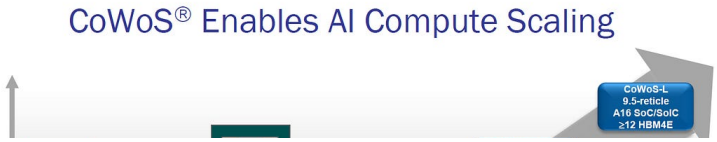


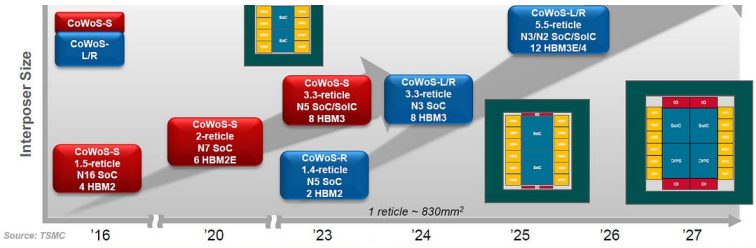
To achieve this, TSMC's **CoWoS-R architecture** employs **GSGSG (Ground-Signal-Ground-Signal-Ground) routing** and multilayer shielding structures within the RDL. Inner ground shields and dielectric isolation layers suppress high-frequency crosstalk and EMI, while **discrete decoupling capacitors** are integrated near the C4 bumps to stabilize power delivery, reduce IR drop, and minimize noise coupling. Because HBM's signal paths are extremely short, **even minor impedance discontinuities can shrink eye diagrams or increase BER**, making precise control over **RDL geometry, dielectric thickness, and copper trace distribution** essential.

As signal channel counts increase, more RDL layers are required, which introduces new manufacturing challenges: **layer-to-layer alignment must be controlled at the micron scale**, and CTE mismatches in dielectric materials can cause warpage or layer slippage. TSMC addresses these issues in CoWoS-R by using an **organic interposer**, whose material elasticity helps absorb stress and reduce deformation mismatch between silicon and substrate, ensuring long-term reliability and stable transmission.

In short, among the multiple bottlenecks in scaling HBM bandwidth, **the PI/SI performance of the RDL has become the critical support point**—the decisive factor determining whether a packaging architecture can pass high-speed verification.

From CoWoS-R Validation to System-Level Packaging Integration





To validate the complex RDL designs and electrical stability described above, **TSMC and its customers widely use CoWoS-R (Chip on Wafer on Substrate – RDL Interposer)** as their **technology verification platform**. Unlike traditional TSV silicon interposers, CoWoS-R uses **organic interposers combined with RDL routing**, offering higher manufacturing yield and faster design iterations. Engineering teams can use this platform to **simulate full signal routing between multiple HBM dies and large SoCs**, performing comprehensive tests such as **eye diagrams, insertion loss, crosstalk, IR drop, and power noise**, allowing fine-tuning of line width, spacing, layer count, and decoupling capacitor placement.

TSMC's CoWoS-R currently supports **minimum 4 µm pitch (2 µm line/space)** RDL layouts, maintaining strong electrical performance in ultra-dense environments. This makes CoWoS-R the **"mid-stage validation platform" for HBM packaging**: before entering mass production with **CoWoS-L** or **CoWoS-S**, electrical and mechanical fine-tuning is carried out at this stage to reduce production risk.

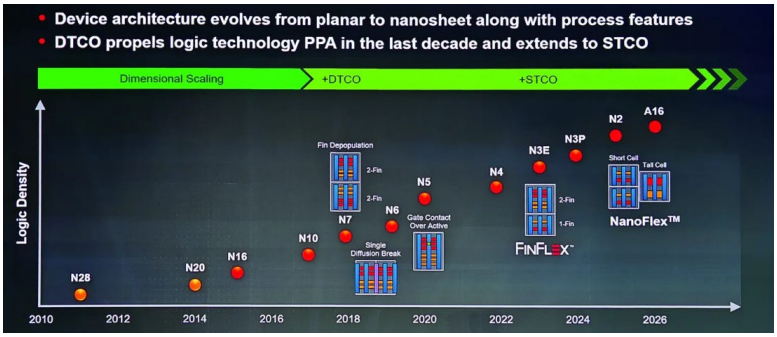
The **end-to-end design flow** — from base die to final packaging — requires **multi-layered optimization**:

- **Base Die I/O and PHY** must be co-designed with the package routing to shorten signal paths and reduce latency.
- **RDL routing and shielding strategies** must balance signal frequency, crosstalk, and power noise.
- **C4 bump arrays and underfill material selection** directly impact die-to-die alignment precision and mechanical strength.
- Additional considerations include **thermal management** (e.g., thermal vias or micro-cold plate structures near high-power regions) and **clock tree jitter control** to ensure system synchronization.

After these design steps, engineering teams use the CoWoS-R platform for **SI/PI simulation and real measurements**, iteratively adjusting RDL routing, shielding layers, and decap placement based on the results. This process not only verifies the stability of high-speed channels but also establishes CoWoS-R as a **critical validation stage for next-generation HBM packaging**.

Ultimately, through this **"design–package–test" closed-loop flow**, TSMC is advancing its advanced packaging platform toward **System Technology**

Co-Optimization (STCO), laying the **high-bandwidth, low-power, and scalable foundation** for the HBM and AI accelerator era.

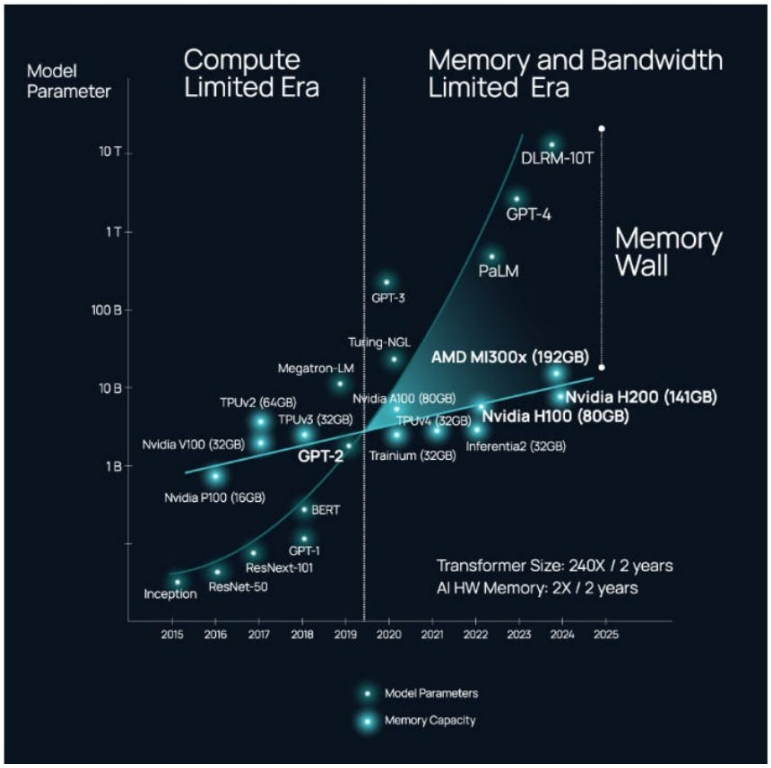


The growing adoption of **HBM IP** further enables widespread integration of HBM into AI GPUs, data center SoCs, and high-performance computing ASICs, while reducing development risk and time-to-market.

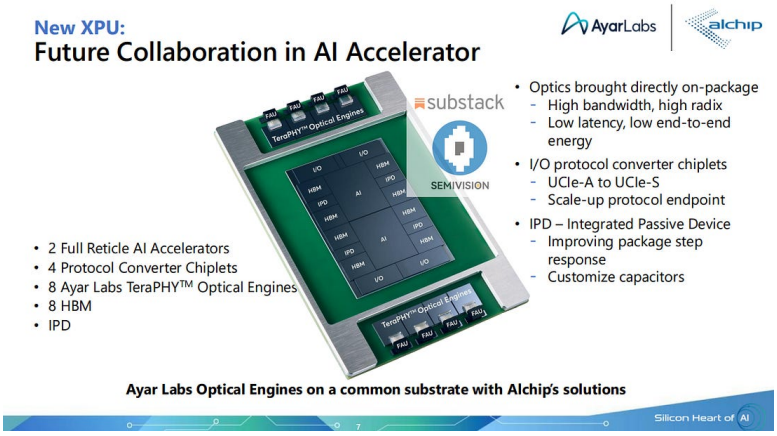
The Memory Wall Problem

Problem Definition

The *Memory Wall* refers to the growing gap between processor speeds and memory access speeds. As Moore's Law advances, CPU/GPU compute performance has improved by tens of thousands of times, while DRAM bandwidth has only increased by about a hundred times. As a result, processors spend a large amount of time idling while waiting for data — a phenomenon known as the "memory wall."



According to Ayar Labs' glossary, the memory wall arises because processor performance improvements far outpace memory bandwidth growth. Additionally, packaging I/O limitations and signal integrity constraints make memory expansion difficult. While logic transistor density and compute capability have advanced rapidly, DRAM scaling has stagnated. HBM provides higher bandwidth, but it is expensive and limited in supply.



Impact on AI and HPC

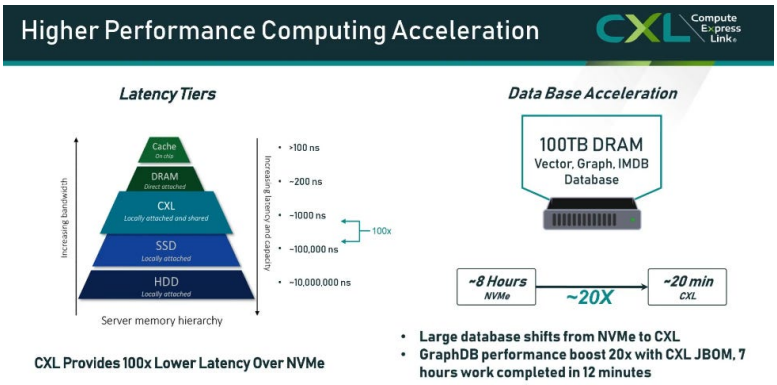
AI training and inference demand massive memory bandwidth and capacity. Modern large language models (LLMs) contain tens of billions to trillions of parameters. For example, training GPT-3's 175 billion parameters in 16-bit precision requires about 3 TB of memory, while trillion-parameter models need around 32 TB.

However, current HBM stacks are limited by chip shoreline I/O and packaging area, which constrain the number of stacks that can be integrated. As model sizes continue to grow, existing architectures struggle to keep up. This creates an urgent need for new memory architectures or higher-density HBM stacking to mitigate the memory wall problem.

Memory Allocation and Expansion under the CXL Architecture

CXL Overview
Compute Express Link (CXL) is an open standard built on the PCIe physical layer that enables high-speed, low-latency, cache-coherent interconnects between processors and devices. CXL consists of three protocols:

- **CXL.io** – Handles basic management and I/O transfers, similar to PCIe.
- **CXL.cache** – Allows external devices to coherently cache host memory.
- **CXL.mem** – Enables the host or device to directly access device memory (e.g., DRAM, Optane, or HBM) using load/store semantics. It supports memory sharing and pooling, allowing dynamic allocation based on workload demands.



CXL devices are categorized into three types:

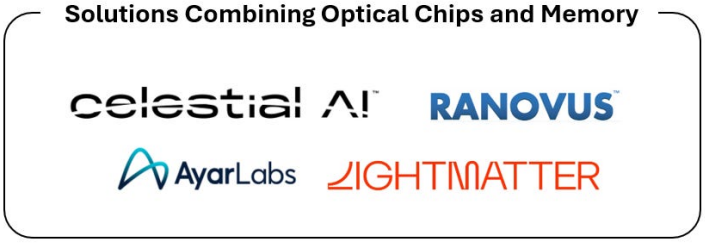
- **Type 1**: Accelerators without local memory (e.g., SmartNICs).
- **Type 2**: Accelerators with local memory (e.g., GPUs, FPGAs).
- **Type 3**: Pure memory expansion devices that provide additional memory capacity to the host.

Memory Expansion and Allocation
Starting from version 2.0, the CXL standard supports switching and memory pooling. Memory pooling allows multiple hosts to share a pool of memory devices and allocate memory dynamically as needed. For example, in a data center, a shared CXL memory pool can be created so that different servers can use portions of the pool based on their workloads—without the need to overprovision DRAM for each individual server. This reduces idle memory and lowers costs.

CXL 3.0/3.1 further enhances flexibility by introducing multi-level switching, point-to-point memory access, and support for up to 4,000 nodes in a single fabric, enabling highly scalable and flexible memory sharing.

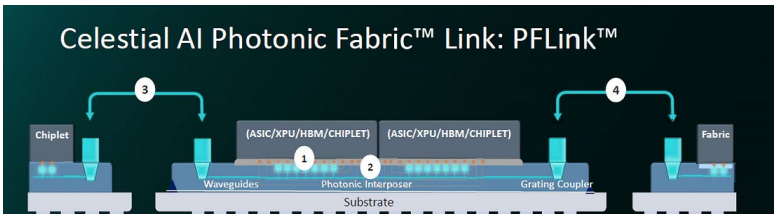
In practice, CXL memory expansion cards—such as Astera Labs' **Leo CXL Smart Memory Controller**—connect multiple DDR5 channels to the host through CXL, offering up to 2 TB of capacity and supporting both memory expansion and pooling. This approach can increase memory bandwidth and capacity for AI and database workloads by more than 50%. By enabling dynamic memory allocation, CXL helps eliminate underutilization and overprovisioning, reducing the constraints imposed by the memory wall on overall system performance.

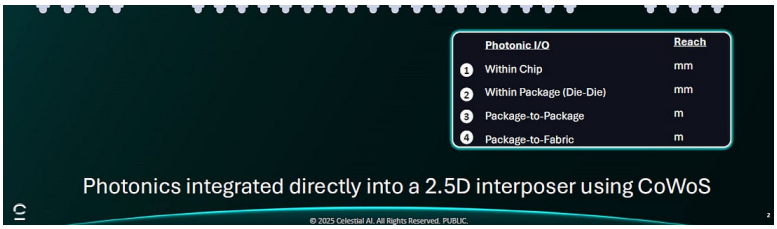
Solutions Combining Optical Chips and Memory



Celestial AI Photonic Fabric

Celestial AI's **Photonic Fabric** uses silicon photonics interfaces to connect compute and memory **within the package**, enabling data transmission at the speed of light. In 2024, the company announced a partnership with hyperscalers to deploy its photonic routing technology, which can deliver data directly to the compute point, supporting the bandwidth demands of **HBM3E and future HBM4**.





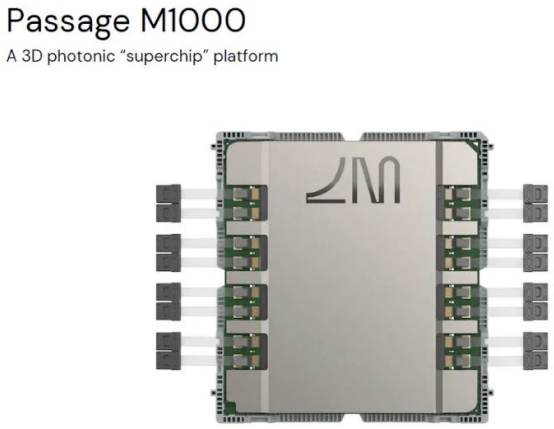
Its memory expansion module integrates **72 GB of HBM (two stacks)** and four DDR5 DIMMs, expandable to **2 TB**. The photonic chip provides **14.4 Tb/s (1.8 TB/s)** endpoint bandwidth with round-trip latency of about **120 ns**, achieving energy efficiency of **6.2 pJ/bit**, significantly lower than NVLink and other electrical interconnects.

Beyond NVLink: Celestial AI's Photonic Interconnect Leadership and Capital Strategy in the Trillion-Parameter AI Era
SEMIVISION · JULY 4, 2025
Read full story

Jim Elliott, VP of Samsung's Memory Business, noted that this technology can connect large shared memory at **HBM speed with nanosecond latency**, effectively eliminating the memory wall.

Lightmatter Passage M1000

In 2025, **Lightmatter** introduced the **Passage M1000**, a 3D photonic superchip. The M1000 is a **multi-die 3D photonic interposer** with a total area of over **4,000 mm²**, offering **114 Tb/s** of optical bandwidth. It enables electrical–optical I/O at any point within the package.



Using a **reconfigurable waveguide network** and **256 optical fibers**, it connects thousands of chips into a single domain, overcoming the limitations of chip shoreline I/O. Built on the **GF Fotonix** platform, the M1000 features **1024 electrical SerDes**, **56 Gbps NRZ modulation**, and **8-wavelength WDM** transmission.

This platform allows HBM stacks and compute dies to be integrated into a large composite package, providing **hundreds of Tb/s** of interconnect bandwidth through optical links — a powerful combination for training large AI models.

Lightmatter: Transforming AI Infrastructure with the Power of 3D Photonics

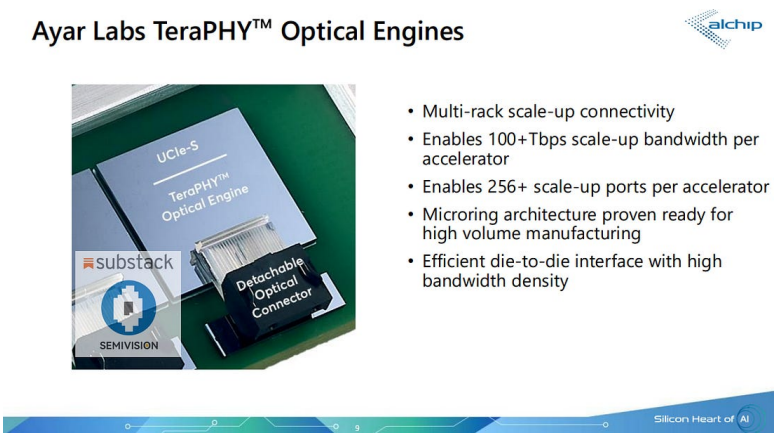
SEMIVISION · AUGUST 29, 2025

Read full story

Ayar Labs TeraPHY Optical I/O

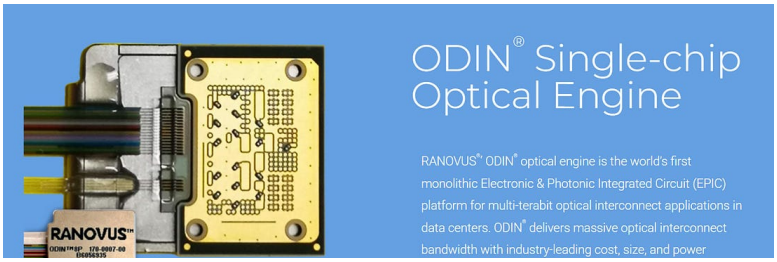
Ayar Labs' **TeraPHY** is a compact, low-power optical I/O chiplet that complies with the **UCIe packaging standard**. Each chiplet contains **8 full-duplex optical ports**, with each port consisting of **16 WDM transceivers**, providing **8 Tb/s bidirectional bandwidth** and only **~10 ns latency**.

Compared to traditional copper interconnects, TeraPHY offers **5–10× bandwidth**, **10× lower latency**, and **4–8× better energy efficiency**. When paired with the **SuperNova light source**, TeraPHY enables **millimeter-to-kilometer scale connections** between chips and memory, bringing multi-Tb/s interconnect capabilities to AI infrastructure.

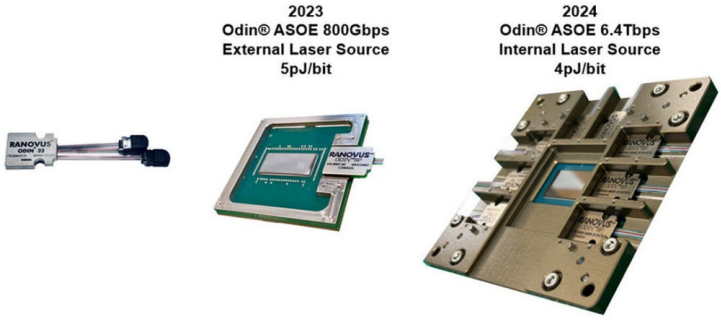


Ranovus (Odin) Multi-Wavelength Optical Platform

Ranovus' ODIN platform leverages a **single-chip electronic–photonic integrated circuit (EPIC)** that integrates **quantum dot multi-wavelength lasers** and **microring resonators**, delivering multi-Tb/s optical interconnects.



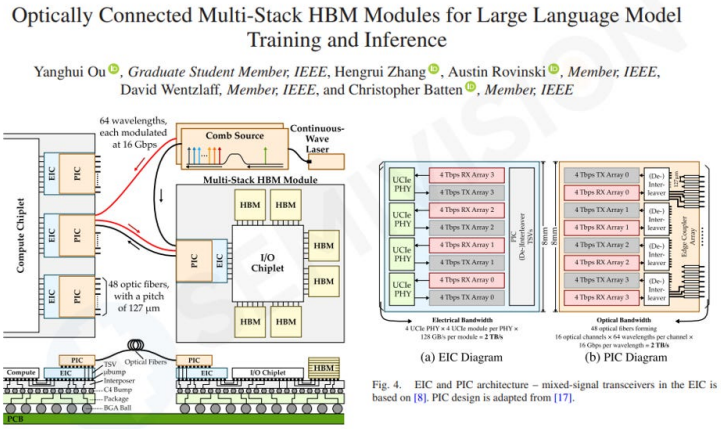
ODIN provides **high-density, low-power co-packaged optics** for AI data centers, supporting next-generation packages that combine XPUs and HBM. Through advanced packaging and co-packaged optics, ODIN enables **high-performance optical links between compute and memory**, addressing both the **memory wall** and **data center power consumption** challenges.



Academic Research: Optical Multi-Stacked HBM

Researchers from Cornell University and other institutions have proposed an optical interconnect architecture for multi-stacked HBM. By co-packaging photonic-electronic interface chips (EIC + PIC) with multiple HBM stacks and connecting them to compute chips via optical fibers, the architecture can scale HBM capacity to 576 GB with 12 TB/s of total bandwidth — all within the existing A100 package footprint.

Simulation results show this approach can **improve training efficiency for trillion-parameter LLMs by 1.4×** and **boost inference efficiency by 4.2×**. This demonstrates that **optical interconnects can break traditional package area constraints**, enabling more HBM stacks beyond the limits set by die perimeters.


Optically Connected Multi-Stack HBM Modules for Large Language Model Training and Inference

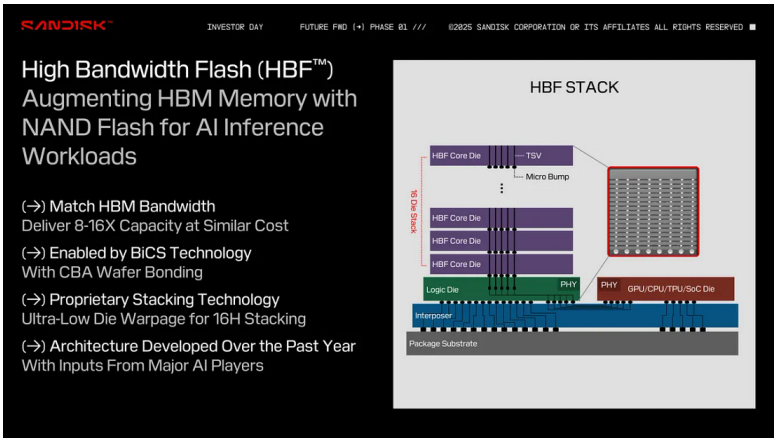
Yanghui Ou, *Graduate Student Member, IEEE*, Hengrui Zhang, Austin Rovinski, *Member, IEEE*, David Wentzlaff, *Member, IEEE*, and Christopher Batten, *Member, IEEE*

Fig. 4. EIC and PIC architecture – mixed-signal transceivers in the EIC is based on [8]. PIC design is adapted from [17].

Fig. 3. Example system architecture.

Will HBM Be Replaced by Flash? — The Potential of HBF/HFM

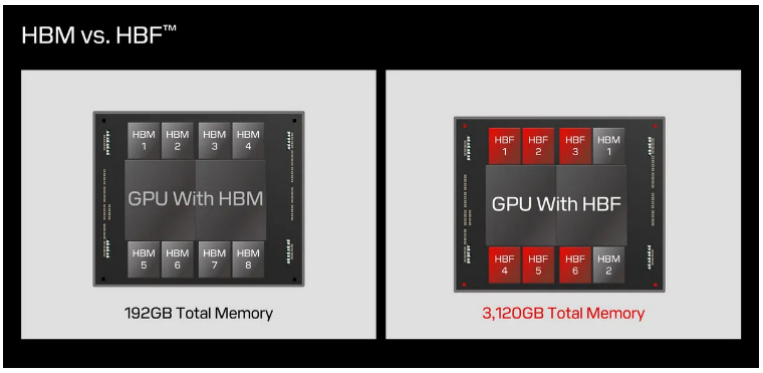
While HBM offers extremely high bandwidth, it is limited in capacity and expensive. To address the capacity issue, SanDisk and SK hynix signed an agreement in 2025 to jointly develop the **High-Bandwidth Flash (HBF)** standard. HBF leverages 3D NAND stacking combined with TSV vertical interconnects, enabling parallel access to multiple flash arrays under a logic control layer. Its design goal is to deliver bandwidth comparable to HBM while increasing capacity by 8–16× at a similar cost. SanDisk plans to release samples in the second half of 2026, with the first AI inference devices equipped with HBF expected to appear in 2027.



According to industry information, first-generation HBF stacks can provide **up to 4 TB of VRAM** on GPUs by parallelizing access across high-capacity 3D NAND arrays to sustain bandwidth. HBF is designed to complement, rather than replace, HBM—targeting **AI inference workloads** that are predominantly read-intensive. However, HBF still faces several challenges:

- **Higher latency** – Flash memory has much higher access latency than DRAM, making HBF unsuitable for training workloads that require nanosecond-level response times.
- **Limited endurance** – NAND flash has write cycle limitations, which may impact workloads with frequent updates.
- **Protocol differences** – Although HBF's mechanical and electrical interfaces resemble HBM, protocol modifications are still needed, meaning it is not fully compatible.

As a result, **HBF/HFM is more likely to serve as a complementary memory layer**, providing massive capacity for AI inference or cold data storage, rather than a full replacement for HBM.



SemiVision's Quick FAQ on HBM & Memory Technologies

Q1: How has HBM evolved across generations, and what are the key features?

A: HBM originated in 2013 with the release of the HBM1 standard by JEDEC. It uses stacked DRAM dies connected through TSVs and a 2.5D silicon interposer, delivering far higher bandwidth than GDDR or DDR. HBM2/2E introduced pseudo-channels and raised speeds to 8 Gb/s per pin. HBM3/3E expanded channels to 16 and supported 64 banks, enabling up to 24 GB per stack. HBM4, launching in 2025, doubles the channel count to 32, with 2,048 I/Os, reduces power by 40%, and adds features like Directed Refresh Management and bus remapping. Future HBM may move toward even taller stacks, higher bandwidth, and hybrid architectures with flash memory.

Q2: What is HBM IP and why is it needed?

A: HBM IP refers to the controller and PHY intellectual property provided by semiconductor IP vendors, allowing SoC designers to integrate HBM interfaces without in-house development. These IP blocks support the latest standards, offering high bandwidth, low latency, and robust data transfer. For example, Synopsys' HBM4 IP achieves 12 Gb/s per pin and over 3 TB/s total bandwidth. GUC's HBM4 IP uses proprietary interposer routing and I/O monitoring to ensure signal integrity. With HBM IP, chip designers can rapidly integrate high-bandwidth memory for AI GPUs, datacenter SoCs, and ASICs while reducing development risk and time.

Q3: What is the "Memory Wall" and why does it exist?

A: The Memory Wall refers to the growing gap between processor performance and memory bandwidth. Over the past two decades, GPU compute performance has grown roughly 60,000×, while DRAM bandwidth has only improved by about 100×. As a result, processors spend significant time idle, waiting for data. Causes include stalled DRAM scaling, limited I/O pins, signal integrity degradation, and the high cost of data movement versus computation. HBM alleviates this with high bandwidth, but stacking limits and supply constraints prevent it from fully meeting trillion-parameter

AI model demands.

Q4: How does CXL address memory allocation challenges?

A: CXL is a cache-coherent interconnect standard built on PCIe, enabling processors and devices to share memory. CXL.io handles discovery and management, CXL.cache allows devices to cache host memory, and CXL.mem exposes device memory (e.g., DDR5 or HBM) directly to the host. Starting with CXL 2.0, memory pooling and switching are supported—multiple CXL memory devices can form a shared pool for dynamic allocation across servers. For example, Astera Labs' Leo CXL controller aggregates multiple DDR5 channels into a single card with up to 2 TB capacity, supporting both expansion and pooling. This enables datacenters to scale memory on demand, reduce stranded resources, and ease memory wall constraints.

Q5: How can optical chiplets and photonics help overcome the Memory Wall?

A: Optical chiplets or photonic interposers use silicon photonics waveguides to link compute and memory modules, offering tens of times more bandwidth than electrical wires with lower latency. Celestial AI's Photonic Fabric treats HBM as a high-speed cache while connecting large DDR5 memory banks optically to compute dies—providing 1.8 TB/s bandwidth, 120 ns round-trip latency, and 6.2 pJ/bit energy. Lightmatter's Passage M1000 uses a 3D photonic interposer with 114 Tb/s bandwidth and 256 fibers, breaking the chip perimeter bottleneck. Ayar Labs' TeraPHY optical I/O chiplet provides 8 Tb/s per chip at ~10 ns latency, while Ranovus' ODIN platform uses quantum-dot lasers and microring resonators for multi-wavelength interconnect. These solutions extend HBM beyond the package or connect multiple memory modules, boosting bandwidth while lowering power and latency—effectively breaking the Memory Wall.

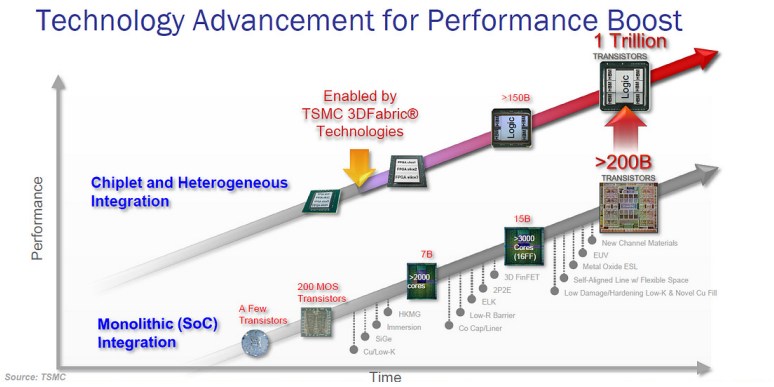
Q6: Will Flash replace HBM in the future?

A: To address HBM's limited capacity and high cost, SanDisk and SK hynix proposed the **High-Bandwidth Flash (HBF)** concept. HBF stacks multiple 3D NAND dies with TSVs, offering HBM-like bandwidth but 8–16× the capacity. First-generation products are expected to enable up to 4 TB of GPU VRAM. However, flash has higher latency and lower endurance, so HBF targets read-intensive AI inference workloads rather than replacing DRAM-based HBM entirely. Long-term roadmaps (e.g., HBM8) envision hybrid **HBM + HBF/HFM** architectures, where HBM acts as a low-latency cache and HBF provides large, cost-efficient capacity for cold data or inference.

Through **3D stacking and silicon interposers**, HBM has revolutionized memory bandwidth but remains limited by stack height and supply constraints. As AI models grow explosively, the Memory Wall has become a core system bottleneck. **CXL** tackles this at the system level with pooling and coherency, while **photonics** addresses it at the package level with

massive bandwidth and low latency. The future likely lies in a **multi-tier memory hierarchy** combining HBM, high-bandwidth flash, and optical interconnects—enabling AI systems to finally break through the Memory Wall.

Technology Advancement for Performance Boost in TSMC's viewpoints



The surging demand for AI chips is driving the entire semiconductor supply chain toward **larger areas, higher bandwidth, and greater power density**. For example, the die size of individual GPUs and AI ASICs is now approaching **800 mm²**, with final package modules often exceeding **100 × 100 mm**. This places unprecedented design pressure on both **interposers** and **ABF substrates**.

These high-power AI chips must simultaneously achieve **high-speed signal transmission**, **low-latency power delivery**, and **efficient thermal management** within the packaging stack.

- **Signal integrity:** To support hundreds of GB/s of chip-to-chip communication (e.g., GPU-to-HBM or GPU-to-GPU), the interposer requires ultra-dense RDL routing and stable electrical characteristics. TSMC's **CoWoS-L** and **CoWoS-R** have emerged as key solutions.
- **Power distribution:** To accommodate the growing number of power and high-speed signal layers, **ABF substrates** must increase their layer counts and improve dielectric performance, driving generational upgrades in **CCL (Copper Clad Laminate)** and PCB materials.

For example, to meet **224 Gb/s+ SerDes interfaces** and **4–8 TB/s HBM bandwidth** demands, CCL materials must simultaneously deliver **low Dk/Df**, **low surface roughness**, **high thermal stability**, and optimized **Z-axis expansion** to minimize loss and latency at high frequencies. However, these electrical enhancements also intensify **thermal pressure** — with AI chips now routinely exceeding **500–1000 W/cm² heat flux**, traditional TIM and air-cooling modules are no longer sufficient. The industry is rapidly shifting toward **liquid cooling**, **micro-channel cold plates (MCLP)**, **SiC**

thermal substrates, and **diamond thin-film heat spreaders**.

TSMC x Nvidia : Breaking the Thermal Wall: How Advanced Cooling Is Powering the Future of Computing

SEMIVISION · OCTOBER 5, 2025

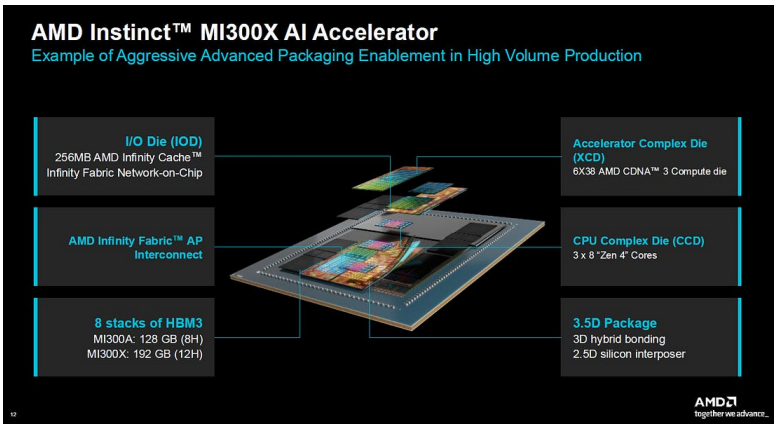
Read full story

Observing TSMC's SiC Strategy： SiC Enters the Advanced Packaging Mainstage

SEMIVISION · SEPTEMBER 21, 2025

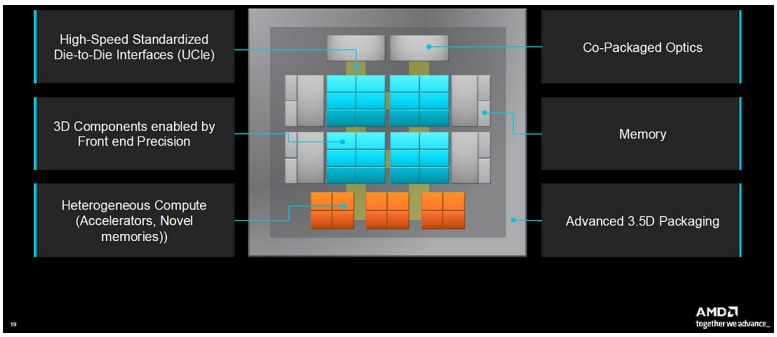
Read full story

As **packaging and thermal modules reach physical limits**, **memory placement and allocation (HBM Allocation)** has become the new bottleneck. Modern AI GPU modules typically integrate **6–8 HBM stacks**, but within the limited interposer area, simultaneously achieving optimal **signal path length**, **power distribution uniformity**, and **thermal balance** is increasingly difficult. SemiVision's earlier analysis shows that HBM performance bottlenecks often stem not from the DRAM itself, but from **RDL routing congestion** and **power noise distribution**.



To address this, TSMC and major GPU vendors are adopting **asymmetric HBM placement**, **localized power delivery networks (PDN)**, and **multi-layer RDL stack-up optimizations** to maximize bandwidth **without increasing interposer size**.

- **NVIDIA** uses a **split interposer** architecture in CoWoS-L packaging to keep HBM signal paths localized, reducing insertion loss from long traces.
- **AMD** and **Broadcom** are experimenting with integrating **Power Redistribution Networks (PRN)** into the substrate layer to further reduce high-frequency noise and IR drop.

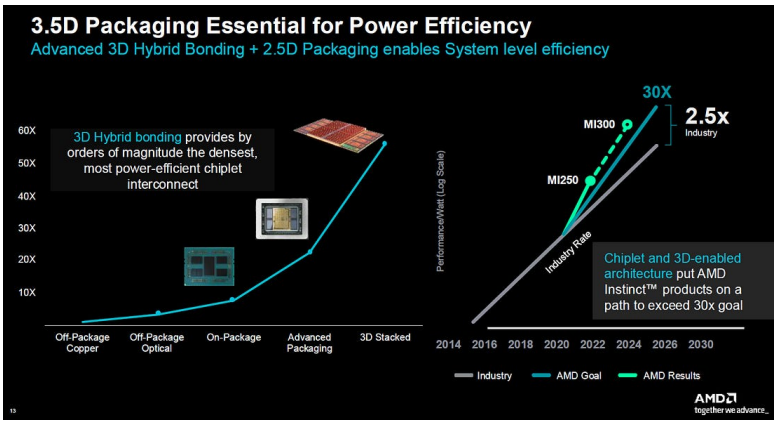


These design trends all point toward a critical shift: **AI chip performance gains no longer depend solely on logic scaling**, but increasingly on the **co-optimization of packaging, substrate, memory, and power delivery**.

In short, as **interposer area is constrained by reticle size and cost**, the key challenge for next-generation AI chip architecture is how to **maximize HBM bandwidth utilization and dataflow efficiency within a fixed package footprint**. This will not only determine the performance of individual GPUs but will also directly impact the **performance-per-watt** and **cost structure** of entire AI systems.

HBM Allocation and Packaging Space Optimization Strategies

As **interposer area** and **ABF substrate** designs approach their physical and cost limits, the ability to **maximize HBM bandwidth and power delivery efficiency within limited space** has become the next key battleground for advanced AI packaging.



In today's high-end GPUs, the logic die already occupies the majority of the interposer area. As a result, the **placement of HBM stacks (allocation)** must be precisely engineered to avoid routing congestion, increased latency, and reduced power efficiency.

1. Asymmetric HBM Placement

Traditionally, GPU modules adopt **symmetric HBM configurations** (e.g., 4+4 or 6+6 around the logic die) to enable uniform signal routing. However, as interposer area becomes constrained and **RDL routing pressure** increases, the industry is moving toward **asymmetric layouts**.

In these designs, **high-bandwidth and latency-sensitive HBM stacks** are placed closer to critical I/O hotspots (such as chiplet-to-chiplet SerDes regions), while less critical memory modules are allocated to secondary zones. This shortens critical signal paths and reduces insertion loss.

The main challenge lies in maintaining **uniform PDN (Power Delivery Network)** and **SI/PI (Signal/Power Integrity)**. TSMC has addressed this on its CoWoS-R/L platforms by introducing **segmented PDN architectures**, allowing each HBM region to have relatively independent power and decoupling networks. This minimizes local IR drop or power noise from impacting global performance.

2. Multi-layer RDL Stack-up and Routing Zoning

With HBM speeds reaching **9.2–12.8 Gbps per pin** (and potentially exceeding 16 Gbps/pin with HBM4), **single-layer RDL** can no longer accommodate the thousands of high-speed signal traces required.

TSMC has adopted **multi-layer RDL stack-ups** that:

- Place **high-speed differential pairs** in the inner layers.
- Use outer layers for power distribution and low-speed control signals.

This structure helps to:

- Reduce coupling and crosstalk for high-speed signals.
- Stabilize impedance and routing uniformity.
- Precisely control delay and jitter.
- Increase routing density without expanding interposer size.

Additionally, TSMC has introduced **Routing Zoning** in CoWoS-R, partitioning RDL layers into functional zones (HBM zone, SerDes zone, Clock/Power zone). This reduces trace length and crosstalk while enabling localized optimization.

3. SI / PI Simulation and Verification Flow

Such dense designs demand rigorous **electrical simulation and validation** from the early design stage, using full-wave 3D simulation tools (e.g., Ansys SIwave, HFSS, Cadence Sigrity) to analyze:

- HBM–SoC **S-parameters** (reflection/insertion loss).

- PDN noise suppression with decoupling capacitor configurations.
- Routing length impact on **latency and jitter**.
- Effectiveness of cross-layer grounding for EMI/crosstalk mitigation.

Following simulation, TSMC and customers use **CoWoS-R as the physical validation platform**, performing **eye diagram measurements**, **BER testing**, and **power noise spectral analysis** to ensure the design can sustain **high yield and reliability** in volume production.

4. System Technology Co-Optimization (STCO)

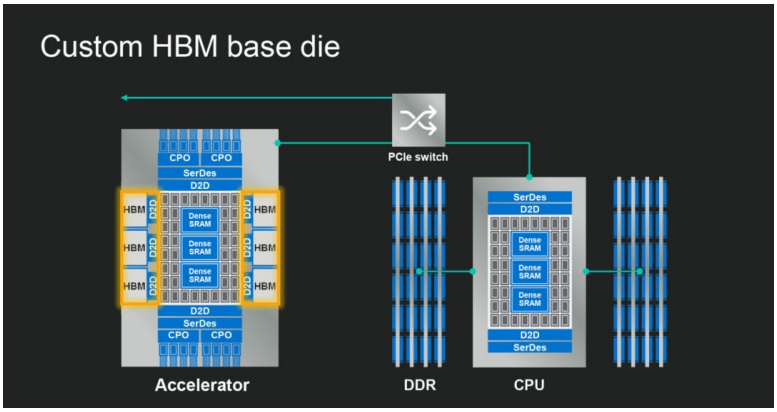
Ultimately, this is no longer the sole responsibility of packaging teams. It has evolved into **STCO (System Technology Co-Optimization)**, where **HBM placement**, **RDL routing**, **PDN distribution**, and **thermal structures** must be **co-optimized with the logic I/O architecture, bandwidth demands, and system-level cooling**.

This is why companies such as **TSMC, NVIDIA, and AMD** have established dedicated **Package Architecture Teams** to bridge front-end design, packaging, and system validation.

In the AI-driven era of **high-speed packaging**, **HBM allocation** has evolved from a simple stacking decision into a **multi-dimensional design challenge** involving electrical integrity, routing architecture, and system thermal management.

This will be a decisive front in the competition among **CPO/OIO, high-speed interconnect, and advanced packaging platforms**.

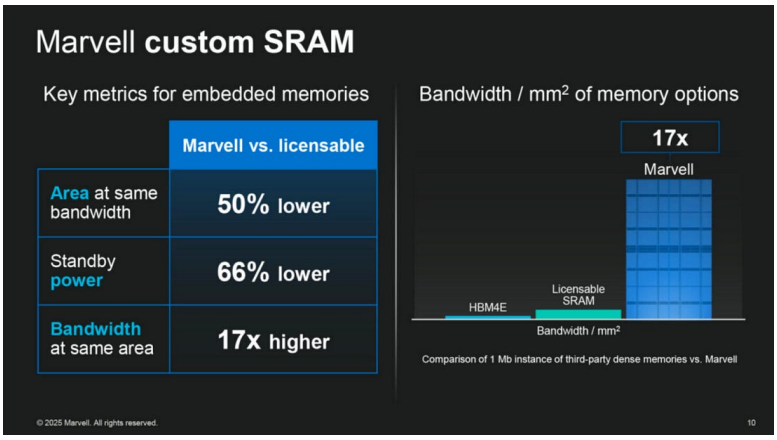
Notably, at **Hot Chips 2025**, Marvell VP of Technology **Mark Kuemerle** presented *"A Revolution in Memory Architecture for the Data Center"*, highlighting the company's innovations in **SRAM and HBM architectures**. Marvell's technical blogs further detailed their **custom HBM** and **CXL** strategies. Below is a summary comparing their perspectives on future memory architectures.



Marvell :SRAM — Boosting Bandwidth and Efficiency in AI/ XPU Devices

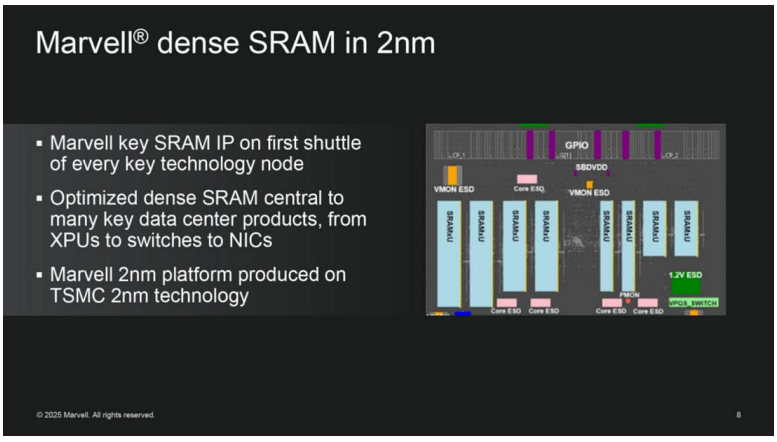
2 nm High-Density SRAM

Marvell's key SRAM IP is among the first to tape out at every major process node, providing developers with early hardware validation platforms. The company's high-density SRAM is manufactured on its 2 nm platform and deployed in XPUs, network switches, and NICs for data center applications.



Lowering Vmin and Enhancing Reliability

To reduce power consumption, Marvell optimized Vmin (minimum operating voltage) on its N2P hardware, achieving the lowest Vmin ever recorded for ultra-dense bitcell memory. This was accomplished through a combination of circuit design techniques including write-assist, stability-assist, high-sigma modeling, and row/column redundancy. These innovations not only reduce power but also improve manufacturing yield.



Benefits of Custom SRAM

Marvell compared its in-house 1 Mb high-density SRAM to third-party memory and found that, at equivalent bandwidth, its solution achieves

50% smaller area, **66% lower standby power**, and **17× higher bandwidth per unit area**. These advantages make SRAM an ideal embedded memory for AI accelerators and network switches.

To deliver such high bandwidth, Marvell uses three strategies: **increasing operating frequency**, **widening data access width**, and **adding more ports**. Experimental results show that, using the same memory array and bitcell, Marvell's SRAM achieves a significantly higher bandwidth-per-area ratio than industry standards across different capacities.

Silicon validation demonstrates that Marvell's custom N3P high-density SRAM maintains strong PVT (process, voltage, temperature) margins, proving its manufacturability.

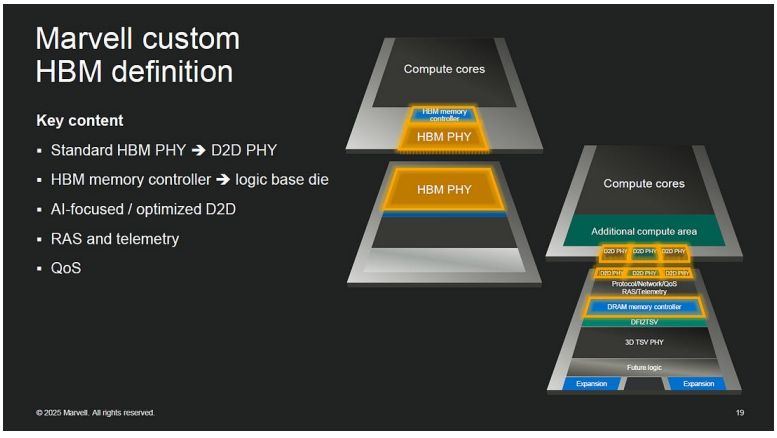
Marvell views high-density SRAM as a **foundational building block** for AI accelerators. By leveraging advanced 2 nm processes, low-Vmin circuit techniques, and superior bandwidth-to-area efficiency, Marvell's embedded SRAM can complement external memory like HBM, playing a critical role in high-performance computing architectures.

Marvell:Custom HBM — Tailoring Memory to Match XPU Requirements

Limitations of Standard HBM

Traditional High-Bandwidth Memory (HBM) consists of stacked DRAM dies and a base die that houses the I/O interface and memory controller. Standard HBM4 designs feature 32 × 64-bit channels, resulting in a large I/O footprint and high power consumption.
In XPU designs, HBM can consume up to **25% of die area** and account for **40% of total cost**. As memory capacity scales from **80 GB to 288 GB and beyond**, the power and area overhead of standard HBM becomes a major constraint.

Marvell's Custom HBM Architecture



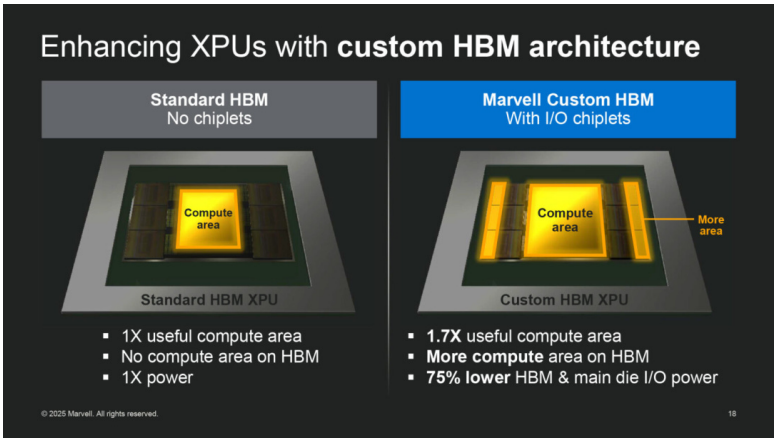
Concept: Marvell proposes a **"Custom HBM"** architecture that

redistributes traditional HBM functionality:

- Replaces the standard HBM PHY with a **die-to-die (D2D) PHY**, moving the HBM memory controller onto the **logic base die**.
- Adds **AI-specific D2D channels**, RAS/telemetry monitoring, and Quality-of-Service (QoS) modules on the logic base.
- This allows memory interfaces to be tailored to the XPU's specific compute characteristics, placing critical logic and transport closer to where it's most effective.

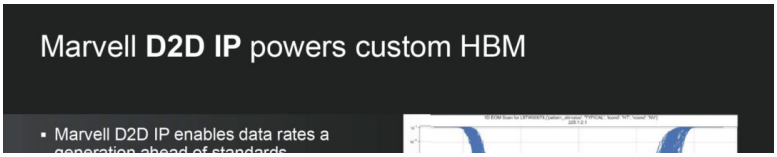
Benefits:

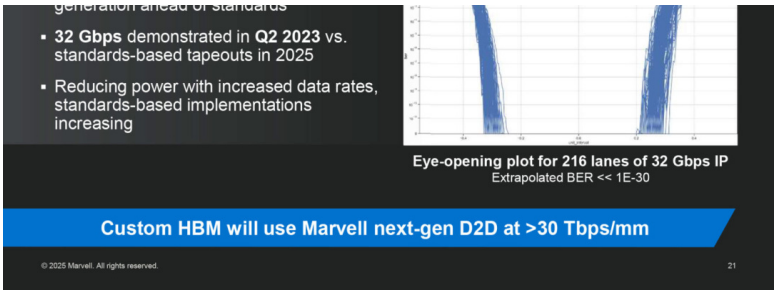
- **More usable compute area:** Standard HBM XPUs allocate a large portion of silicon to I/O. With Custom HBM, Marvell reports a **1.7× increase in usable compute area**, with the option to place compute chiplets directly on the HBM base for higher compute density.
- **Reduced I/O power:** By shifting interfaces to D2D links, I/O power between HBM and the main die can be cut by **~75%**.
- **Scalable memory–compute integration:** Mapping standard HBM channels to D2D links allows future expansion with additional memory, logic, or accelerator chiplets.



Example: Marvell notes in its blog that Custom HBM can **increase on-package memory by 33%**, **cut interface power by 70%**, and **free up 25% of chip area** for compute logic. Depending on the use case (e.g., edge inference vs. large-scale training), designers can adjust HBM controller and PHY sizing to reduce I/O footprint or boost channel count.

D2D Technology Support





Marvell emphasizes its D2D interface IP as the enabler of Custom HBM.

- In Q2 2023, Marvell demonstrated **32 Gbps** D2D transfer speeds—well ahead of industry standards scheduled for volume production in 2025.
- In August 2025, Marvell announced the **industry's first 64 Gbps/pin bidirectional D2D interface IP**, offering **>3× the bandwidth density** of UCIe, with intelligent power management to reduce power spikes under burst traffic.

These D2D advances underpin Custom HBM, enabling higher bandwidth within a smaller I/O footprint while lowering error rates and power consumption.

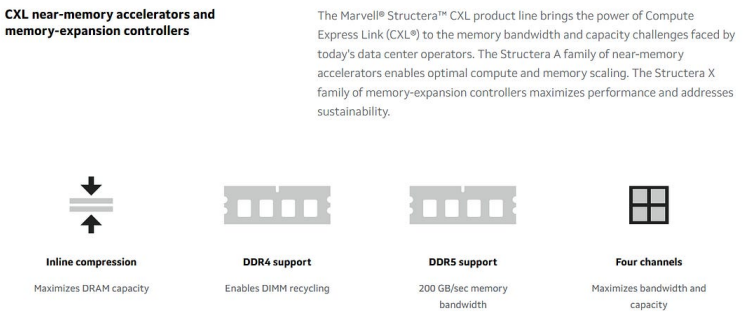
Comparison with Standard HBM
Marvell compared standard HBM with Custom HBM in its December 2024 blog, highlighting dramatic reductions in **interface area** and **power**, along with fewer **µ-bumps** and power/ground bumps required.

Custom HBM not only meets current **HBM4** specifications but also scales toward **HBM5** and **HBM-Next**, making it a key architectural direction for AI accelerators. Marvell argues that standard HBM is too rigid in terms of area and power, whereas Custom HBM—by **redesigning the base die and I/O interface**—enables tighter memory-compute integration, increased usable compute area, and lower power.

The **core enabler** is the high-speed D2D interface and the ability to **customize the memory controller** to fit specific XPU requirements. Marvell predicts that by **2028**, around **25% of accelerated compute chips** will adopt custom architectures, including **Custom HBM**, **custom XPUs**, and **CXL controllers**.

Marvell : CXL — Expanding and Sharing Memory Through Compute Express Link

CXL Overview and Principles
Marvell notes in its blog that since the invention of magnetic memory in the 1950s and DRAM in the 1970s, there have been few major breakthroughs in the memory field. **CXL (Compute Express Link)** represents the next major leap.



> **Marvell: CXL Near-Memory Compute and Expansion**

CXL devices use the existing PCIe physical interface to open a **parallel channel to the processor's memory bus**, providing additional lanes and higher data throughput to relieve the burden on the congested memory fabric.

Structera A: Near-Memory Accelerator

Near-memory accelerator (Structera A) — Marvell's Structera A is a new type of CXL near-memory accelerator that integrates **16 Arm Neoverse V2 cores**, delivering **200 GB/s of memory bandwidth** through **four DDR5-6400 channels**, each supporting two DIMMs.

Benefits:

- Plugging one Structera A into an x86 server increases **core count by 25%** (from 64 to 80), boosts **bandwidth by 50%** (400 Gbps → 600 Gbps), and expands **memory capacity from 2 TB to 6 TB**, with only a **100 W** power increase.
- Energy per GB/s of transfer drops by **17%**.
- Installing two accelerators increases core count by **50%**, doubles bandwidth, and boosts capacity **sixfold**.

Rack-level scaling:
Installing **40 Structera A units per rack** can add **3,840 processing cores** and **24 Tbps of memory bandwidth** to existing servers — **without expanding the data center footprint**, thereby reducing infrastructure and cooling costs.

Structera X: Memory Expansion and DDR4 Reuse

Memory Expansion Controller (Structera X) — CXL enables servers to attach external memory expansion devices. Structera X supports **DDR5 or DDR4**, with up to **200 GB/s bandwidth**. The DDR5 version supports **up to 3 DIMMs per channel**, for over **6 TB capacity**; the DDR4 version supports **3 DIMMs/channel**, enabling up to **4 TB capacity**.

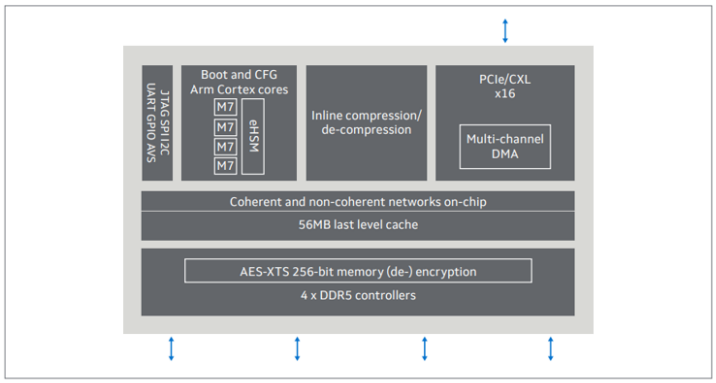
MARVELL

Product Brief

Structera™ X 2504 Memory-Expansion Controller

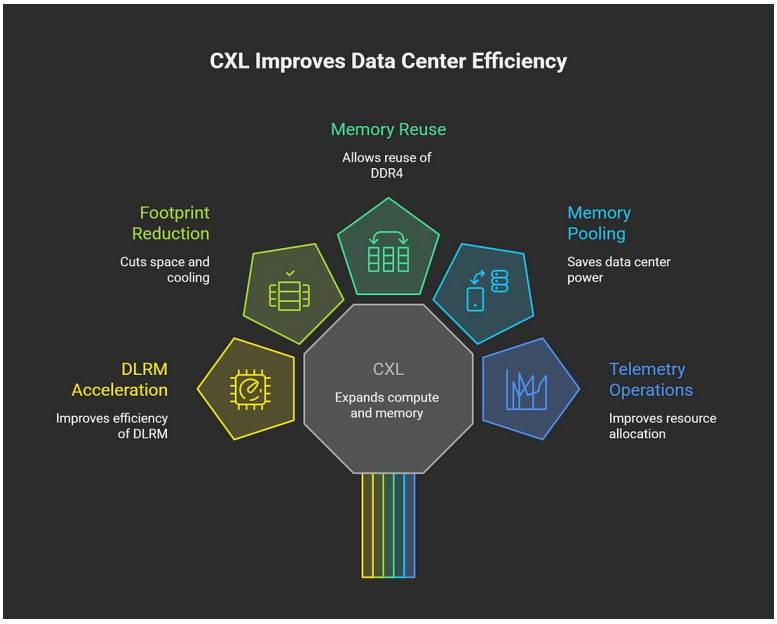
CXL 2.0 DDR5 4-channel expander
P/N MV-SLX25041-A0-HF350AA-C000

Block Diagram



DDR4 reuse example:
Next-generation servers may only support DDR5, but there remains a large pool of **decommissioned DDR4 DIMMs**. Structera X can connect **12 DDR4 DIMMs** through CXL, offering **6 TB capacity**, or **up to 12 TB** with LZ4 compression. This reduces costs, extends the life of existing memory, and cuts electronic waste.

Five Key Impacts of CXL



1. **Accelerating DLRM and Inference Workloads**
   CXL alleviates memory bandwidth bottlenecks. Structera A provides extra cores and high bandwidth near memory, improving the efficiency

of **Deep Learning Recommendation Models (DLRM)** and other inference tasks.

2. **Reducing Data Center Footprint and Infrastructure Costs**
   By expanding compute and memory through CXL, **one rack** can achieve the capacity of many servers, cutting space and cooling demands.
3. **Reusing Idle Memory**
   CXL enables the reuse of DDR4 memory through Structera X, allowing 12 DIMMs to be configured per server. Compression further boosts capacity, reducing the need for expensive DDR5 procurement.
4. **Improving Asset Utilization via Memory Pooling**
   CXL allows **two processors to share a single Structera X memory expansion**, forming a **memory pool**. Microsoft estimates **~25% of server memory sits idle** because it is tied to a single CPU; CXL pooling can **save 5–9 TWh** of data center power annually by reducing this stranded capacity.
5. **Telemetry-Driven Operations**
   Future CXL devices will include telemetry to monitor memory pools and distributed systems, improving **resource allocation** and **predictive maintenance**.

Marvell's Perspective
Marvell sees **CXL as a key technology** to overcome both memory bandwidth and capacity bottlenecks.

- **Structera A** offers near-memory compute and bandwidth for AI inference workloads.
- **Structera X** provides flexible memory expansion, supporting DDR5 while enabling cost-effective DDR4 reuse.
- CXL's **memory pooling and telemetry capabilities** reduce idle resources and improve sustainability.

Together, these features enable **more scalable, efficient, and flexible AI infrastructure**, marking a major architectural shift in data center memory systems.

Integrated Perspective: Marvell's Memory Strategy

From Marvell's presentations and blog posts, the company's **memory architecture strategy** can be summarized into the following key pillars:

1. Optimizing Every Tier of the Memory Hierarchy

- **On-chip:**
  Marvell leverages **2 nm high-density SRAM** combined with advanced circuit techniques to lower power and area while delivering ultra-high bandwidth. These SRAM blocks act as **caches or scratchpads for XPUs**, improving core utilization and efficiency.

- **Stacked-chip level:**
  Marvell proposes **Custom HBM**, which re-architects the HBM base die and I/O interface and pairs it with Marvell's proprietary **die-to-die (D2D)** technology. This approach frees up compute area, reduces power consumption, and supports future HBM standards, enabling **more flexible and application-tailored AI accelerators**.
- **System level:**
  Through **CXL**, Marvell enables **near-memory acceleration and memory expansion**, using its **Structera A/X** products to support various workloads such as inference, training, and databases. At the same time, **memory pooling and flexible resource allocation** are achieved across servers.

2. Lowering Total Cost of Ownership (TCO)

- By reducing **SRAM Vmin**, cutting **I/O power in custom HBM**, and enabling **memory reuse through CXL**, Marvell significantly reduces overall power consumption, contributing to lower **data center TCO**.
- CXL expands **core counts and memory capacity** without requiring new physical infrastructure. Reusing DDR4 DIMMs through Structera X further saves on **procurement and disposal costs**.

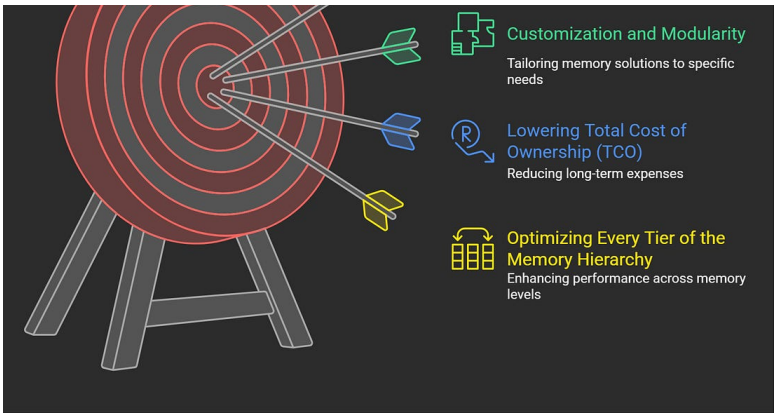
3. Customization and Modularity

- Marvell anticipates a future with **greater customization across XPUs, HBM, CXL controllers, and NICs**. Custom HBM is part of this trend, allowing chip designs to be **fine-tuned to specific application needs**.
- By leveraging **chiplet and D2D technologies**, compute and memory can be modularized for **easier upgrades, flexible scaling**, and optimized trade-offs between power, area, and performance.

4. Memory-Centric Data Center Evolution

Marvell's combination of **high-density SRAM**, **custom HBM**, and **CXL architectures** outlines a **memory-first roadmap** for data center evolution:

- from **bitcell innovations at the silicon level**,
- to **interface customization in stacked memory**,
- to **cross-server memory sharing at the system level**.



Marvell's message is clear: **"Memory is the only thing that matters."** This comprehensive optimization strategy not only addresses AI accelerators' demand for **high-bandwidth, low-power memory**, but also reduces data center energy use and cost, underscoring the **central role of memory architecture innovation** in the future of computing infrastructure.

From Custom SRAM to Optical SerDes: How Marvell Builds the Data Highways for AI Chips

SEMIVISION · JULY 9, 2025

Read full story

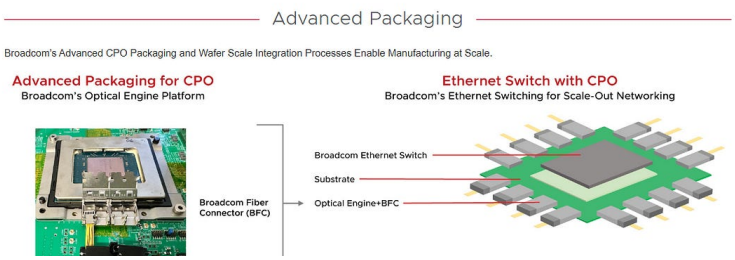
We observe that **Broadcom**, **Marvell Technology**, and multiple AI chip suppliers are actively helping major **cloud service providers (CSPs)** adopt and integrate **advanced packaging technologies**. This effort has evolved far beyond traditional packaging manufacturing—it is now transforming into **full system-level platform collaboration**.

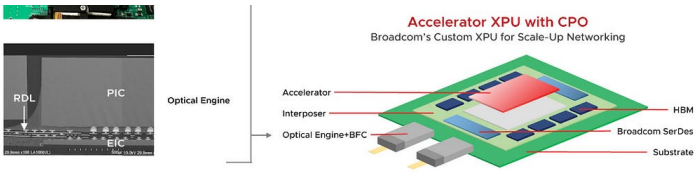
Broadcom's CPO Strategy and Its Implications for the Future of Optical Interconnects

SEMIVISION · JUNE 4, 2025

Read full story

These semiconductor vendors are not only supporting customers in **chiplet integration** at the packaging level, but are also **co-developing customized memory subsystems**, such as **custom HBM base dies**, **custom SRAM**, and **high-speed cache architectures**. These tailored designs optimize **power consumption, bandwidth, and latency**, aligning precisely with the **diverse architectural requirements** of different cloud computing platforms.





More strategically, **Broadcom and Marvell have incorporated Optical I/O technologies** into their **Advanced Package Platform roadmaps**, enabling **higher-bandwidth and lower-latency** data transmission between **AI chips, memory, switches, and accelerators**. This marks a fundamental shift: advanced packaging is no longer just a physical integration vehicle for a single chip—it is becoming the **interconnect backbone of high-performance computing systems**, directly determining the **performance and energy efficiency** of AI data centers.

Overall, this **joint development model of "customized memory + packaging + optical interconnect"** signals that the industry is rapidly moving from **single-chip competition** toward **platform-level strategic integration**. It also reflects the **deepening partnerships between CSPs and the semiconductor supply chain**, which will shape the next generation of AI infrastructure.

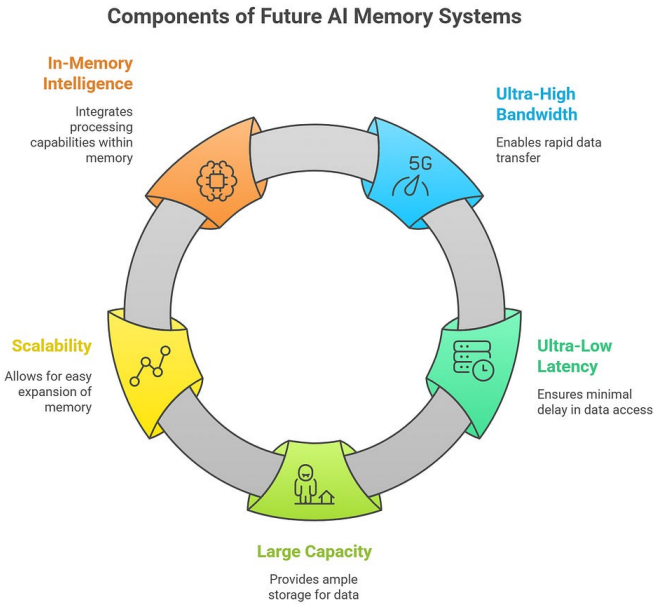
Aspect	Trend	Rationale / Driving Factors
Higher Bandwidth	Memory bandwidth is becoming a critical bottleneck	As model size and parameter counts grow, Transformer-based architectures require intensive parameter read/write and activation exchanges during forward and backward propagation. Insufficient bandwidth slows overall throughput.
Lower Latency	Certain operations demand extremely low access latency (e.g., control paths, weight indexing, sparse access)	Even with large bandwidth, excessive latency can stall critical steps, reducing performance.
Larger Capacity / Multi-tiered Memory	Growing model sizes and longer context windows require more memory for KV caches, activations, and intermediate states	Models such as long-context LLMs, graph neural networks, and reinforcement learning workloads need to store large amounts of intermediate data.
Scalable Memory	On-chip or in-package memory is limited, requiring external memory expansion	Memory pooling and distributed memory architectures can extend capacity beyond the package.
Memory–Compute Integration (Near- / In-Memory Computing)	Reduce data movement latency and energy by performing certain computations near or inside the memory	Logic units, compression/quantization modules, or accelerators can be embedded directly in SRAM/DRAM.
Efficiency / Power Optimization	Memory power and energy efficiency are becoming critical under high bandwidth and throughput demands	If high-bandwidth memory consumes too much power, the overall system may exceed thermal or energy limits.

In addition, some studies have specifically addressed the **bandwidth bottlenecks of CXL-based memory in LLM inference**, proposing a **CXL-NDP (Near-Data Processing)** architecture. This approach performs operations such as **compression, quantization, and bit-plane layout processing** directly inside the CXL memory module, thereby improving the utilization of "effective bandwidth."

Experiments showed that this method can **increase throughput by 43%** and **extend context length by 87%**. This clearly reflects that future AI memory requirements will not only depend on **raw bandwidth and capacity**, but also on **intelligent processing at the memory-module level** to enhance efficiency.

In summary: Future AI chips will demand memory systems that combine

ultra-high bandwidth, **ultra-low latency**, **large capacity**, **scalability**, and **in-memory intelligence**.



Amid the rise of generative AI, the choice of compute architecture is undergoing a profound transformation.

Traditionally, **CPUs** have handled general-purpose control and logic processing, making them suitable for multitasking and unstructured workloads. **GPUs**, originally designed for graphics rendering, excel at parallel computing. With the advent of the CUDA platform and the popularization of deep learning, GPUs have become the **de facto core for AI training**. **FPGAs** offer reconfigurable logic flexibility and can be reprogrammed after manufacturing, but they come with higher development complexity and cost.

ASICs (Application-Specific Integrated Circuits) represent a different direction—**sacrificing flexibility for extreme efficiency**. ASICs are designed for specific algorithms and applications, implementing only the necessary circuitry to achieve their functions. As a result, they offer significant advantages in **power consumption and cost**.

As AI models grow larger and **inference power consumption and latency become bottlenecks**, more companies are turning to ASIC architectures to achieve **higher energy efficiency and predictable deployment costs**. The GPU-dominated compute revolution is gradually evolving into a **"GPU + ASIC coexistence" era**.

Advantages and Limitations of ASICs in AI Inference

AI workloads can be broadly divided into **training** and **inference** phases.

- **Training** emphasizes flexibility and generality. GPUs dominate this phase because they support a wide range of deep learning frameworks and dynamic model development.
- **Inference**, in contrast, involves fixed model architectures, and the main challenge is to **achieve higher throughput with lower power and cost**—precisely where ASICs excel.

The greatest strength of ASICs is **efficiency**. Because their architecture is tailored to specific algorithms or models, ASICs can **eliminate unnecessary general-purpose modules**, shorten data paths, and minimize power usage. Once mass production begins, their **unit cost can be lower than GPUs or FPGAs**. Moreover, ASICs' tightly controlled and closed circuitry gives them **advantages in hardware security and operational stability**.

However, ASICs also have **notable drawbacks**.

- Their **development cost is extremely high**—including design, EDA licensing, IP procurement, and tape-out—often reaching tens or even hundreds of millions of NT dollars.
- They lack flexibility; if the underlying AI model changes, the ASIC may not support new algorithms.
- The **rapid iteration cycle of AI** means ASIC development timelines can lag behind market needs, requiring careful evaluation of **application lifetime and return on investment** before committing.

As a result, ASICs are **not intended to replace GPUs**, but to **complement them** by providing optimized solutions for **specific, stable, and large-scale applications**. For workloads such as **recommendation systems, speech recognition, or large language model inference clusters**, ASICs can deliver significant benefits.

Comparison and Complementarity of Compute Architectures

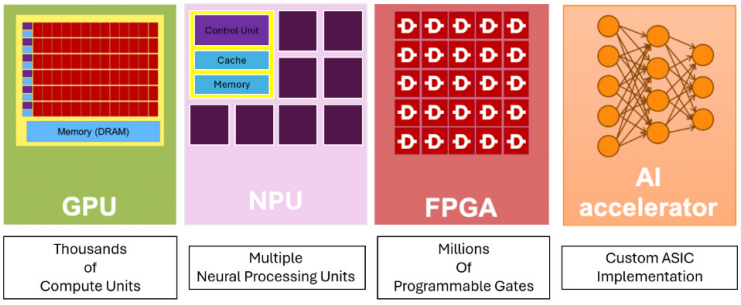
As compute architectures evolve, a **complementary landscape** has emerged among **GPUs**, **ASICs**, and **next-generation high-bandwidth accelerators** (such as LPUs or specialized memory-processing chips).

- **GPUs** remain the dominant platform due to their **flexibility and general-purpose programmability**, making them well-suited for a wide range of deep learning frameworks and precision modes.
- **ASICs**, on the other hand, excel in **low-precision inference**, **power-constrained edge deployments**, or **large-scale rollouts**, offering superior energy efficiency in these targeted use cases.

Modern GPU systems are also rapidly advancing. For example, **NVIDIA's NVLink technology** provides up to **1.8 TB/s of chip-to-chip interconnect**

bandwidth, dramatically reducing communication latency in multi-GPU systems. This keeps GPUs at the forefront for large-scale training and distributed inference workloads.

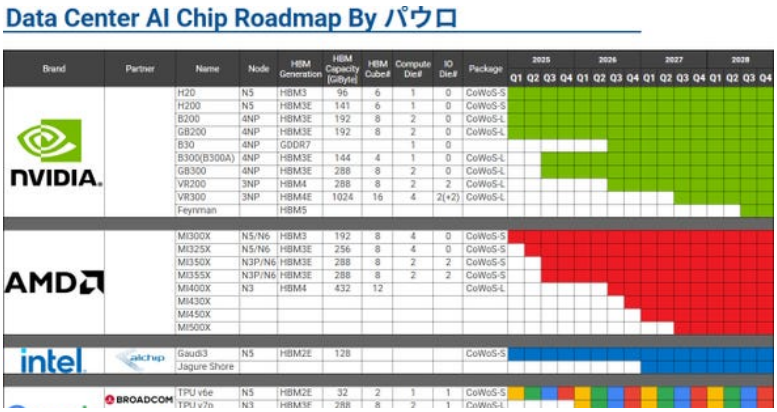
In practice, many companies are adopting a **"GPU for training, ASIC for inference"** dual-track strategy, balancing flexibility and efficiency. This **hybrid deployment model** is increasingly becoming the mainstream approach for both **cloud and edge AI**.

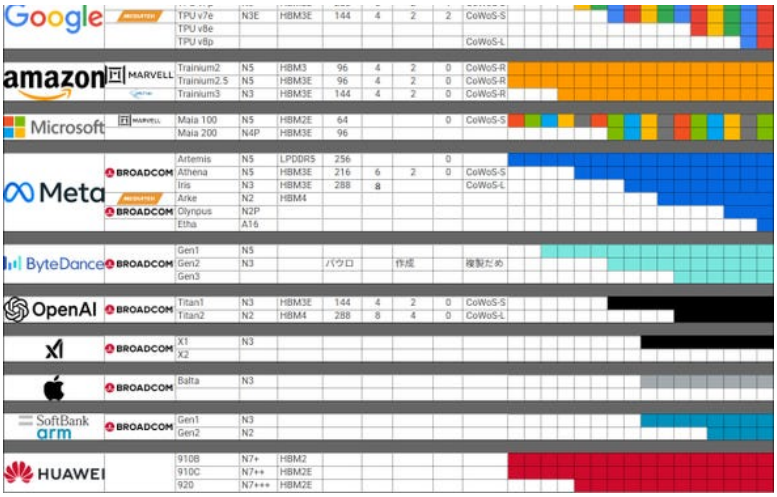


Criteria	CPUs	GPUs	FPGAs	ASICs
Processing Peak Power	Moderate	High	Very High	Highest
Power Consumption	High	Very High	Very Low	Low
Flexibility	Highest	Medium	Very High	Lowest
Training	Poor at training	The only production-ready training hardware	Not efficient	Potentially best for training, but not available yet
Inference	Poor at inference, but sometimes free	Average for inference	Best for inference	Not inference focused

Why are major tech giants investing in custom ASICs?

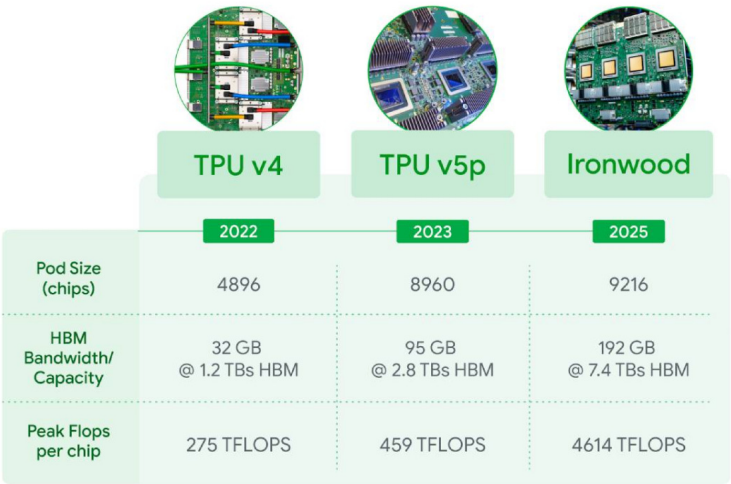
As the **cost structure of AI** shifts, major cloud and device companies—including **Google, Amazon, Meta, Tesla, Alibaba, and Huawei**—are all developing their own ASIC chips. This is driven not only by technical optimization but also by **cost control and strategic business considerations**.





The traditional chip supply chain involves **high licensing, manufacturing, and marketing costs**. When AI deployments scale to **hundreds of thousands of GPUs**, even minor efficiency gains can translate into massive operational savings.

- **Google's TPU**, **Amazon's Inferentia**, **Tesla's Dojo**, **Alibaba's Hanguang**, and **Huawei's Ascend** chips are all designed for specific AI workloads, reducing power consumption and data movement latency while improving performance-per-watt.



Furthermore, **inference is increasingly becoming the primary source of AI revenue**. Industry data shows that **inference-related revenue now accounts for over 40% of NVIDIA's data center revenue**. This gives companies a strong incentive to **develop in-house ASICs**, reducing long-term dependency on GPUs and mitigating supply chain and cost risks.

Taiwan and China's Deployment in the AI ASIC Sector

Key Players in the Asian Supply Chain.
In Taiwan, **Alchip** has built a strong position in the high-performance ASIC segment through its expertise in custom AI SoCs, 3DIC, and **hybrid bonding** design flows. **GUC (Global Unichip)** leverages its close partnership with TSMC to gain a competitive edge in advanced-node SoC integration, becoming a key partner for global chip design and manufacturing.

Broadcom, with its strength in networking and interface IP, has solidified its leadership in AI data center interconnect chips. **Marvell** focuses on high-bandwidth, low-latency ASIC products by combining HBM and **co-packaged optics (CPO)** in its designs.

On the China side, companies such as **Alibaba**, **Cambricon**, and **Huawei** continue to invest in in-house AI chip development, gradually expanding their reach into inference and edge applications.

These developments indicate that competition in AI ASICs has transcended individual companies — it is now a contest of **entire supply chain integration and process ecosystems**.

ASIC WAR : The Race Toward Specialization and Systemization

The rise of ASICs signals that AI computing has entered an era of specialization and system-level optimization. The competition is no longer just about transistor density or process nodes — it's about who can achieve system-level optimization across memory bandwidth, interconnect efficiency, and packaging integration.

GPUs will continue to dominate the general training market, but ASICs will play a critical complementary role in large-scale inference and specialized workloads. Future AI infrastructure is likely to be a **heterogeneous ecosystem** comprising GPUs, ASICs, optical interconnect modules, and **CXL-based memory architectures**.

For chip design companies, entering the ASIC domain requires not only co-design capabilities between circuits and algorithms, but also the ability to navigate challenges across **packaging, process technologies, EDA, IP, and interconnect protocols**. Only companies with **system thinking** can stand out in this next-generation compute race.

The Impact of the AI Chip Era on Memory and HBM

Architectural Differentiation of AI Chips: The Prelude to the Memory–Compute Bottleneck

CPU, GPU, FPGA, and ASIC Architectural Roles

CPUs and GPUs are mostly based on the **von Neumann architecture**, where computation and memory share a single bus for data transfer. This provides high generality but relatively low efficiency.

FPGAs and ASICs, on the other hand, typically adopt **Harvard architectures or custom designs**, bringing compute units closer to memory to optimize power consumption and performance for specific tasks.

GPUs leverage massive parallel processing units combined with **HBM memory** and high-bandwidth interconnects such as **NVLink** to deliver exceptional training performance. NVLink 5.0 provides up to **1.8 TB/s per card**, while PCIe 5.0 ×16 is limited to about **128 GB/s**.

ASICs are optimized for specific workloads (e.g., matrix multiplication) through dedicated hardware pipelines. They deliver **higher energy efficiency and lower latency** than GPUs, with lower power and cost — but they lack generality and involve high development overhead.

Memory Requirements for AI Training and Inference

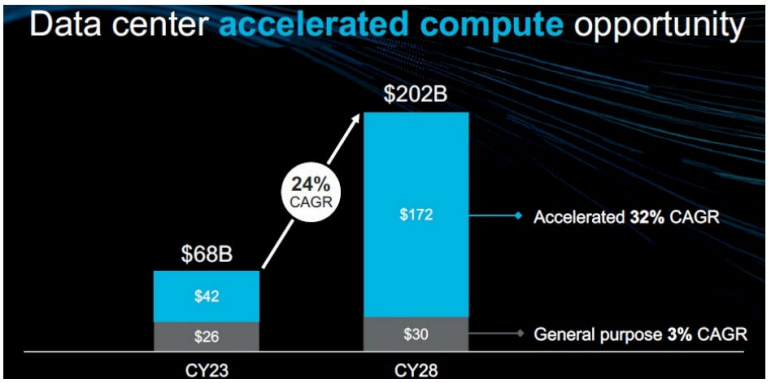
Large model training is often constrained by **memory bandwidth**, not compute power. Transformer attention mechanisms require storing the relationships between all tokens, with memory requirements growing linearly with sequence length. During inference, large **Key-Value (KV) caches** must be maintained.

To reduce training and inference latency, both **memory capacity and bandwidth per accelerator** must be expanded. Upgrading to newer HBM generations (e.g., HBM3 → HBM3E) or increasing the number of HBM stacks can **boost inference performance for models like Llama 2 and GPT-3 by 40 %–90 %**, without changing core counts or clock speeds.

Large-Scale Clusters and CSP In-House ASIC Development

AI training and inference clusters continue to scale aggressively. For example, Meta built **two 24,000-H100 GPU clusters in 2024**, and xAI plans to deploy a **100,000-H100 cluster**. Stacking massive numbers of GPUs or accelerators drives unprecedented VRAM demand.

To **reduce hardware cost and power consumption**, cloud service providers (CSPs) are developing their own AI ASICs, such as **Google TPU**, **Microsoft Maia 100**, and **Amazon Inferentia**. Although individual ASIC cards may have lower peak compute than an H100, **they offer superior energy efficiency and TFLOPS-per-dollar** for inference workloads. Furthermore, ASIC clusters can be paired with **custom software stacks and interconnect architectures**, potentially achieving higher overall efficiency than GPU clusters.

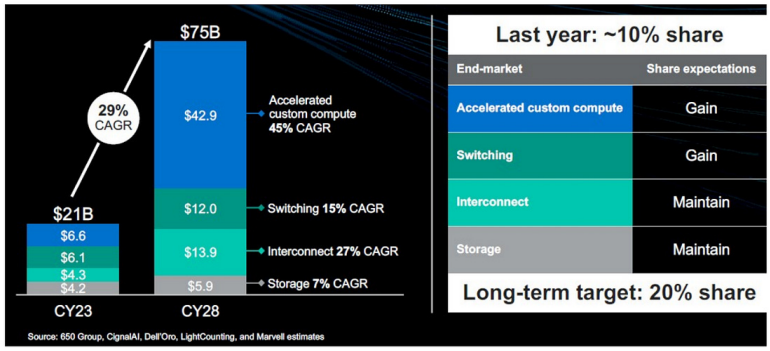


Marvell and other chip design service firms forecast that custom AI accelerators will grow from 16 % of the accelerator market in 2023 to 25 % by 2028, reaching a market size of $42.9 billion, with a 45 % CAGR over five years.

The overall data center silicon market is expected to expand from **$21 billion in 2023 to $75 billion by 2028**, representing a **29 % CAGR**. Within this,

- **Accelerated custom compute** grows from **$6.6 billion → $42.9 billion** (45 % CAGR),
- **Switching** expands from **$6.1 billion → $12.0 billion** (15 % CAGR),
- **Interconnect** grows from **$4.3 billion → $13.9 billion** (27 % CAGR), and
- **Storage** grows from **$4.2 billion → $5.9 billion** (7 % CAGR).

McKinsey further estimates that by 2025, **ASICs will account for 40 % of data center inference and 50 % of training workloads, and up to 70 % at the edge**. Last year, accelerated custom compute accounted for roughly **10 % market share**, with expectations to **gain share over the long term toward a 20 % target**.



HBM: The Key Memory Technology Behind Generative AI

1. Basic Structure and Advantages of HBM

HBM (High Bandwidth Memory) is built by stacking multiple DRAM dies on top of a logic die and connecting them through Through-Silicon Vias (TSVs). Its advantages include:

- **Ultra-high bandwidth**: HBM provides TB/s-level memory bandwidth via a wide data bus—over 20× faster than conventional DDR. Because the memory stack is located close to the logic die rather than on a distant DIMM module, the data transmission distance is much shorter, significantly reducing latency and energy consumption.

- **Power and area efficiency**: By stacking DRAM vertically beside the processor, HBM shortens data paths and achieves much higher memory capacity per unit area.

For each new generation of GPUs or ASICs, the critical performance improvement often comes from increased HBM capacity. For example, moving from Nvidia H100 to H200 or B200 to B300, the core count and frequency remain unchanged, but the HBM stack increases by 50 %, and the memory generation advances from HBM3 to HBM3E.

2.HBM Manufacturing and Competitive Landscape

- **Process differences**:
  SK hynix uses advanced **MR-MUF** (Molded Reflow Underfill) technology, filling the gaps between stacked dies with liquid epoxy molding compound at room temperature, which improves yield and thermal performance. SK hynix currently holds over 60 % of the HBM market.
  In contrast, Samsung and Micron use **TC-NCF** (Thermo-Compression Non-Conductive Film), which requires 300 °C processing and higher pressure, increasing the risk of die warpage.
- **Process nodes and base dies**:
  HBM4 will use **1b nm or 1c nm DRAM** dies and **4 nm logic base dies**, which handle stack control, I/O interface management, and custom logic integration. If 1c nm development lags, SK hynix—with its 1b nm yield advantage—could further widen its lead.
- **Hybrid bonding**:
  Cu-Cu hybrid bonding enables sub-10 µm pitch, lower resistance, and improved thermal performance. However, the required equipment and cleanroom investments are extremely costly. Large-scale adoption is expected with **HBM5 20-Hi stacks around 2028–2029**.
- **China's catch-up**:
  CXMT (ChangXin Memory Technologies) began HBM2 mass production in 2024 and is developing HBM3, with plans to launch HBM3E by 2027—shortening the technology gap to about four years.

3. HBM Supply-Demand Outlook and Market Size

- **Supply and demand**:
  Global HBM demand is estimated at **1,128.6 million GB in 2024** and **2,293.0 million GB in 2025**, versus supply of **1,068.1 million GB** and **2,210.9 million GB**, leaving gaps of about **5.4 %** and **3.6 %**, respectively.
- **Shift to new generations**:
  HBM3/3E's production share will jump from 33 % in 2023 to 81 % in 2024 and reach 89 % by 2025, signaling the rapid phase-out of older generations.
  Nvidia and SK hynix signed a **priority supply agreement** for HBM3E. Mass production of 12-high HBM3E started in Q3 2024, with first-

generation 12-high HBM4 entering production in 2025.

- **Market scale**:
  TSMC's CoWoS output in 2024 is around **252,000 wafers**, while global HBM demand is approximately **40 million units**, representing a market worth **US $9.14 billion**. Nvidia accounts for 58 % of demand, Google TPU for 15 %, AMD for 14 %, and Chinese companies for about 7 %.

Why Cloud Service Providers Develop Their Own AI ASICs

Cloud service providers (CSPs) face two major cost pressures:

1. The high price of premium GPUs like the H100.
2. The massive power consumption required for large-scale model training and inference.

ASICs offer a way out: by tailoring the hardware to specific workloads (e.g., matrix multiplication, tensor operations), they achieve higher performance-per-watt than GPUs.

Although North American cloud players' AI ASICs typically offer **lower peak compute than Nvidia's H100**, their cost and power advantages result in better overall **performance-per-dollar**. For example, **Google's TPU v6** and **Microsoft's Maia 100** reach **90 % and 80 % of H100's non-sparse performance**, respectively.

For **inference**, power, latency, and cost are decisive factors. CSET's analysis shows that ASICs outperform CPUs by **100–1000× in efficiency** for inference tasks.
McKinsey forecasts that by 2025, ASICs will account for **40 % of inference** and **50 % of training workloads** in data centers, indicating large-scale adoption of custom accelerators by major CSPs.

ASIC Growth Drives HBM Demand

Even though ASICs improve energy efficiency, their **high compute throughput still requires high-bandwidth memory**. Chips like TPU and Maia typically integrate HBM stacks to achieve **multi-TB/s bandwidth**, paired with high-efficiency interconnects to minimize data stall time.

Companies like **Marvell** are re-architecting **HBM I/O and co-packaged optics (CPO)** to reduce energy and support more stacked HBM per die. Their innovative design moves the HBM interface circuitry from the XPU die edge to the **base die under the stack**, cutting **I/O power by 70 %** and freeing **25 % of die area** for additional compute units. This optimization boosts the number of HBM stacks per XPU by **33 %**, helping CSPs lower total cost of ownership.

Meanwhile, ASIC design houses such as **Alchip** and **GUC** are using **TSVs and hybrid bonding** to build 3DIC solutions that further enhance the **bandwidth and energy efficiency** between compute dies and HBM. These

technologies provide the infrastructure for CSPs to **massively deploy custom AI chips**.

HBM Competitive Landscape and Supply Chain Evolution

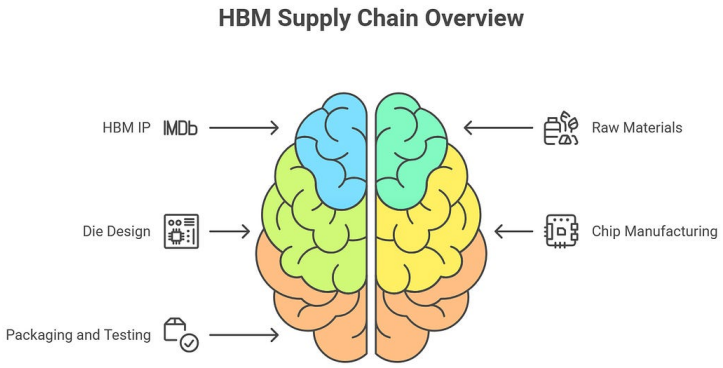
Company	Strengths / Characteristics	Potential Challenges
SK hynix	Advanced MR-MUF packaging enables HBM3E yields approaching 80%; collaborating with TSMC on HBM4 development and base die improvements; primary HBM3E supplier for NVIDIA's B-series GPUs.	Pressure to continuously expand capacity while maintaining high yields; facing competitive catch-up from Samsung and Micron in the HBM4/5 generations.
SAMSUNG	Vertically integrated with in-house DRAM and foundry capabilities, enabling a one-stop strategy; using 1c nm DRAM and 4 nm logic base dies in HBM4 development.	Lagging in MR-MUF packaging; HBM3/3E not yet qualified for NVIDIA's supply chain; 1c nm process progress has been below expectations.
micron	Actively ramping up HBM3E production; TC-NCF packaging yield currently around 40%–60%; plans to expand capacity by 2025.	Yield performance trails SK hynix; must solve thermal and process challenges for 16-Hi HBM stacks.
CXMT	Mass-produced HBM2 in 2024 and developing HBM3; plans to introduce HBM3E by 2027 and push for volume production.	Significant technology gap to global leaders; must overcome capital intensity and patent/IP restrictions.

In addition, the **HBM supply chain spans five critical segments**:

1. **HBM IP**,
2. **Raw materials**,
3. **Die design**,
4. **Chip manufacturing**
5. **Packaging and testing**.

Because HBM adopts a **3D stacked structure**, demand for **upstream equipment**—such as **TSV (Through-Silicon Via) processing tools**, **wafer-level packaging equipment**, and **bonding machines**—is rising rapidly.

On the **materials side**, key elements include **EMC underfill compounds**, **composite molding materials**, and **TSV buffer layers**. Major suppliers are building **competitive barriers** through **patent portfolios** and **tight control over their supply chains**, shaping the evolving competitive landscape of the HBM industry.

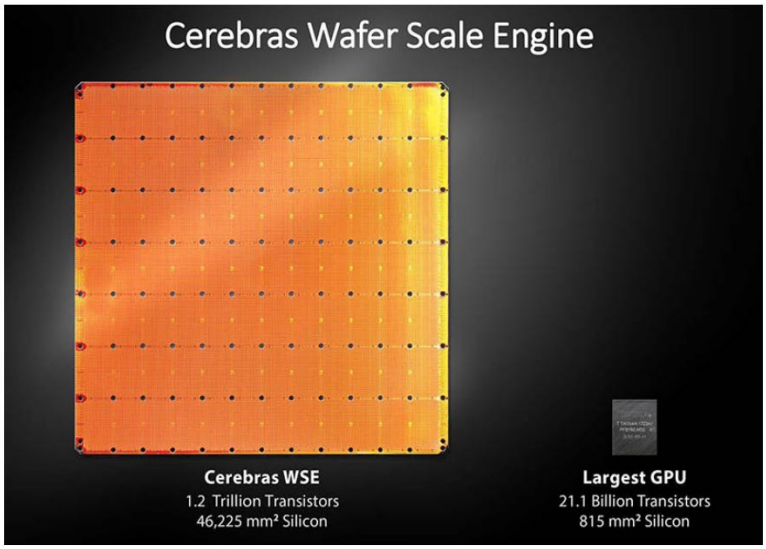


Key Takeaways

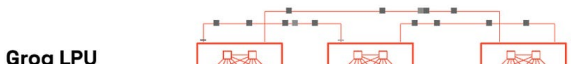
Memory wall and new interconnect technologies:
Current HBM relies on stacking and silicon interposers to deliver bandwidth, but this approach comes with **high costs and thermal challenges**. Future architectures may abstract memory using **photonic interconnects** or **CXL-based memory pooling**, enabling efficient data sharing and distribution across multiple accelerators. It will be critical to examine the role of **CXL 2.0/3.0** in AI data centers and the potential of **photonic wiring** to overcome **"shoreline limitations."**

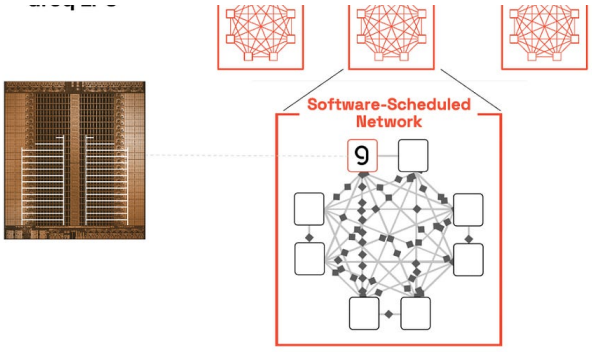
Hybrid bonding and back-end integration:
Hybrid bonding can increase inter-layer interconnect density and is expected to be introduced in the **HBM5 generation**. However, its impact on cleanroom class requirements, wafer grinding equipment, and overall manufacturing cost still requires in-depth analysis.

Cerebras :Near-memory computing and HBM alternatives



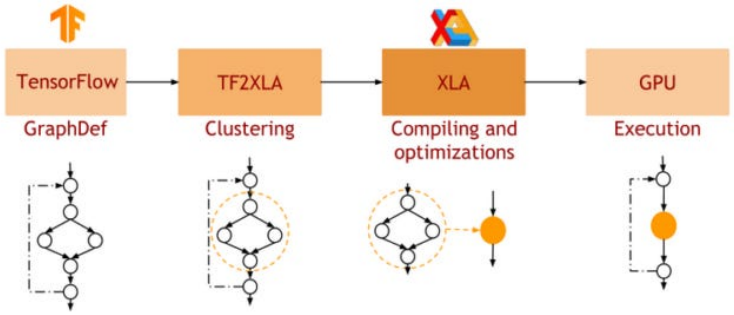
Companies like Cerebras (wafer-scale processors) and Groq (near-memory architectures) use massive on-chip SRAM to minimize data movement. It's worth exploring whether Processing-in-Memory (PIM) or Large Processing Units (LPU) could partially replace HBM in certain AI workloads.China's HBM development and geopolitics: CXMT's HBM program and China's semiconductor self-sufficiency push could reshape global supply chains and pricing. Meanwhile, U.S. export controls and licensing restrictions add further uncertainty. How these geopolitical factors will affect the global HBM market remains a key question.





HBM and photonic packaging integration:
To improve energy efficiency, vendors are investigating **co-packaging HBM with photonic dies or CPO (Co-Packaged Optics)**. For example, **Marvell** has proposed rearchitecting HBM I/O to integrate optical transceivers—potentially a major direction for the **post-HBM era**.

Memory expansion protocols and software ecosystems:
Beyond hardware, **compiler and framework support for HBM management** is crucial for GenAI workloads. It's important to study how **PyTorch 2.0**, **TensorFlow XLA**, and other frameworks optimize **HBM paging and data pipelines** to fully exploit available bandwidth and capacity.

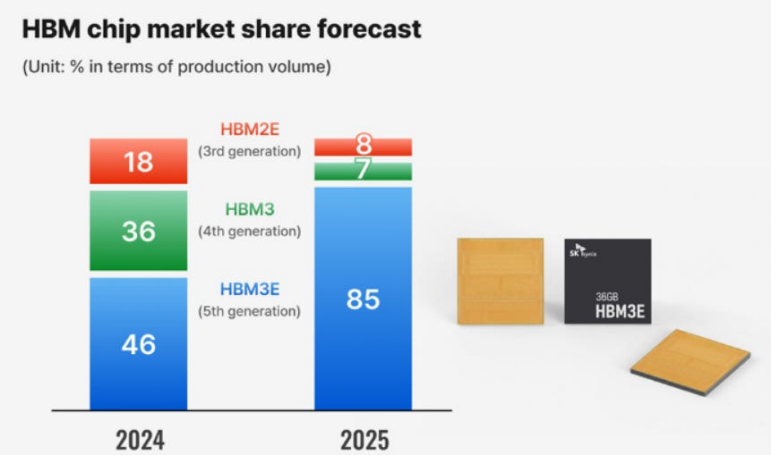


AI wave amplifies memory demand: Training and inference for large models are fundamentally constrained by **memory bandwidth and capacity**, making HBM an indispensable component of AI accelerators. **Upgrading to newer HBM generations or increasing stack heights** can significantly boost model throughput without changing compute power.

CSP custom silicon drives HBM demand: Cloud hyperscalers like **Google** and **Microsoft** are designing custom ASICs to optimize energy efficiency and cost, but **still rely on HBM** to provide high-speed cache and bandwidth. The industry is exploring **HBM I/O redesign and 3D IC integration** to reduce power and area overheads.

Evolving competitive dynamics: SK hynix currently leads the HBM

market thanks to its **packaging technology and yield advantages**. **Samsung** and **Micron** are working aggressively to catch up, while **Chinese vendors** are accelerating their entry. Future competition will center on **process nodes**, **packaging innovation**, **hybrid bonding**, and **photonic interconnect technologies**.



Emerging topics to watch: Breaking the **memory wall** will require **new interconnect** and **near-memory computing** technologies. **Hybrid bonding**, **electro-optical packaging**, and **China's semiconductor policies** will also play critical roles in shaping the industry.

This synthesis aims to provide readers with a clear understanding of the **interplay between AI chips and HBM**, as well as **emerging trends that will define the next wave of memory innovation**.

Summary of Explosive Growth of the HBM Market

1. Surge in Production and Consumption

- **GPU shipments:** 1.638 million units in 2023 → 7.133 million units in 2027
- **ASIC shipments:** 2.556 million units → 6.949 million units
- **Average GPU memory usage:** 80 GB → 422 GB
- **Average ASIC memory usage:** 40 GB → 364 GB

2. Market Size Expansion

The total HBM market is projected to skyrocket from **$3 billion to $53 billion**, with a **97% CAGR**.
Although unit prices will decline, the sharp increase in per-device capacity and adoption rates will massively expand the overall market size.

3. Technology Node Transition: HBM4 / HBM4e

- HBM4 will become mainstream starting in **2026**.

- HBM4e will further **boost bandwidth**, **increase stack height**, and **raise packaging integration complexity**.
- **Packaging and test supply chains** will be upgraded in tandem.

4. DDR4 / DDR5: A Supply–Demand Inflection Point

- **DDR4 investments are gradually shrinking**, and by 2026 a **10–15% supply gap** is expected.
- Although DDR4 demand is declining due to server upgrades, **long-tail markets** (automotive, industrial, TV, set-top boxes, etc.) will maintain a stable baseline.
- The transition toward **DDR5 and HBM** is firmly underway, pushing DDR4 into a **low-supply, structurally high-price** phase.

5. HBM4 / HBM4e Process and Base Die Trends

- By **2028**, demand for **HBM4e base dies** is projected to surpass **80 million units**.
- HBM4 will use **12 nm processes**, while HBM4e will leverage **5 nm nodes**, underscoring the **deep integration between memory manufacturing and leading-edge logic foundries**.
- The memory industry is evolving from **"component manufacturing" to "co-designed systems"**, integrating SoC + memory + interposers + packaging.
- This also signals a **supply-chain power shift**: HBM is no longer just a DRAM product but a **core system component** that determines performance and integration strategy.

6. Semiconductor Equipment Spending & Process Upgrades

- **DRAM equipment spending:** expected to reach **$32.7 billion in 2026**, up 10% YoY.
- **NAND equipment spending:** will surge in 2025 and remain strong in 2026 (9%–30% growth).

7. Key technology drivers include:

- Increased EUV layers (advanced nodes)
- HBM4/4e packaging and testing
- Wafer dicing and warpage control
- High-k materials and thermal compression bonding
- Introduction of mobile HBM (LLW DRAM)

These trends indicate that **equipment investment is shifting toward more advanced nodes and complex packaging**, raising the technical barrier for market participants.

8. NOR Flash: The "Hidden Demand Driver" for Edge AI

- The NOR Flash market is tightening due to **packaging capacity constraints** and **the explosion of IoT and Edge AI devices**.
- By **H1 2026**, NOR supply shortages could reach **mid-single-digit percentages**.
- **Wearables, sensor modules, and automotive electronics** will push NOR prices higher, making it a strategically important but often overlooked segment of the memory market.

9. Structural Characteristics of the New Memory Cycle

Unlike previous rebounds driven by **supply cuts**, the current memory upcycle is fueled by **structural AI-driven demand growth**.

Even if consumer markets remain weak in early 2026, **AI demand will support pricing**. A supply shortage is expected to last **4–6 quarters**, with **server DRAM and enterprise NAND spot markets** becoming particularly tight.

Structural vs. Cyclical Cycles — Will AI Reshape the Traditional Memory Boom–Bust Pattern?

Traditional Cycles vs. the AI Shift

Historically, the **DRAM/NAND memory industry has been highly cyclical**, driven by end-market demand (e.g., PCs, smartphones) and inventory adjustments. These **2–3 year "cyclical waves"** typically follow a familiar pattern: when demand surges, suppliers expand capacity and prices rise; eventually, oversupply leads to sharp price declines.

However, with the rise of **Artificial Intelligence (AI)**, this classic cycle is being **reshaped by structural growth forces**. Industry observers note that the current **"supercycle" in memory** is fundamentally different from past inventory-led booms: demand is now being **directly pulled by AI compute workloads**, lifting baseline consumption and creating a more durable, higher-value growth trajectory. In other words, AI is **shifting memory demand from single-market cycles (e.g., PCs, mobile) toward multi-sector synchronized structural expansion**.

Market Data and Forecast Models

Multiple market studies support the view that **future cycles will be longer and less volatile**:

- A U.S. investment bank report revealed that **OpenAI plans to consume up to 900,000 DRAM wafers per month by 2029**, roughly **half of total global DRAM capacity in 2025**. A single AI customer could support what used to be half of the entire industry — structurally lifting the demand baseline.

- According to **TrendForce**, AI-driven demand has pushed **global DRAM inventory turnover down to 3.3 weeks**, a **seven-year low**, far below the historical norm of ~10 weeks — highlighting acute supply tightness.
- **UBS analysts** project that OpenAI's in-house AI chips will use **HBM memory**, with a single project consuming **500,000–600,000 DRAM wafers per month between 2026–2029**.
- OpenAI's **"Stargate" supercomputer** alone is expected to consume **900,000 DRAM wafers per month**, equivalent to **~40% of global supply** — an unprecedented level of demand concentration.

At the same time, major DRAM vendors (Samsung, SK hynix, Micron) are **pivoting part of their production to HBM** and **upgrading process nodes (e.g., 1c nm)** to address AI-driven growth, as near-term capacity expansion is difficult.

Some executives even predict prolonged tightness: **Phison CEO K.S. Pua** forecasts **a decade-long NAND undersupply starting 2026**, while **Morgan Stanley** believes DRAM price increases — fueled by AI demand — will **continue through H1 2026**, offsetting macro weakness.

Toward a Longer, Smoother Supercycle

Two forces are making the **memory cycle longer and less volatile**:

1. **Strategic AI Data Center Demand:**
   DRAM/HBM and enterprise NAND consumption is increasingly tied to **long-term AI infrastructure investments**, not short-term consumer electronics cycles. This **smooths demand curves**, reducing sudden drops.
2. **Disciplined Supply Expansion:**
   After severe losses in the last downturn, memory makers are now **more cautious with CapEx**, expanding capacity more slowly. This keeps supply–demand healthier for longer.

Combined, these factors point to a **"tight, high plateau" cycle**, characterized by **moderate, sustained growth** rather than the sharp boom–bust swings of the past.

For example, multiple forecasts point to **multi-year DRAM upcycles**, with **structurally higher prices** driven by AI adoption. Of course, risks remain: if AI growth slows or CapEx surges too quickly, **cyclical corrections could still occur**, but their **amplitude and frequency will likely be lower** than in previous decades.

Strategic Implications for Memory Players

Memory companies should **treat this as a structural growth wave**, not just another short-term upcycle. Strategic shifts include:

- **Move from short-term to long-term planning**: Maintain disciplined capacity expansion during booms to avoid future oversupply.
- **Prioritize high-value products**: Focus capacity on **HBM and enterprise SSDs**, which have higher margins and strategic importance for AI.
- **Use market modeling to plan ahead**: Align capacity and raw material procurement with the **future wafer demand of major AI customers** like OpenAI, locking in long-term contracts.
- **Collaborate with governments and industry**: Build real-time **economic indicator systems** (e.g., AI server shipments, cloud CapEx) to guide supply planning and inventory policies.

By adopting these strategies, the memory industry can **capitalize on the AI-driven supercycle** while **avoiding the boom–bust trap** of previous decades — achieving **steady, high-value growth in the AI era**.

HBM4e and Advanced Logic Process Competitiveness — Process Dependence and Strategic Collaboration in Next-Gen Memory

Dependence of HBM4/4e on Advanced Logic Nodes

The fourth generation of High Bandwidth Memory (HBM4/4e) is widely regarded as a **critical enabler for the AI era**, requiring significantly higher I/O interface width and speed. To achieve this, the **base die** architecture of HBM has undergone a major technological shift.

In previous generations (HBM1–3), base dies were typically manufactured using DRAM fabs' **in-house planar DRAM processes**, serving mainly as passive signal pass-through layers. However, with **HBM4**, the data interface width has **doubled from 1024 bits to 2048 bits**, and per-pin data rates are approaching **10 Gbps**, exposing the **speed, signal integrity, and power limitations** of traditional DRAM processes.

To overcome these challenges, **leading vendors (Samsung, SK hynix)** have decided to move **HBM4 base die production** to **advanced foundry logic nodes** (e.g., 12 nm or even 5 nm FinFET). For instance, **TSMC** has announced support for **N12FFC+ (12 nm)** and **N5 (5 nm)** processes for HBM4 base die manufacturing. A 5 nm base die provides **higher logic density** and **lower power consumption**, ensuring stable control circuitry operation in ultra-high-speed environments.

As a result, **HBM4/4e is deeply dependent on leading-edge logic processes**, pulling the memory industry into an unprecedented **"process race."**

Cost and Yield Challenges of 5 nm Base Dies

While moving to **5 nm** processes boosts performance, it also brings **significant cost and yield challenges**:

- **Cost Structure:**
  5 nm foundry wafers are extremely expensive, partly due to EUV lithography tools that cost **$200–300 million per unit**. Outsourcing the base die to advanced foundries significantly increases manufacturing costs. Industry reports indicate that the **base die accounts for ~20% of total HBM4 manufacturing cost** at TSMC, pushing up overall HBM pricing.
  Recent market reports suggest that **SK hynix's HBM4 pricing has risen ~70% over HBM3E**, with **12-Hi HBM4 stacks priced around $500 each**, compared to ~$300 for HBM3E — largely reflecting the cost of advanced logic processes.
- **Yield Ramp:**
  Advanced nodes like 5 nm typically experience a **yield learning curve**. If base die yields are low, unit costs increase sharply and can **drag down the overall stack yield**, as **any defective base or DRAM die causes the entire HBM stack to fail**.
  South Korean industry sources note that Samsung initially planned to use 8 EUV layers for its 1c DRAM node but reduced this to 6–7 layers to balance cost and yield. Similarly, achieving reasonable yields for 5 nm base dies is a **major challenge**.
  Samsung may absorb some costs through its own fabs, while **SK hynix and Micron must rely on TSMC**, facing higher foundry prices. Micron even **delayed its adoption of advanced nodes to HBM4e**, waiting for cost structures to stabilize.

Strategic Collaboration Models Between Foundries and Memory Vendors

To address these challenges, memory makers and foundries are entering **deeper strategic collaborations** than ever before. Key models include:

- **Foundry Alliances & Capacity Lock-Ins:**
  Micron has formally partnered with **TSMC** to manufacture HBM4e base dies, targeting **mass production in 2027**. Micron essentially purchases advanced logic capacity and then packages the finished HBM at the wafer level. This guarantees Micron access to 5 nm/3 nm technology, while TSMC secures long-term capacity utilization — a **mutually beneficial lock-in**.
- **Joint Development & Custom Design:**
  SK hynix signed an **MOU with TSMC in April 2023** to co-develop HBM4, focusing on tighter integration between the base die and CoWoS packaging. A **joint "One Team"** R&D group works on optimizing base die performance, lowering power, and co-tuning the interface with TSMC's advanced packaging.
  This **goes beyond a pure foundry relationship**, resembling **co-development**, where the memory vendor provides HBM specs and TSMC contributes process and packaging expertise — enabling SK hynix to stay ahead in HBM4 performance while reinforcing TSMC's role in the AI packaging ecosystem.

- **Vertical Integration:**
  Samsung Electronics, which operates both DRAM and foundry divisions, pursues **full vertical integration**. Its HBM4 base dies are produced in-house using 5 nm/4 nm lines, combined with **I-Cube packaging** technologies. This allows Samsung to **control costs and supply**, avoiding reliance on rival foundries. However, it also bears the **full R&D and yield risk** internally, demanding strong technical capabilities.
- **Customer-Driven Customization:**
  Looking forward, **customized HBM base dies** will become common. Major AI players like **NVIDIA and OpenAI** may design their own base dies for HBM4e, then rely on memory vendors to stack DRAM and foundries for fabrication. Samsung plans to use **VCS (Vertical Cu-Post Stack)**, while SK hynix is developing **VFO (Vertical Fan-Out)** packaging to support such customized workflows.
  This **tri-party collaboration model** — AI company + memory vendor + foundry — enables tailored HBM optimized for specific accelerators.

Strategic Implications

As HBM moves into the **advanced process era**, memory vendors must **adopt flexible, multi-pronged strategies** to stay competitive:

1. **Collaborative R&D:** Early engagement with foundries (e.g., SK hynix + TSMC) can shorten the yield ramp and tailor processes to HBM requirements.
2. **Cost-Sharing Agreements:** Long-term capacity contracts with foundries can secure better pricing and transfer part of the cost to AI customers via pre-orders.
3. **Vertical Integration & Customization:** Vendors can either follow Samsung's self-sufficiency model or build co-development capabilities with key customers.
4. **Talent & Technology Roadmaps:** Investing in 5 nm and below expertise, including **3 nm and High-NA EUV**, ensures leadership as logic-memory integration deepens.

Through such strategies, memory vendors can **leverage advanced nodes while mitigating risks**, positioning themselves strongly in the **HBM4 era of AI memory competition**.

Edge AI and Mobile HBM Technology – The Outlook for High-Bandwidth Memory in Smartphones, XR, and Automotive Devices

The Concept and Demand for Mobile HBM (LLW DRAM)

As AI applications **expand from the cloud to the edge**, devices such as **smartphones, AR/VR headsets (XR)**, and **automotive AI systems** are facing rapidly increasing demands for **high-bandwidth, low-latency memory**. Traditional mobile DRAM solutions like **LPDDR5X** offer limited

bandwidth and are no longer sufficient to handle the massive data throughput required for **on-device AI inference**.

To address this, the industry has proposed a "**mobile version of HBM**," also referred to as **Low Latency Wide I/O DRAM (LLW DRAM)** or **LPW DRAM**. The core idea is to **widen the memory I/O interface** (i.e., implement a wide bus) while **lowering the per-channel speed**, achieving both **high data transfer rates** and **low power consumption**. In other words, this approach brings the **HBM wide bus concept** to mobile devices but optimizes it for **battery-powered environments**.

At **ISSCC 2025**, Samsung revealed its plan to launch **next-generation low-power wide I/O DRAM (LPW/LLW DRAM)**, with the first mobile products expected to debut in **2028**. Meanwhile, rumors suggest **Apple** is considering integrating mobile HBM technology into its **20th-anniversary iPhone**, slated for **2027**, to dramatically boost on-device AI capabilities.

This indicates that incorporating **HBM-class memory** into smartphones, wearables, and automotive systems is **no longer a distant vision**, but rather a key industry focus for the **next 2–3 years**.

Adoption Timeline and Thresholds

The **commercialization window for mobile HBM** is expected to fall between **2026 and 2028**. Major memory manufacturers such as **Samsung** and **SK hynix** are actively developing their respective LLW DRAM solutions. Samsung has publicly stated that it will release the **first mobile device equipped with LPW DRAM in 2028**.

Market rumors also suggest that **Huawei** may attempt to **beat Apple to market**, potentially unveiling a flagship smartphone using HBM-like memory as early as **2025–2026**, although this has yet to be confirmed.

Based on current trajectories, **around 2027** is likely when leading vendors will **pilot mobile HBM** in **ultra-premium smartphones** or **MR headsets**, targeting use cases such as **on-device LLM inference** and **real-time computer vision**.

In the **automotive domain**, high-end autonomous driving platforms require extremely high memory bandwidth. Some are currently adopting **GDDR6**, but **next-generation automotive AI chips (2027+)** may turn to **mobile HBM** to achieve both **higher bandwidth** and **lower power consumption**. For example, future iterations of **Tesla's FSD chips** or other custom automotive accelerators may integrate mobile HBM as part of their architecture.

Overall, the **key barriers** to mobile HBM adoption are **technical maturity** and **cost absorption within flagship products**. Both are expected to be gradually overcome toward the **end of this product cycle**, paving the way for **wider deployment of mobile HBM between 2026 and 2028**.

Breakthrough Conditions: Power, Packaging, and Cost

To bring **HBM-class high-performance memory** into **mobile and XR (Edge) devices**, several technical bottlenecks must be overcome:

Power Consumption & Thermal Management

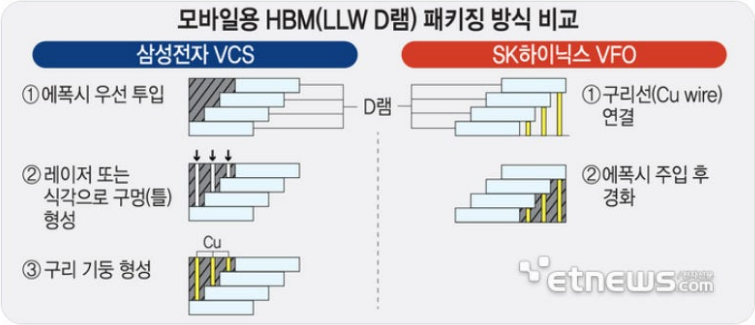
Mobile devices are strictly limited by battery capacity and compact form factors, making the **power overhead of high-bandwidth memory** a key challenge.

LLW DRAM (Low-Latency Wide I/O DRAM) addresses this by **increasing the number of I/Os while lowering per-pin frequency**, resulting in a large aggregate bandwidth with **lower per-pin power**. Samsung estimates that the new LPW DRAM can **deliver over 200 GB/s bandwidth while reducing power consumption by ~54% compared to LPDDR5X**.

As memory and SoC are tightly packaged, thermal density increases significantly. Advanced heat dissipation measures—such as **vertical thermal channels**, **high-performance heat spreaders**, or **using through-silicon copper posts in 3D packaging for heat conduction**—are required. If power can be kept within acceptable limits and heat is effectively managed, **mobile HBM becomes viable**.

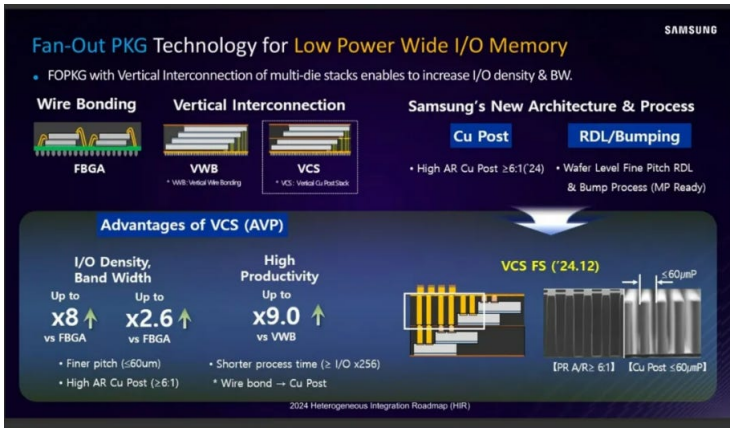
Packaging Technologies

Mobile devices cannot use the **large silicon interposers** found in data-center GPUs due to size and cost constraints. To solve this, vendors are developing **new chip stacking and interconnection techniques**:



- **Samsung VCS (Vertical Cu-Post Stack)** and **SK hynix VFO (Vertical Fan-Out)** replace the large interposer with **vertical copper pillars or fan-out packaging**, tightly bonding wide-bus memory to the SoC.

- Samsung also proposed **Vertical Wire Bonding (VWB)** at ISSCC, replacing traditional looping wires with straight vertical interconnects to reduce latency.



These are essentially **simplified 2.5D/3D packaging methods** that directly connect memory dies to processors through ultra-short interconnects. They enable **wide I/O integration within a small footprint** while maintaining manageable package height and reliability.

However, **packaging yield is a critical challenge**. Hundreds or even thousands of micro-interconnects must be bonded simultaneously and remain reliable over the device's lifetime. This demands **precise alignment**, **advanced materials**, and **process innovation**. In short, **mature packaging technology is a prerequisite for mobile HBM deployment**.

Cost and Capacity

At current technology maturity, **mobile HBM is significantly more expensive than LPDDR**, both due to **advanced packaging** and potential **silicon stacking** steps. Its **cost per bit may be several times that of LPDDR**.

Analysts point out that mobile HBM faces **thermal constraints and 3D stacking yield issues** in slim devices, further increasing costs. Commercial viability typically depends on **maturity, yield ramp-up, and scale effects** to bring costs down. Initially, mobile HBM will likely appear in **ultra-premium flagships or specialized devices**, with prices expected to decline after **2028** as production scales.

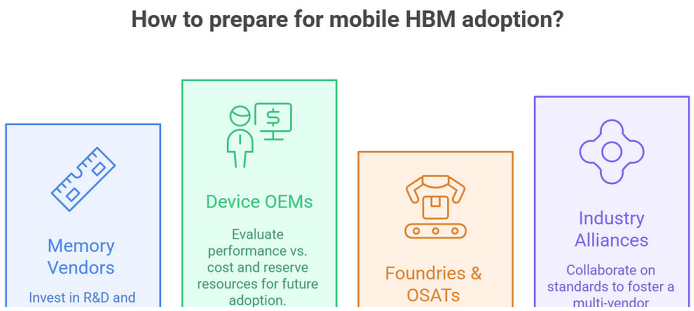
In terms of **capacity**, mobile HBM is expected to range from **a few gigabytes up to several tens of GB**, much larger than conventional mobile DRAM, to support on-device AI models. Achieving this within limited physical space may require **taller DRAM stacks (e.g., 16-Hi)**, which imposes additional **packaging and power management challenges**.

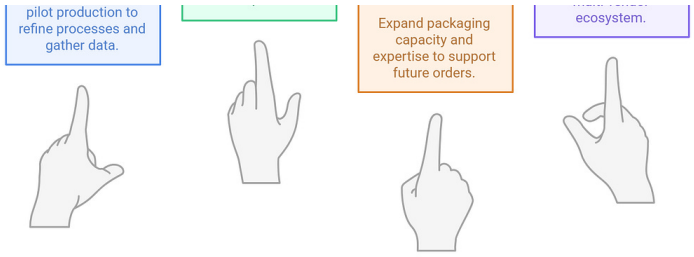
Strategic Implications for the Edge AI Era

To prepare for the **Edge AI boom**, industry players should **start building mobile HBM capabilities early**:

- **For memory vendors:**
  - Invest in **low-power wide-I/O DRAM R&D** and collaborate closely with mobile SoC makers to define interface standards.
  - Consider **small-scale LLW DRAM pilot production** for AI smartphones and AR glasses developer kits to accumulate real-world data and refine processes.
- **For device OEMs:**
  - Evaluate the **performance gains vs. cost trade-offs** of mobile HBM early.
  - **Reserve board space and power budgets** in 2026–2027 flagship roadmaps to enable quick adoption once the technology matures.
- **For foundries & OSATs:**
  - **Expand advanced packaging capacity** and build relevant engineering expertise in advance to support future mobile HBM orders at scale.
- **For industry alliances:**
  - Collaborate on **JEDEC or open interface standards** for mobile HBM to foster a multi-vendor ecosystem, accelerate cost reduction through volume, and avoid vendor lock-in.

With these strategies in place, the industry will be well-positioned to **capitalize on the Edge AI wave**, using **mobile HBM** to meet the rapidly growing demand for **high-bandwidth memory in intelligent edge devices**.





Supply Chain Bottlenecks and Capacity Expansion — Key Factors in HBM / Advanced Memory Manufacturing

Advanced Packaging Capacity Bottlenecks:
For HBM and other advanced memory products, not only front-end wafer processing but also back-end **testing and packaging** are critical to determining whether production can scale smoothly. At present, the most severe constraint is **the shortage of advanced packaging capacity for HBM**.

HBM requires 2.5D silicon interposers—such as TSMC's CoWoS technology—to integrate multiple DRAM stacks with processors. However, TSMC's CoWoS capacity is limited. Since 2023, surging demand for AI chips has made CoWoS packaging the **central bottleneck** in the AI supply chain.

TSMC plans to expand CoWoS capacity to **75,000 wafers per month in 2025**, and to **90,000–110,000 wafers per month by 2026**—roughly double its 2023 level. But demand is rising even faster. Reports indicate that despite these expansions, **cloud giants like Amazon and Microsoft are still facing multi-year delays** in securing AI chip packaging slots.

This "**sweet bottleneck**" is pushing the industry to accelerate **next-generation packaging technologies**, such as **panel-level packaging (CoPoS)**, which TSMC aims to pilot in 2026 and mass-produce by 2028. By replacing silicon wafers with large panels, CoPoS can dramatically increase the number of chips packaged per batch.

Between 2026 and 2028, however, conventional CoWoS capacity is expected to remain tight, becoming the **primary limiting factor** for HBM volume ramp. For DRAM makers, even if wafer capacity is sufficient, lack of timely packaging services can stall shipments. As a result, many companies are exploring **alternative solutions**:

- **Samsung** is ramping its in-house **I-Cube** packaging,
- **Intel** and OSATs are collaborating on **EMIB** and **FOLOB** technologies, all aimed at reducing dependency on TSMC.

Intel's Next Frontier: Redefining Chiplet Integration Through Advanced Packaging

SEMIVISION · MAY 1, 2025

Read full story

Testing and Yield Challenges:
The **testing stage** is another often-overlooked bottleneck. HBM testing is far more complex than conventional DRAM, involving:

- Known Good Die (KGD) screening for each die before stacking,
- TSV electrical testing,
- High-speed interface stress testing, and
- Extensive burn-in at the package level.

Before dicing, each DRAM wafer undergoes **probe testing** to identify good dies, which requires higher tester precision and throughput. Once stacked into HBM, **packaged devices must undergo stress and burn-in tests** to weed out latent defects.

Due to HBM's **wide I/O and large capacity**, testing each device is time-consuming, creating significant demand for ATE (Automated Test Equipment). Leading ATE vendors have flagged this surge:

- Advantest noted during its 2024 financial briefing that **HBM testing capacity must be expanded**,
- It expects HBM-related tester shipments to rise significantly in the second half of the year as capacity improves.

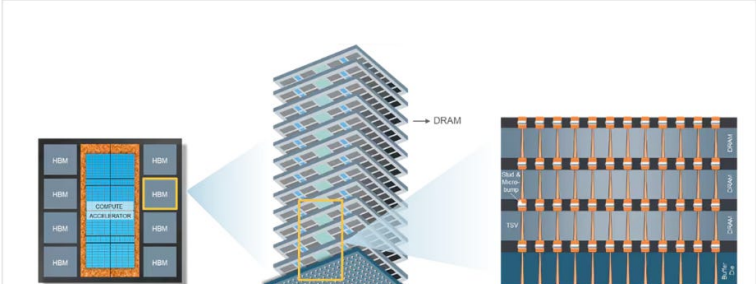
Advantest Leading the AI Testing Wave: Earning the Title of the "ASML of the Test Industry"

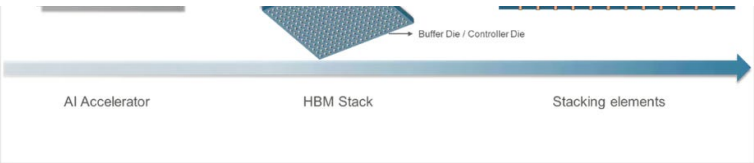
SEMIVISION · AUGUST 10, 2025
Read full story

If test equipment supply cannot keep pace, even completed HBM dies will face delays in validation and shipment, forming another bottleneck.

To support the expected **HBM ramp in 2026–2028**, IDM and foundry players must:

- **Expand high-end tester fleets**,
- **Adopt more efficient test flows**, such as wafer-level burn-in to reduce post-packaging fallout, and
- Strengthen yield management.





HBM stacks involve **dozens of small dies**, meaning that any error in the chain can sharply reduce yields. The supply chain must therefore invest in **advanced inspection tools**—such as 3D X-ray and AOI (automated optical inspection)—and ensure sufficient capacity at these inspection nodes during peak production periods. Otherwise, **inspection itself could become a new choke point**.

In short, **packaging and testing** are emerging as critical determinants of HBM supply growth. Without parallel expansion of these back-end capabilities, even breakthroughs in DRAM fabrication won't translate into real shipment capacity.

TSMC: Acceleration of Taiwan's "Local-to-Local" Semiconductor Strategy

SEMIVISION · SEPTEMBER 25, 2025

Read full story

Materials and Process Equipment Bottlenecks in Memory Device

The manufacturing of advanced memory also depends on whether the supply of critical materials and specialized process equipment can keep up with capacity expansion. For example:

High-k dielectric materials:
To shrink DRAM capacitor size and improve charge retention, major DRAM makers began introducing high-k materials (e.g., HfO_2 hafnium oxide) as capacitor dielectrics starting at the 1α/1β nodes. High-k requires precise ALD (atomic layer deposition) processing, mostly provided by a limited set of suppliers (e.g., JSR in Japan, Applied Materials in the U.S.). If high-k materials are widely adopted in HBM and other high-end DRAM, the demand for these chemicals and ALD tools will surge. Because production of such materials is complex and supplier concentration is high, any tightening of material supply during the 2026–2028 ramp-up phase could disrupt wafer manufacturing. In addition, yield ramping for high-k processes takes time; early production runs may have lower yields, indirectly constraining output.

Materials challenges are not limited to DRAM. ABF substrates, silicon interposers, and high-density glass cores used in advanced packaging are also potential chokepoints. For example, ABF substrate capacity expansion typically takes over a year to build new facilities, and there are few short-term substitutes. If substrate supply is insufficient, HBM packaging schedules may be delayed, creating shipment bottlenecks.

Thermal Compression Bonding (TCB) and equipment:
As noted earlier, TCB is a critical process for accurately aligning and bonding memory dies onto interposers or substrates. TCB requires specialized tools to heat and press each die individually, making it relatively slow. SK hynix's MR-MUF batch reflow process is more efficient, but it too requires dedicated molding materials and equipment. During the HBM capacity ramp, the availability of high-pin-count, high-precision bonding and reflow tools will be a key determinant of whether expansion can keep pace. If manufacturers and OSATs fail to procure enough TCB tools or vacuum reflow ovens by 2026, bottlenecks could form at the packaging stage.

Moreover, tuning these tools to mass-production standards takes time, and training skilled operators is another hurdle. These constraints are often underestimated but very real—for example, during the 2024 NVIDIA H100 boom, packaging lines experienced workforce and equipment shortages, forcing urgent recruitment and extra shifts. To avoid similar problems in 2026–2028, the supply chain must plan ahead for equipment and manpower needs.

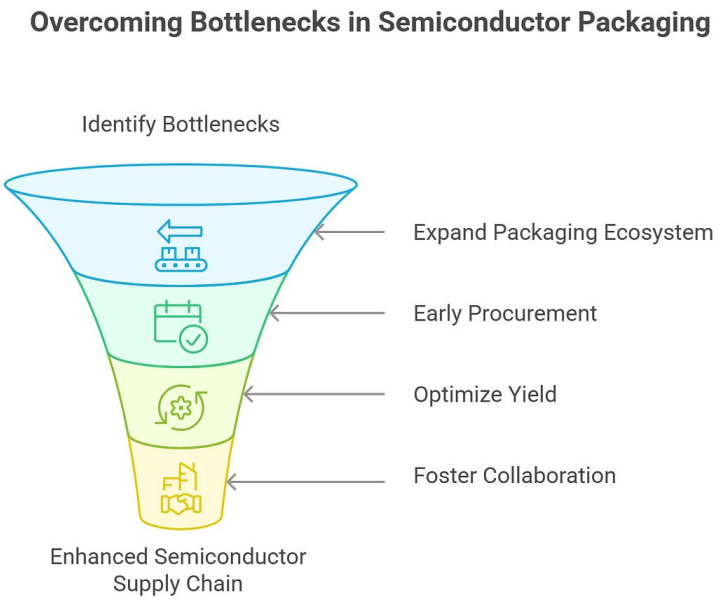
Advanced lithography and other bottlenecks:
Although packaging and testing dominate the discussion, front-end challenges may also constrain output. Starting at the DRAM 1γ/1δ nodes, EUV lithography is being introduced; if ASML cannot deliver sufficient EUV tools on time, wafer output will be capped. Additionally, TSV (through-silicon via) etching and plating—key steps in HBM die fabrication—depend on deep etch and electroplating tools. Delays in the delivery or ramp-up of these specialized tools can limit capacity expansion, particularly for Chinese manufacturers who may face restrictions on acquiring the most advanced etch equipment, impacting their HBM yield and output.

Thus, the entire value chain must systematically examine capacity bottlenecks across all stages—including but not limited to packaging, testing, materials, and equipment—and address weak links to ensure smooth ramp-up.

Impact on 2026–2028 capacity ramp:
Packaging capacity is likely to be the first and most critical bottleneck; before new technologies are widely adopted, HBM volume growth may be limited by packaging line expansion speed. Testing equipment and processes are a secondary bottleneck: less severe than packaging, but still capable of slowing shipments if capacity isn't expanded early. On the materials and process side, as long as suppliers coordinate effectively, increased stockpiling and capacity expansion can typically meet demand—but specific materials like ABF substrates or specialty gases pose notable risks.

Overall, while annual growth in HBM and advanced memory is expected to be high between 2026 and 2028, the degree of supply chain coordination

will determine the actual output curve. Poor handling of bottlenecks could lead to “demand waiting on supply,” capping sales below expectations; conversely, coordinated expansion could enable the industry to meet explosive market demand more smoothly.



Strategies to overcome these bottlenecks:

- **Expand the packaging ecosystem:**
  Advanced packaging capacity is currently concentrated in a few players such as TSMC. Memory and chip design companies should explore second sources, e.g., collaborating with OSATs like ASE to develop CoWoS-like solutions, or adopting Intel’s EMIB/FO packaging as alternatives. Governments can support local OSAT upgrades to diversify supply and reduce risk.
- **Early procurement of equipment and materials:**
  For potential bottlenecks like test equipment, high-k materials, or ABF substrates, companies should place long-term orders 2–3 years in advance and build strategic inventories. For long lead-time equipment such as EUV scanners or ALD tools, CapEx planning must be meticulous to avoid last-minute shortages.
- **Yield and process optimization:**
  Companies can invest in more efficient testing methodologies (e.g., wafer-level burn-in, AI-assisted data analysis) to increase test throughput, and continue improving packaging processes such as adopting MR-MUF batch bonding to speed up assembly and improve yields. A 1% yield increase can translate into hundreds of millions of

dollars in added output—well worth prioritizing.

- **Industry collaboration and government support:**
  Overcoming bottlenecks requires coordinated action across the ecosystem. Industry associations can organize supply chain summits to align capacity plans, while governments can offer tax incentives and fast-track permits for advanced packaging and testing expansions.

Through these measures, the memory industry can avoid local bottlenecks during the critical 2026–2028 window and fully convert technological breakthroughs into real market supply capacity, capitalizing on the AI boom rather than being constrained by it.

Materials and Process Innovations – Impact of New Technologies on Yield, Cost, and Capacity

EUV Adoption and Multi-Patterning:

Extreme ultraviolet lithography (**EUV**) represents a major innovation in advanced semiconductor manufacturing and is gradually being applied to **DRAM production**. The concept of "**EUV layering**" refers to using EUV lithography to pattern fine features that previously required multiple exposure steps with 193 nm ArF immersion, effectively increasing patterning density with fewer steps. For memory devices, this enables **smaller cell sizes** (e.g., reduced $4F^2$) and **higher circuit density**.

Attendees gathered at the event to commemorate the industry-first introduction of the High NA EUV system at the M16 fab

SK Hynix has been actively acquiring extreme ultraviolet (EUV) lithography machines from ASML to produce advanced chips. A 2021 contract involved a $4.3 billion investment over five years, and more recently, the company installed the industry's first commercial High-NA EUV system at its M16 fab for memory production. SK Hynix is also reportedly planning to acquire about 20 additional EUV systems by 2027 to double its EUV capacity.

> **SK hynix Introduces Industry's First Commercial High NA EUV**

However, EUV introduces significant **cost and yield challenges**. On one hand, EUV scanners are extremely expensive — each tool costs roughly **$200–300 million** — and have relatively limited wafer throughput per hour. Adopting EUV thus requires a surge in upfront capital expenditure and fundamentally reshapes the manufacturing cost structure.

To control costs, companies like **Samsung** have reportedly sought to **reduce the number of EUV layers** in DRAM: for example, Samsung's 1c DRAM node was initially planned to use 8–9 EUV layers but was later scaled back to 6–7 layers to save on equipment investment.

On the other hand, **initial EUV yield tends to be low**, due to factors such as mask defects and stochastic lithography errors. Overcoming these issues requires improvements in **mask protection**, **pattern correction**, and extensive **pilot production data**. As a result, during early adoption phases, DRAM yields may temporarily decline, raising per-chip costs.

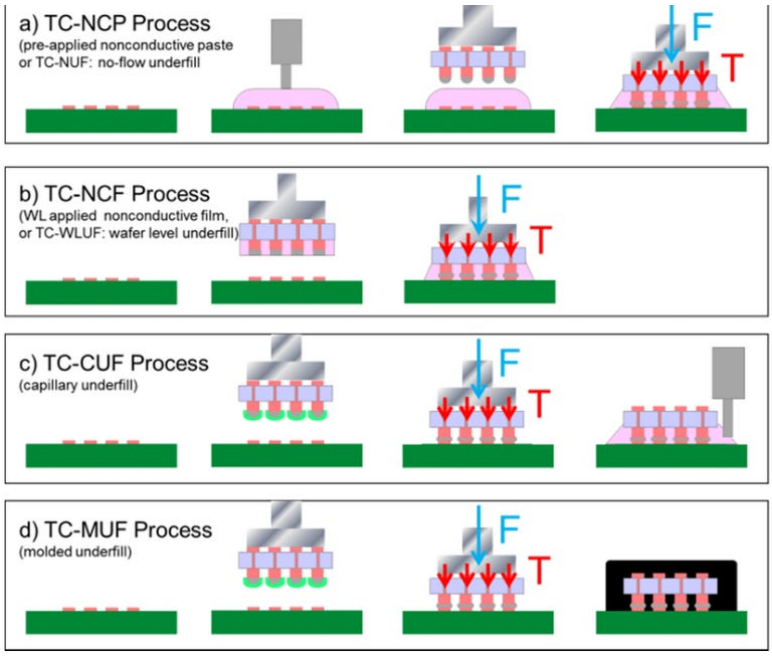
In the long run, however, EUV simplifies the process flow by **reducing multi-patterning steps** and their associated overlay errors, while improving **critical dimension (CD) control**, ultimately leading to **higher final yields** and **better performance consistency**. Once yields ramp up and capital depreciation is amortized, the **unit cost of EUV-patterned DRAM may actually fall**, extending the scalability of Moore's Law in the memory domain.

In terms of **capacity planning**, while EUV scanners have lower raw throughput than traditional ArF tools, a single EUV tool can replace **multiple ArF multi-patterning steps** once the process matures. Thus, as long as the number of EUV tools keeps pace with demand, EUV should not become a bottleneck; however, **limited EUV tool availability could constrain output**.

Overall, EUV represents a **"high upfront cost, high long-term benefit"** investment. Manufacturers must endure a period of **lower yield and higher per-unit cost** during the yield ramp phase, but once this transition is complete, they can reap significant benefits from **process simplification, improved pattern fidelity, and reduced overall cost** in advanced DRAM production.

Thermal Compression Bonding (TCB) and Advanced Stacking:

Innovations in thermal compression bonding (TCB) are having a significant impact on the **yield and cost structure** of HBM and other 3D-stacked memory technologies. Traditionally, to ensure precise alignment of each die, manufacturers have used **die-by-die thermal compression bonding**, ensuring reliable fine-pitch micro-bump connections. However, this approach involves long process times and low equipment utilization, which drives up assembly costs and limits throughput.



A new generation of techniques, such as **large-area simultaneous bonding** (e.g., SK hynix's **MR-MUF process**), enables multiple dies to be bonded in a single reflow step.

In terms of **yield**, die-by-die TCB offers excellent precision, but each bonding step introduces a potential failure point, and with increasing stack height, the cumulative risk grows significantly. MR-MUF's one-shot bonding reduces the number of repetitive alignment steps, theoretically lowering accumulated alignment errors and helping to maintain interconnect consistency across the entire stack height. In practice, SK hynix claims that its advanced MR-MUF process can keep **12-Hi HBM** stacks within height tolerances while also improving thermal characteristics. This indicates that the company has overcome the technical challenges of bonding multiple dies simultaneously, achieving both high reliability and yield.

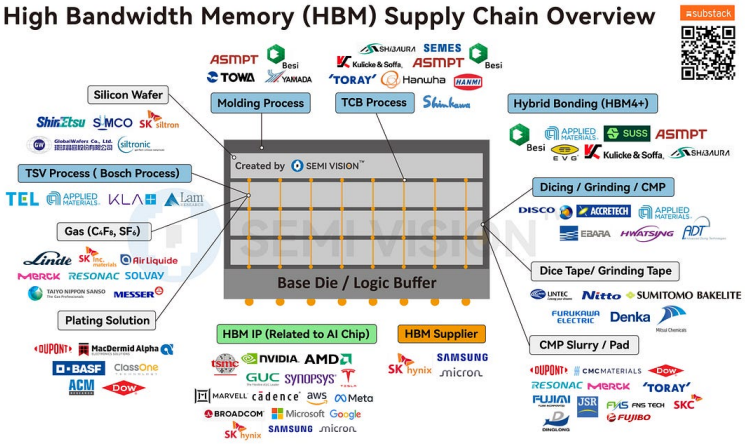
From a **cost perspective**, MR-MUF eliminates the time-consuming die-by-die bonding steps, drastically increasing packaging throughput and lowering the per-unit assembly cost. While conventional TCB offers precision, it is expensive and scales poorly — costs rise sharply with each additional stack layer, making it unsustainable for future high-volume production.

Looking forward, if thermal compression bonding continues to evolve toward **batch and automated processing** — for example, bonding multiple HBM stacks across an entire wafer in a single thermal compression step — it could massively boost packaging capacity and reduce packaging's share of the overall cost structure.

This has **direct implications for capacity planning**: manufacturers that

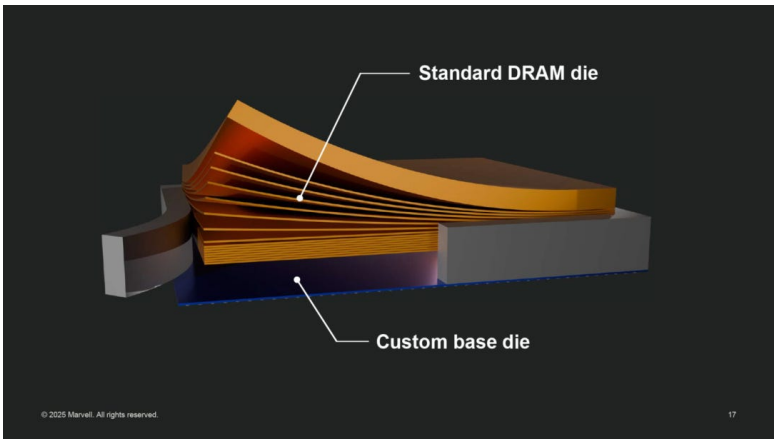
adopt high-throughput bonding technologies will require fewer packaging lines and less manpower, enabling them to deploy more capacity within limited fab space to meet peak demand. In contrast, companies that remain dependent on traditional TCB methods may face **packaging becoming a production bottleneck**.

In short, **innovations in thermal compression bonding** offer the dual benefits of **higher yield per unit** and **lower assembly costs**, making them an inevitable trend for next-generation advanced memory.



5 nm Base Die and Logic Packaging Innovations:

As discussed earlier in the HBM section, the introduction of 5 nm base dies (logic dies) has enabled a major leap in HBM performance, but it also brings significant implications for yield and cost structures. Each HBM stack now includes a base die containing several billion transistors—essentially a small SoC—whose yield directly determines the overall product yield.



If initial 5 nm yields are only around 50%, half of the base dies would be

scrapped. Combined with potential DRAM stack losses, the final product yield could be even lower. This presents a substantial supply challenge for HBM in 2025–2026. As a result, memory vendors and foundries must work closely to improve base die yields through strategies such as optimizing EDA fault-tolerant design, applying more relaxed design rules than those used for logic SoCs to trade performance for yield, or even performing localized wafer-level repair where technology allows.

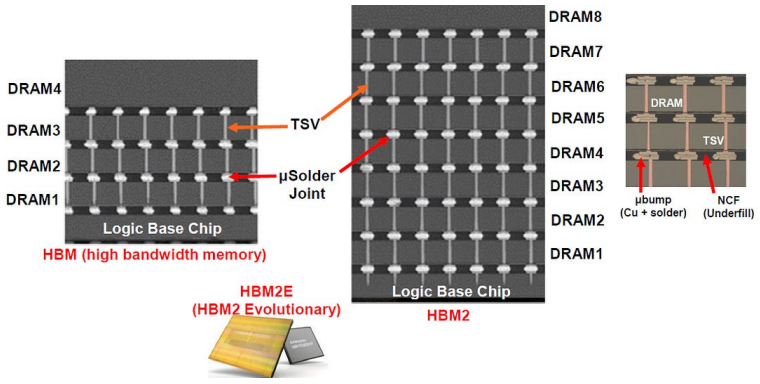
Every 1% yield improvement translates into thousands of additional HBM stacks—an enormous economic gain that the industry is increasingly aware of. From a cost perspective, the base die accounts for roughly 20% of total HBM cost. Higher yields are essential to lower the per-unit cost.

Different collaboration models will also impact cost structures. For example, if companies like OpenAI or NVIDIA design their own base dies and outsource manufacturing, economies of scale could lower costs, but heavy customization may make capacity sharing difficult, potentially requiring higher prices to offset inefficiencies.

On the capacity side, 5 nm production is already in high demand among multiple customers. Memory makers must secure capacity agreements with foundries to ensure stable base die supply and avoid being squeezed out by CPU or GPU orders during peak periods.

There are also technical roadmaps exploring **HBM without a base die**, such as distributing logic functions within the DRAM layers or integrating them into the underlying interposer. If successful, this could simplify manufacturing and eliminate one 5 nm die, improving both yield and cost. However, in the near term, 5 nm base dies will remain mainstream.

Therefore, rapidly climbing the yield learning curve and driving down costs will be critical priorities for HBM suppliers around 2025.



High-k Dielectrics and Other Process Innovations

The use of **high-k materials in DRAM capacitors** represents a major technological breakthrough. Micron was the first to commercialize high-k capacitors at the 1α node, enabling cell capacitance to be maintained even

as feature sizes shrank. This had direct implications for **yield and reliability**: high-k dielectric layers are thinner and exhibit more complex dielectric behavior, which can initially lead to higher leakage currents and increased variability, requiring extensive process tuning. In fact, Micron's initial yield issues during early 1α ramp-up were closely tied to the challenges of integrating new materials.

High-K Material	*K-value*	*Gap (eV)*
Si	3.9	1.1
SiO2	3.9	9
Si3N4	7	5.3
Al2O3	9	8.8
Ta2O5	22	4.4
TiO2	80	3.5
ZrO2	25	5.8
HfO2	25	5.8

However, after iterative refinements, high-k capacitors ultimately **improved product reliability**, since maintaining sufficient capacitance ensured longer data retention times.

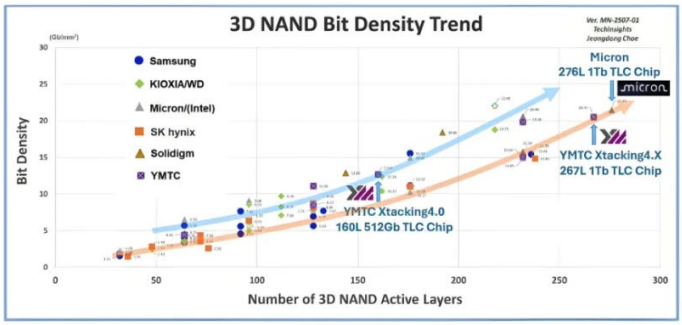
From a **cost structure** perspective, adopting high-k involves adding several ALD deposition and annealing steps. While these increase process complexity, they represent a small fraction of total DRAM manufacturing cost. In the long run, high-k enables greater scaling and cell density without needing larger cells or complex 3D structures, thereby **lowering the cost per bit**. In the short term, however, R&D expenses rise due to new equipment purchases and materials development.

On the **capacity side**, ALD processes are slower than conventional CVD, which could become a throughput bottleneck if equipment is not scaled up. In practice, manufacturers typically deploy multiple tools in parallel and optimize deposition cycle times to maintain line cadence. Thus, while high-k is a new technology, **capacity challenges can be managed through equipment scaling**.

Other Process Innovations

EUV in 3D NAND: As 3D NAND surpasses 200 layers, traditional 193 nm lithography requires multiple patterning to form high-aspect-ratio staircase holes. Early signs suggest EUV may be adopted to improve pattern fidelity and alignment, potentially reducing overlay errors and rework rates—benefiting both yield and cost.

3D NAND Bit Density Trend (TLC, QLC)



Direct and hybrid bonding: These advanced packaging methods eliminate micro-bumps and allow direct chip-to-chip connection, lowering interconnect impedance and increasing bandwidth. However, they require **extremely tight surface planarity and alignment control**, making yield critical. If successful, hybrid bonding can simplify HBM interconnects and improve both yield and performance—but if poorly controlled, it can introduce new defects.

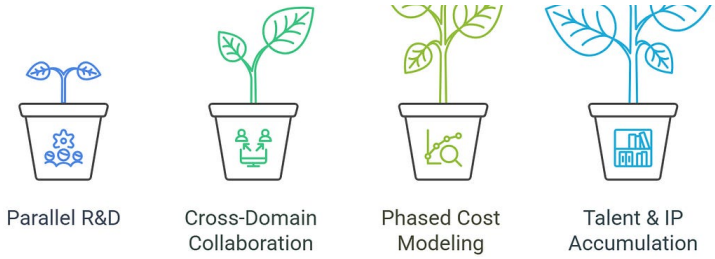
Overall, every new material or process innovation introduces **short-term variability and yield risk**, but offers significant long-term performance and cost advantages once stabilized.

Overall Impact and Conclusions

Materials and process innovations act as a **double-edged sword**: in the short term, they often cause yield drops and cost increases; in the long run, they enable better performance-to-cost ratios and expand production capacity limits. Companies must balance **yield, cost, and capacity** carefully.

In the AI era, **standing still means falling behind**—adopting new technologies is risky but inevitable. A smart strategy involves **small-scale pilot runs and rapid iteration**: introduce new processes on select lines, surface problems quickly, fix them rapidly, and gradually expand deployment. Close collaboration with the supply chain—especially materials and equipment vendors—is crucial for stable ramp-up and fast feedback cycles. Once stabilized, these innovations yield **higher yields, lower cost per bit, and more flexible capacity**, enabling the industry to meet AI's massive memory demands in the coming decade.

Risk Management and Road-Mapping for New Process Technologies



- **Parallel R&D and manufacturing**: Engage manufacturing teams early during R&D to ensure new processes consider yield and cost. Use pilot lines to validate EUV and hybrid bonding feasibility, and establish **yield KPI triggers** to drive structured improvements if early yields fall below targets.
- **Cross-domain collaboration**: New materials and equipment often involve external partners. Form **joint teams with lithography and materials suppliers** to address EUV yield issues, or work with packaging houses to co-optimize hybrid bonding. Data sharing accelerates problem-solving and shortens ramp time.
- **Phased cost modeling**: Build cost projections for each innovation—short-term investment vs. expected cost crossover point—to plan deployment pace. For example, apply high-k first to high-capacity products to amortize costs; ramp EUV tools gradually based on learning curves.
- **Talent and IP accumulation**: New technologies require new expertise. Companies should build in-house teams for EUV lithography, 3D packaging, and materials science while strengthening IP portfolios. Mastering these technologies early ensures **future market power** and protects against easy replication by competitors.

Strategic Implications for the AI Era

Under the wave of EUV, multi-stacking, advanced packaging, and high-k innovation, early adopters must be ready for **initial turbulence** but can gain long-term strategic advantage. Through careful planning and rapid optimization, companies can **turn new technologies into competitive weapons**, achieving yield improvement, cost optimization, and large-scale capacity expansion—fueling the AI economy of the next decade.

The **AI chip industry is undergoing an irreversible power realignment**. The GPU-centric, single-vendor supply chain is fragmenting, giving way to a new landscape driven by **OpenAI, custom ASICs, diverse HBM sourcing, and advanced packaging**. HBM is no longer a mere auxiliary memory but a **strategic asset** influencing chip design, packaging capacity, supply chain configuration, and geopolitics.

Samsung and SK hynix, by controlling memory capacity and technology cadence, are becoming **critical nodes** in the future AI ecosystem. TSMC

and OSATs secure their positions through advanced packaging integration. While NVIDIA remains ahead, its **overwhelming grip on the industry is loosening**.

2026 will mark a structural turning point, shifting the AI supply chain from a "GPU economy" to an "ASIC × HBM economy." Memory, packaging, power, interconnect, and capacity allocation will become the key battlefields over the next five years. Whoever controls **pricing power in this supply chain realignment** will dominate the next AI supercycle.

This is not merely a technological race—it is a **strategic, financial, and geopolitical contest**.