William Blair

AI INSIGHTS

Equity Research
Technology, Media, and Communications | Semiconductor and Infrastructure Systems

January 22, 2026
Industry Report

Sebastien Naji +1 212 245 6508
snaji@williamblair.com

Ana Bilbao +1 312 364 8598
abilbao@williamblair.com

Total Recall

How AI Is Supercharging Memory Demand

Contents

Introduction

While so much of the AI platform shift has been focused on computing demand, the traditionally slower growth and less exciting memory market is starting to have its day in the sun. Advancements in AI compute are increasingly hindered by memory bottlenecks, revitalizing memory as an area of innovation as vendors race to develop high-bandwidth solutions capable of addressing the data-intensive needs of AI models. With AI inference demand still in a phase of early adoption, we believe the next several years will be characterized by a stepwise increase in global memory capacity with new technologies emerging to deliver large volumes of data to AI computing infrastructure.

The William Blair technology team has previously explored AI-driven advancements in compute, detailing how the AI platform shift require a deeply integrated network of next-generation technologies (From Chips to Systems: How AI Is Revolutionizing Compute and Infrastructure) and offering a GenAI primer (Generative AI: The Next Frontier of Innovation). In this report, we offer a detailed explanation of past, present, and next-generation memory solutions and the evolving market dynamics as AI fundamentally changes this previously quiet industry.

In conjunction with this report, we are initiating coverage of three companies: Micron, Rambus, and Silicon Motion.

Key Takeaways

Bandwidth the Key AI Memory Bottleneck
AI performance is increasingly limited by memory bandwidth, since large language models (LLMs) require moving enormous volumes of data for relatively little computation.. Training and inference involve repeatedly streaming massive amounts of data—e.g., model parameters, context, and KV cache—through accelerators with relatively low data reuse. As a result, performance is often limited by how fast data can be delivered rather than how many compute floating-point operations per second (FLOPS) are available. Since the 1990s, processor performance has scaled at a much faster rate than memory bandwidth, leading to the development of the "memory wall." While DRAM bandwidth has improved by a factor of 1.6x every two years over the last two decades, compute FLOPs have increased 3x every two years, resulting in a compute-memory bandwidth performance gap that has reached a more than 600:1 ratio. As models have gotten larger over time, GPUs issue thousands of memory requests in parallel, quickly saturating memory interfaces. When memory bandwidth is insufficient, expensive compute units sit idle and drive down the utilization of GPU clusters.

HBM Improves Memory Access for AI Accelerators
To help address the memory wall problem, memory vendors have developed a new type of dynamic random-access memory (DRAM), high-bandwidth memory (HBM). HBM helps address the memory wall by increasing effective memory bandwidth—DRAM memory stacks are integrated in-package with the processor and connected via thousands of short, low-power wires on a silicon interposer (TSVs, or through-silicon vias). Because of its wide interfaces (16 times wider than traditional double data rate [DDR] channels), HBM enables GPUs to quickly access large volumes of data. With HBM4, vendors are doubling the memory bus capacity to 2,048 bits, which in addition to an increase in pin speeds (beyond 11 Gbps for HBM4) should help drive a significant improvement in memory bandwidth for new AI accelerator deployments in 2026.

Memory Vendors Shifting Focus Toward More Profitable HBM
The success of HBM, first introduced by SK Hynix and followed by offerings from Micron and Samsung, has pushed the major memory vendors toward HBM because it delivers higher revenue and profit per wafer. Today, a single HBM3E stack (used in Nvidia Blackwell GPUs) delivers roughly

1.2–1.3 TB/s, rising to more than 2.0 TB/s with HBM4, versus only about 50–100 GB/s per DDR5 DIMM. Economically, HBM sells for ASPs that are roughly 3-4 times that of traditional DDR DRAM, and despite higher TSV and packaging costs and lower yields, gross margins are materially higher because AI customers value bandwidth far more than raw capacity. We estimate HBM gross margins are in the 55% to 65% range today, compared to historical DRAM gross margins between 25% and 45%. This is driving a dramatic mix shift in capacity allocation at the major memory vendors. SK Hynix, which controls approximately 60% of the HBM3 market, now allocates more than half of its DRAM wafer output to HBM, and Samsung is restructuring its DRAM roadmap to prioritize HBM despite qualification challenges. This is resulting in tighter capacity for more traditional DRAM solutions, with companies exiting lower-profit segments entirely to refocus resources—in December 2025, Micron announced it was terminating its Crucial consumer flash business.

HBM Base Die Becoming a Critical Value Capture Battleground

The base logic die in HBM has evolved from a relatively simple interface into a strategically important compute-adjacent component with room for significant customization. Thus far, HBM base dies have relied on older memory nodes and handled basic functionality (e.g., managing electrical connections, translating high-level GPU instructions, distributing signals across TSVs); most intelligence resides in the GPU's memory controller. With HBM4, there is a shift to a more capable and performant base die built on leading processor node technologies (Micron and SK Hynix will utilize TSMC). In the upcoming generation of HBM, the memory controller will be integrated into the base die. For processor designers like Nvidia, AMD, and Marvell, this opens the door to deeper codesign, where customizing the base die for a larger memory beachfront or to offload functionality from the core processor can drive better performance for the chip system. As more vendors focus on taping out custom HBM base dies for their own processors/infrastructure stacks, HBM vendors may over the medium to long term see even this part of the market become increasingly commoditized.

AI Memory Momentum Extends to DRAM Writ Large

While HBM captures the spotlight, the AI buildout is also pulling through substantial demand for high-performance non-HBM DRAM. Large AI clusters are built on a combination of many chips beyond the GPU, including CPUs, network interface cards (NICs), data processing units (DPUs), and storage controllers. These solutions pull through significant memory demand—e.g., each Nvidia GB300 CPU requires 280 GB of DDR5 memory (more than 20 TB per NVL72 rack). In addition, newer chip designs are incorporating other types of DRAM; Nvidia's new Rubin CPX GPU, which addresses long-context prompts, swaps usage of more expensive HBM for larger capacity GDDR7. While bit demand growth should be up double digits over the next few years, capacity for traditional DRAM is tight. With memory vendors focusing more of their capacity on HBM (which requires 3 times as many DRAM wafers for the same capacity), ASPs for traditional DRAM solutions have seen tremendous upward pressure (TrendForce estimates 55%-60% ASP growth for conventional DRAM in the first quarter of 2026). We expect high prices for DRAM memory to remain the norm for the next few years as new manufacturing capacity takes at least another 18 months to come online.

NAND Storage to Benefit From AI

As inference becomes the dominant AI workload, demand is also expanding toward NAND and storage to support persistent, large-scale model data. Inference workloads require fast access to massive number of model weights, embeddings, retrieval-augmented generation (RAG) datasets, and—critically—persistent or reusable KV cache, all of which are quickly exceeding the capacity of HBM or system DRAM. This is driving higher attach rates of high-performance NAND SSDs. Nvidia's Inference Context Memory Storage Platform (announced in January 2026 at CES) reinforces the increasingly important role of SSDs, by decoupling context from HBM/DRAM and enabling storage as an active tier. It uses BlueField DPUs to expand inference context to memory form-factors that have a lower cost per bit than DRAM, while maintaining system performance through

software optimizations. Sustained growth for flash storage should benefit NAND and SSDs from vendors like Micron, SK Hynix, Samsung, Kioxia, and SanDisk. As inferencing is adopted by enterprises, McKinsey estimates a 35% CAGR in the enterprise SSD (eSSD) market from 2024 to 2030.

Memory Controllers and Interfaces Becoming More Sophisticated

Memory controllers and interface technologies have evolved from simple interfaces to critical, high-value system components. Today, memory management technology increasingly takes on compute-adjacent logic and serves as an active gatekeeper of AI performance. In addition, memory controllers/interfaces must be protocol agnostic and able to manage heterogenous connections to a broad mix of processors (CPUs, GPUs, and DPUs). In DRAM, while controllers remain integrated into the processor, higher speeds and new form factors (DDR5, MRDIMMs, HBM) require increasingly complex PHYs and timing logic, driving up the value of memory interface technologies. Meanwhile, in NAND, the rise of less expensive and denser solutions (like quad-level cell, or QLC) has forced controllers to compensate for weaker media characteristics through heavier firmware, signal processing, and adaptive algorithms. This growing intelligence shifts differentiation and value capture toward the controller/interface layer, allowing vendors like Rambus, Silicon Motion, Phison, and Marvell with strong IP, firmware, and system-level expertise to command higher margins and embed themselves more deeply in customer platforms.

Next-Gen Memory Solutions Designed to Address Memory Wall

While innovations like HBM have made significant strides in scaling memory bandwidth and capacity, they fall short of completely filling in the memory wall gap. Next-generation solutions have emerged to help improve this critical bottleneck. These include compute-in-memory (allocating processing directly to the memory chip), neuromorphic chips (modeled after human neurons, uses spiking neural networks to process information), high-bandwidth flash (which stacks NAND into an HBM-like structure), and resistive RAM (storing data as resistance rather than voltage). In addition, the use of Compute Express Link (CXL) is emerging as higher-bandwidth memory interconnect fabric that allows CPUs and accelerators to share access to memory over PCIe. We highlight companies like Anaflash, Astera, BrainChip, Cerebras, Credo, Crossbar, EnchargeAI, SynSense, Syntiant, Upmem, and Weebit Nano as helping push the innovation envelope in next-gen memory solutions.

Brief History of Memory

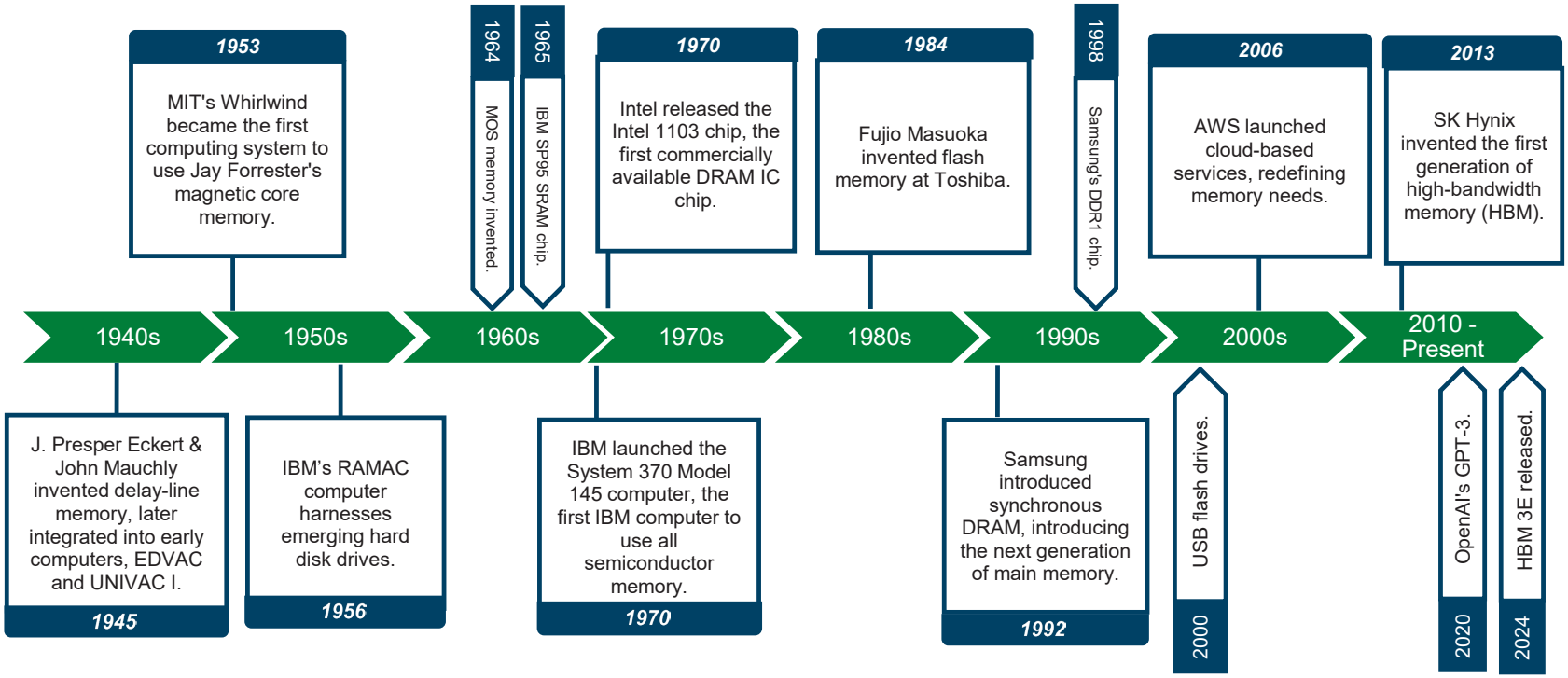
Early History of Memory Technologies

Since the advent of computing, memory has been integral to the execution of the most elementary programs. The advancement of memory is primarily measured through the expansion of bits and the advancement of read and write speeds. “Bits,” or binary digits, are the most fundamental unit of data storage; think of a bit as the unit cell of computer memory. Bits are essentially patterns of two possible values—i.e., 0 or 1—capable of storing all data. As this data grow in complexity, the number of bits increases, though the possible value of the bit is always either 0 or 1. The more recognizable unit of storage—bytes—is simply a sequence of 8 bits (e.g., 00000001). The speed at which these bits can be coded into the memory is the write speed, and the speed at which they can be retrieved and delivered to the processor is the read speed.

The earliest and most rudimentary form of memory developed in the 1940s was a paper punch card. The card functioned as a standardized grid storing 960 bits of data. While costs were low enough to keep punch cards relevant for decades, this method was inevitably cumbersome as cards were limited in capacity and write-once: data could not be re-programmed or rewritten. Given that it would require tens of thousands of punch cards to store data at the megabyte scale (enough to store a singular, high-quality iPhone picture), use declined as compute needs advanced.

6 | Sebastien Naji +1 212 245 6508

Exhibit 1
Total Recall
History of Memory Systems



Sources: William Blair Equity Research, Computer History Museum (CHM), Intel, IBM

In 1944, J. Presper Eckert's development of delay-line memory became the first widely accepted memory system, notably used in early computers such as EDVAC. Delay-line memory relied on pressure wave propagations for reading and writing the data. There were two key detriments to this system: 1) the sequential nature (or cyclic-access memory) limited CPU access to 1 bit at a time, and 2) storage density was restricted by the physical limit of wave propagations. Delay-line memory was rendered obsolete in the 1950s, illustrating two important pillars of subsequent memory technology development: movement toward random-access memory (RAM) and developing a scalable, high-capacity method of bit storage.

Magnetism and Striving for RAM

Aptly named, magnetic core memory used magnetic donuts (or cores) connected by wire interconnects; reading and writing were completed using electrical currents in place of delay-line memory's wave propagations. Key advancements of magnetic core memory directly solved the shortfalls of delay-line memory, i.e., random access to any bit (introducing RAM) and scaled capacity achieved by stacking the grids. Magnetic core memory was first applied in 1953 to MIT's Whirlwind computer, capable of storing roughly 4 KB of memory, about one-millionth of the memory of a modern PC. Magnetic core memory remained the industry standard for two decades, cutting the cost per bit to 100th of the initial cost and inciting contemporary memory access methods, RAM.

Magnetism was separately yet similarly leveraged by Eckert in 1951 to inscribe the binary code into the proprietary memory storage system he developed for UNIVAC, magnetic tape memory. The cheap and compact nature of the magnetic tape permitted data archiving, but access required the rotation of hundreds of feet of tape. This so-called seek issue was temporarily ameliorated with the development of the magnetic drum, which used a revolving barrel to reduce seek times. Though only experiencing a short stint of commercialization, the magnetic barrel directly led to its successor, the hard disk.

Evolution of Storage: Advancing to Solid State Solutions

Developed in 1956, the hard disk's configuration of a central axis with rotating platters mirrored that of the magnetic drum. By nature of the disk structure, the large surface area allowed stacking to exponentially expand the capacity. Read and write heads traveling alongside the length of the stack could wedge between the spinning disks, reducing the seek time to find the indicated bit. While speed was limited to the rotation time on the disk, this mechanism was relatively fast for storage, enabling its dominance by the 1970s. The floppy disk operated in much the same manner, but on a different magnetic medium. By the 1990s, the floppy disk was the preferred mechanism of portable storage.

To briefly diverge from magnetic inscription, optical storage devices (most notably CDs and DVD), became ubiquitous for music- and entertainment-specific storage in the 1990s. Bits were stored as divots that reflected light signals captured by the optical detector. The current landscape of memory storage is defined by solid state technologies. By replacing the moving parts in the device with integrated circuits (ICs), seek time became essentially negligible, underscoring the prominence and subsequent rise of semiconductor memory.

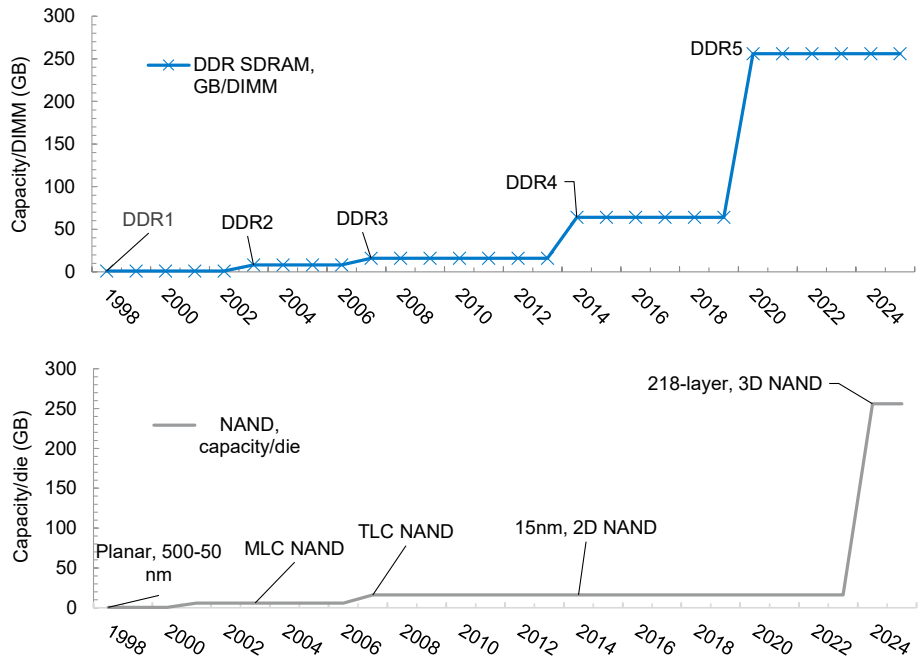
Rise and Reign of Integrated Circuits

Semiconductor memory entered the scene in 1964 with Fairchild Semiconductor's (now Onsemi) development of metal-oxide-semiconductor field-effect transistors (MOSFETs), a technological advancement that permanently tilted the scales toward semiconductor memory. Semiconductor memory stores bits of data within MOS memory cells composed of one or more transistors on a silicon chip. Each bit of data is located at a specified address that is quickly recognizable and retrievable by the CPU. The ongoing dominance of semiconductor memory commenced in the early 1970s.

Depending on the configuration of the memory cell, RAM can either be static (SRAM) or dynamic (DRAM). Commercial use of SRAM commenced in 1965 with IBM's introduction of the SP95 SRAM chip. Development of DRAM memory cells followed with the 1970 debut of the first commercial DRAM IC chip, the Intel 1103. DRAM development was further specialized with the introduction of synchronous dynamic RAM (SDRAM) by Samsung in 1992.

In 1998, DRAM technologies were revolutionized with the introduction of Samsung's Mbit DDR SDRAM. Intuitively, DDR SDRAMs are capable of double the bandwidth, quickly becoming the preferred and most used form of volatile main memory. Successive generations of DDR SDRAM—i.e., DDR2 (2003), DDR3 (2007), DDR4 (2014), and DDR5 (2020)—have represented a substantial bandwidth improvement (see exhibit 2). DDR5 boasts memory capacity in the range of 64 to 256+ GB, roughly a million-times capacity improvement from delay-line memory with kilobits of capacity. The most cutting-edge main memory solution, optimized for AI-specific use-cases, is HBM, introduced by SK Hynix in 2013, then implemented in AMD Fiji GPUs two years later.

Exhibit 2
Total Recall
Dramatic Generational Gains Across DRAM and NAND



Source: William Blair Equity Research, TechInsights

While main memory is almost entirely volatile RAM, it is essential to acknowledge the concurrent development of nonvolatile, read-only memory (ROM). The modern basis of nonvolatile memory was established with the development of floating gate memory cells in 1967 by Dawon Kahng and Simon Sze at Bell Labs (formerly Alcatel-Lucant, now Nokia Bell Labs). This floating gate technology permitted Toshiba's development of NOR and NAND flash in 1984 and 1987.

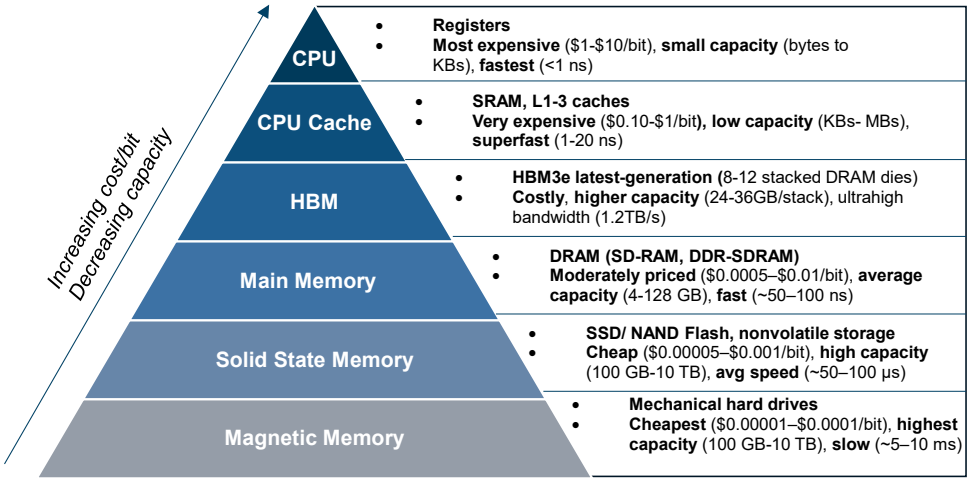
NOR and NAND, named after the logic gates they use (NOT OR and NOT AND, respectively), differ in their memory cell connections and thus their reading and writing capabilities. NAND, the most popular nonvolatile flash memory, is sequential access by nature of the serial connection of the memory cells, sacrificing reading speed for data density. NOR, on the other hand, can access data in parallel but has lower storage density than NAND.

In 2013, Samsung was the first to commercialize 3D NAND flash with its 24-layer V-NAND technology. Instead of continuing to shrink planar (2D) NAND transistors—which was becoming physically difficult—Samsung stacked memory cells vertically, massively improving density and endurance and allowing continued growth in flash capacity.

Bits and Bytes: The Memory Hierarchy

The memory hierarchy (exhibit 3) is the way computing systems organize memory into layers that trade off speed, capacity, cost, and proximity to the processor to deliver high performance efficiently. At the top are registers and caches (L1–L3), which are extremely fast but tiny and expensive, sitting directly on or near the CPU/GPU. Next is main memory (DDR or HBM), which is much larger and less expensive per bit but slower and farther away. At the bottom are storage layers (solid-state drive [SSD], then hard disk drive [HDD] or object storage), which offer massive capacity at a very low cost but with orders-of-magnitude higher latency.

Exhibit 3
Total Recall
The Memory Hierarchy



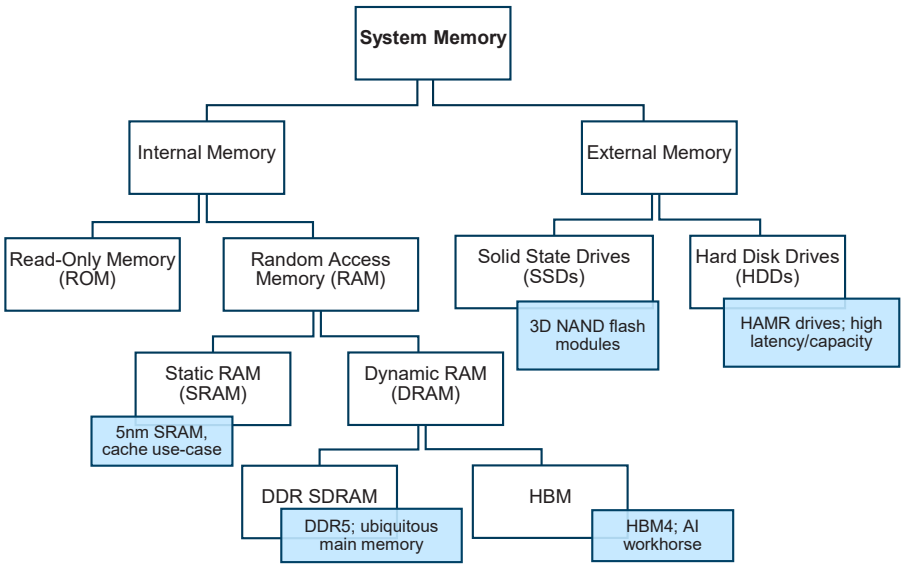
Sources: William Blair Equity Research, Engineering LibreTexts

The most fundamental distinction in the memory hierarchy is between internal and external memory. As the names suggest, internal memory stores data inside the server/computer and external memory relies on data stored in an auxiliary location.

Note that, while related, the separation between internal and external memory is not synonymous with volatile versus nonvolatile systems. Volatile memory—the most ubiquitous form of main memory—requires power to maintain the stored data. In contrast, nonvolatile memory can retain information permanently in the absence of power. Volatile memory tends to be used for internal

memory (DRAM), while nonvolatile memory is typically used for SSDs and storage. That said, nonvolatile internal memory does exist in the form of ROM and embedded flash memory systems. Likewise, volatile external memory also exists, though these systems are thinly used.

Exhibit 4
Total Recall
System Memory Flowchart



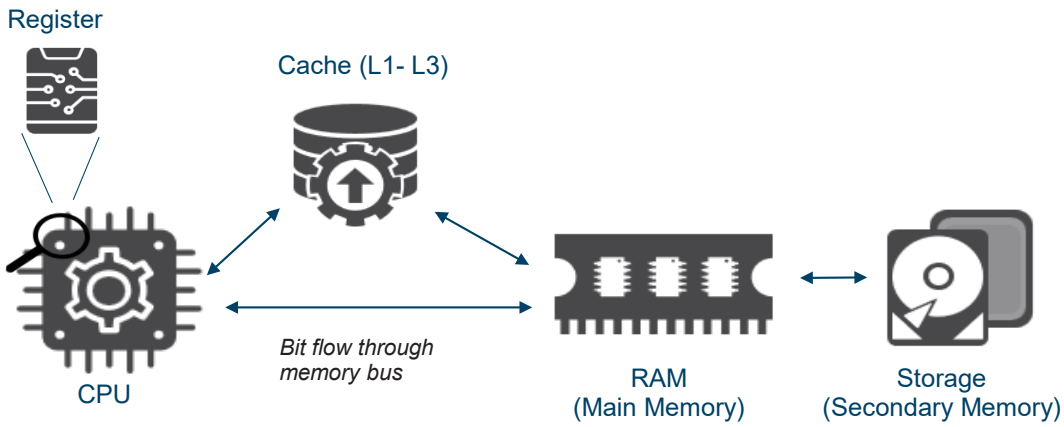
Source: William Blair Equity Research, Princeton Department of Computer Science

Internal Memory

Internal memory is built into the computer/server and used for everyday operations. Main memory is composed of bytes. Each byte represents a sequence of eight bits, the most fundamental memory unit. A bit can be one of only two values, either a zero or one (determined by the voltage of each memory cell). Each byte (i.e., 00000001 or 0100010) represents a distinct address that is retrievable by the CPU during the "fetch" stage of running a program.

Traditionally, when a CPU needs data, it issues a load request that is first checked against its on-chip caches (L1, then L2 and L3). If the data is not found there, the request is forwarded by the integrated memory controller to main memory at the specified address. The retrieved data are returned over the memory interconnect to the processor, placed into the cache hierarchy, and then consumed by the execution units. Caches are small, fast memory structures that exploit data locality, enabling the CPU to avoid repeatedly accessing slower main memory.

Exhibit 5
Total Recall
Data Retrieval Process



Sources: William Blair Equity Research, Journal of Information Systems Engineering & Management

Registers

Registers are the smallest and fastest type of memory systems, with much lower latency coming at the expense of a much higher price per bit. The registers are located inside the CPU cores themselves, with each core assigned a set of registers. To quantify the "superfast" speed between the CPU and registers, we can use a CPU clock cycle, which is the time it takes to perform one operation. The register's latency corresponds to less than 1 CPU cycle or roughly 0.3 to 1 nanosecond of latency.

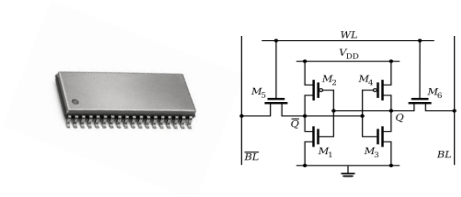
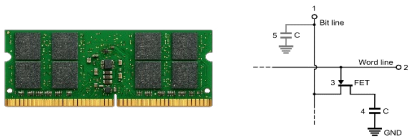
For a 64-bit x86 CPU architecture there are 16 general-purpose registers, each capable of storing 64 bits (or 8 bytes), totaling 128 bytes per CPU. To contextualize, 128 bytes are not enough to store an email or webpage. Even a high-resolution image requires storage on the megabyte scale. By design, registers are adept at temporarily storing small bits of pertinent information. Since registers occupy prime real estate inside the CPU, every dimension increase in memory capacity directly limits valuable square footage for compute.

Cache and SRAM

A cache is a small and temporary memory element located on die or near the CPU. The cache essentially sits between the CPU and main memory. The cache is subdivided into different levels based on its distance to the CPU. An L1 cache boasts ultra-fast speeds comparable to registers located in CPU itself, an L2 and L3 cache is typically located on die but further from the cores, and an L4 cache is typically off die and features the highest latency.

Static random-access memory is the architecture predominantly used for caching. A typical SRAM memory cell, storing one bit, is composed of six transistors interlocked in a flip-flop circuit. In contrast, DRAM is made up of one transistor and a capacitor. Because capacitors slowly leak energy, the value must be refreshed every few milliseconds or the data are lost. With SRAM, because data are stored in stable cells made up of stable transistors, the value remains constant, or static. SRAM still requires power but is a more stable and faster form of memory access. However, SRAM's six-transistor architecture, which enables its low latency, also makes it a bulkier and less dense form of memory, driving up the cost per bit.

Exhibit 6
Total Recall
SRAM vs DRAM

Static Random Access Memory (SRAM)	Dynamic Random Access Memory (DRAM)
	
• Primarily used for cache memory • Faster read and write speeds (1-10 nanoseconds) • Smaller capacity (a few MBs) • Less power consumption, no refreshing of charges • Higher cost per bit ($0.10-1 per MB) • 6 transistors, larger cell size	• Primarily used for main memory • Slower speeds (10-100 nanoseconds) • Larger capacity (several GB) • Higher power consumption, requires refreshing • Lower cost per bit (~$0.0005-$0.02 per MB) • 1 transistor & 1 capacitor, 4-6x higher density than SRAM

Source: William Blair Equity Research

SRAM and Nvidia's Groq acquisition

SRAM has received increased attention over the last few months, following Nvidia's acquisition of AI accelerator company Groq. A core part of Groq's LPU chip architecture is the heavy reliance of on-chip SRAM to store model weights and activations close to its compute units, enabling deterministic, low-latency execution without the stalls and unpredictability of off-chip memory accesses. By keeping most working data in SRAM, Groq avoids the need for complex caches or high-bandwidth external memory systems, which simplifies scheduling and allows the compiler to precisely orchestrate data movement. The downside is that SRAM scales poorly in capacity. Compared to DRAM or HBM, SRAM is extremely area intensive, expensive per bit, and power hungry, which limits the total model size that can fit on a single chip and drives up die size and cost. As a result, we expect that Groq's core technology is likely to be used for inference at the edge, where memory capacity needs are significantly lower than in the data center.

pSRAM

A niche variant of SRAM is pSRAM or pseudo-static RAM. Internally it is built like DRAM (i.e., capacitor-based memory), but it includes an integrated self-refresh controller, which replaces the need for special refresh commands or DRAM controllers—allowing it to mimic SRAM behavior. It tries to blend the benefits of the two, with a lower cost per bit and simpler to use than raw DRAM. Edge devices include embedded systems, IoT, and automotive, and tend to be the primary use-case for pSRAM. Edge AI semiconductor company Ambiq Micro, for example, leverages pSRAM to act like HBM as part of its Atomiq SoC.

DRAM

DRAM, often referred to as main memory, consists of memory cells solely containing a single capacitor and transistor—a much more compact architecture compared to the 6-transistor SRAM. DRAM typically sits on the board next to the processor.

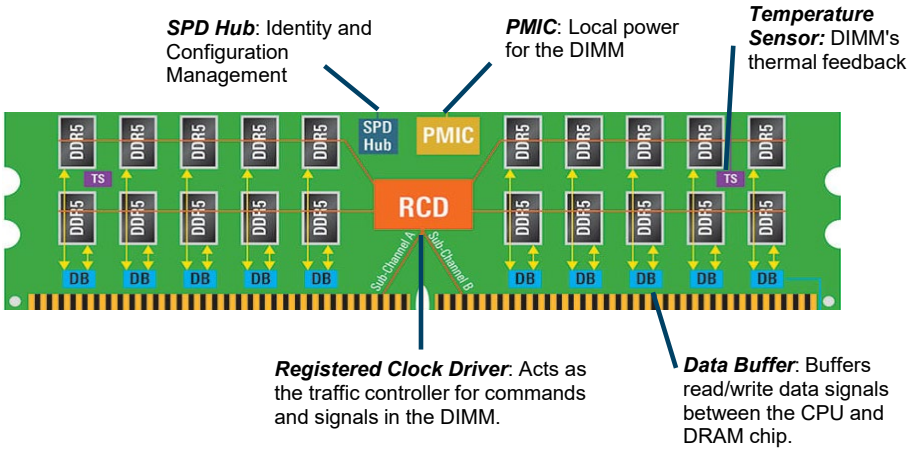
Since DRAM stores bits using a capacitor (like a battery holding charge), charge leakage erodes the bit values over time. To address this, a refresh circuit periodically reads and rewrites all DRAM memory cells. This refresh aspect, coupled with the detachment from the processor, is a major

contributor to DRAM's lower latency (about 150-300 nanoseconds) and heightened power sensitivity. That said, its simpler architecture enables higher storage capacity and a better price per bit than SRAM or caches.

DIMMs

The most common form-factor is the RAM stick, which connects into the computer/server board. The dual in-line memory module (DIMM) is the physical circuit board that combines multiple DRAM chips into a compact unit. Unlike older single in-line memory modules, which shared electrical contacts on both sides, DIMMs have separate electrical contacts on each side of the module, enabling a wider data path and higher capacity. Each DIMM plugs into a dedicated memory slot on the motherboard and communicates with the CPU through a dedicated memory controller. DIMMs are designed to be easily replaceable or upgradable, making them a modular, standardized way to expand system memory.

Exhibit 7
Total Recall
DDR5 Chipset



Source: William Blair Equity Research and Renesas

UDIMM (unbuffered DIMM) is used for lower capacity consumer applications, such as PCs and desktops. RDIMMs (registered DIMMs) are designed for server applications—they use a buffer chip to time and redrive signals from the CPU, preserving signal integrity in high data rate use-cases.

MRDIMMs

MRDIMM (multiplexed rank DIMM) is a new DDR5-based module that boosts memory bandwidth by using a built-in multiplexer chip to "rank-multiply" the DRAM. This allows the memory controller to access multiple ranks as if they were fewer ranks, effectively increasing data throughput without increasing pin count or controller complexity.

An MRDIMM is like a highway that connects several neighborhoods to the city center. With a regular RDIMM (registered DIMM), the road can only let cars from one neighborhood through at a time, so each group has to wait its turn before the next can go. MRDIMMs add a traffic controller—like a smart switch—at the entrance of the highway that can quickly alternate between neighborhoods, letting cars from multiple areas use the road back-to-back with almost no wasted time. This keeps the road busier and delivers more cars per minute without making the road any wider or building a second one, effectively boosting bandwidth without raising the memory clock speed. MRDIMMs are still in the early innings of adoption, with widespread commercial deployment expected in the second half of 2026.

LRDIMMs

Load-reduced DIMMs (LRDIMMs) are another type of specialized RDIMM aimed at reducing the electrical load on the CPU; LRDIMMs sacrifice some latency for huge gains in capacity, making them appropriate for enterprise servers that can take a small hit to latency. LRDIMMs address the same bottleneck as MRDIMMs: the strain on the memory controller of the CPU to drive signals to all ranks on the DIMM module. The LRDIMM adds an additional data buffer on top of the registered clock driver (see exhibit 7 on page 13) to slow down and redrive the signal, maintaining the signal integrity in high data rate environments. The data buffer moderates the signal from the CPU to each rank.

The use of a data buffer reduces latency as communication between the CPU and memory is now moderated. The increased latency is compensated for by gains in capacity. For example, if the CPU distributed the signal to eight memory ranks in an RDIMM with the use of a data buffer in the LRDIMM, the CPU could direct the same signal to one data buffer, which in turn distributes the data to the same ranks, thus increasing capacity by eightfold for each buffer. In DDR4 and preceding generations, LRDIMMs were the standard for high-capacity enterprise servers. As the shift to DDR5 continues, MRDIMMs are incrementally gaining market share from LRDIMMs.

Ongoing transition from DDR4 to DDR5

There are different types of DRAM, with the most widely used form being DDR synchronous DRAM (DDR-SDRAM). DDR-SDRAM has the capability to transfer data on both the rising and falling edge of the CPU clock cycle, implying double the rate of data transfer relative to single-edge systems.

The fourth generation of DDR-SDRAM (referred to as DDR4) remains the most broadly deployed architecture despite commercialization of DDR5 starting in 2020. Given the stickiness of DDR technologies, shifts to new generations happen every five to seven years, with intra-generational improvements taking place every two to three years.

Early adoption of DDR5 was hindered by a multitude of supply-side dynamics culminating in low initial market penetration. These hurdles included 1) initial ASPs that were 2-3x that of DDR4 for the same capacity, 2) small initial performance gains, 3) yield issues, 4) shortages of other chipset components (e.g., power management ICs), and 5) lack of compatible CPUs until 2022-2023 with the launch of Intel's Alder Lake for consumer applications and Intel Sapphire Rapids and AMD Genoa in the server market.

Exhibit 8
Total Recall
DDR4 vs DDR5

	DDR4 (2014)	DDR5 (2020)
Speed	1.6-3.2 GT/s 0.8- 1.6 GHz clock	4.8-8.4 GT/s 1.6-4.2 GHz clock
Max Die Capacity	16 Gb	64 Gb
IO Voltage	1.2 V	1.1 V
Power Management	On motherboard	On DIMM PMIC
Channel Architecture	1 (64-bit) data channel per DIMM	2 (32-bit) data channels per DIMM

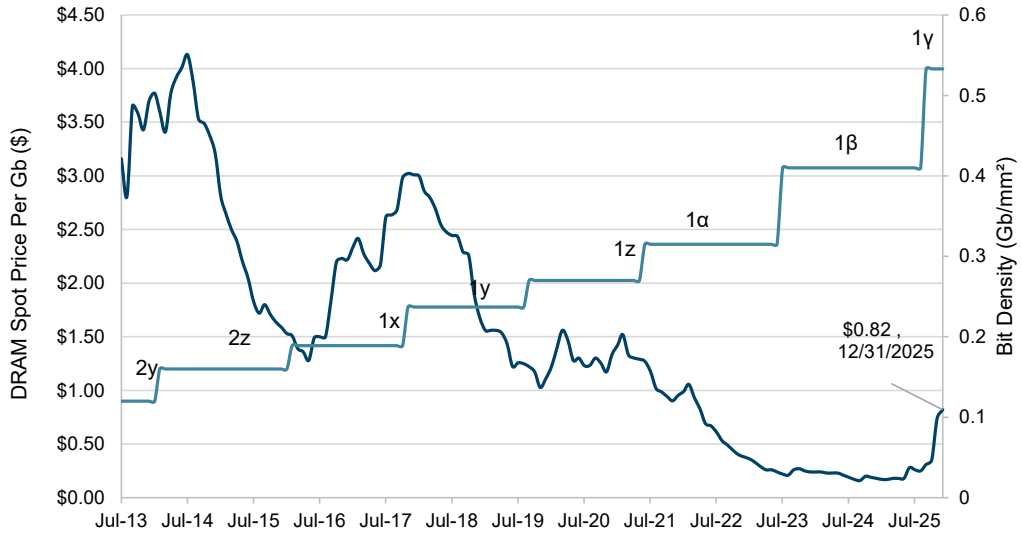
Sources: William Blair Equity Research

While initial penetrations of DDR5 were lower than expected (8% in 2022), shipments of DDR5 inflected in the second half of 2024, largely due to traction in new servers. DDR5 is forecast to represent 58% of exabit shipments in 2025, according to IDC.

DRAM and DDR are broadly viewed as commodity solutions, where long-run price erosion is the baseline, but cycles can cause temporary dislocations in pricing. We are seeing this today with the massive rise in memory demand having driven up pricing per GB for a wide range of solutions, new and old. Prior to the current dislocation in pricing across the memory market, DDR5 ASPs were priced at a 30%-40% premium, inclusive of increased dollar content per DIMM due to more companion components in DDR5.

Nonetheless, as exhibit 9 highlights, DRAM pricing erosion over the last decade has been driven by density scaling outpacing demand growth, with cyclical spikes repeatedly failing to reset the long-term price floor. As bit density per mm^2 rises stepwise with each process generation (2z → 1β), the cost to supply a bit falls persistently, forcing spot prices per Gb to trend downward despite intermittent upcycles (e.g., 2017–2018, 2021). Each recovery peaks lower in real terms and is followed by sharper declines as new capacity and higher-density nodes flood the market, indicating that technology-driven supply elasticity dominates demand-driven pricing power.

Exhibit 9
Total Recall
DRAM Spot Prices Against Backdrop of Node Development



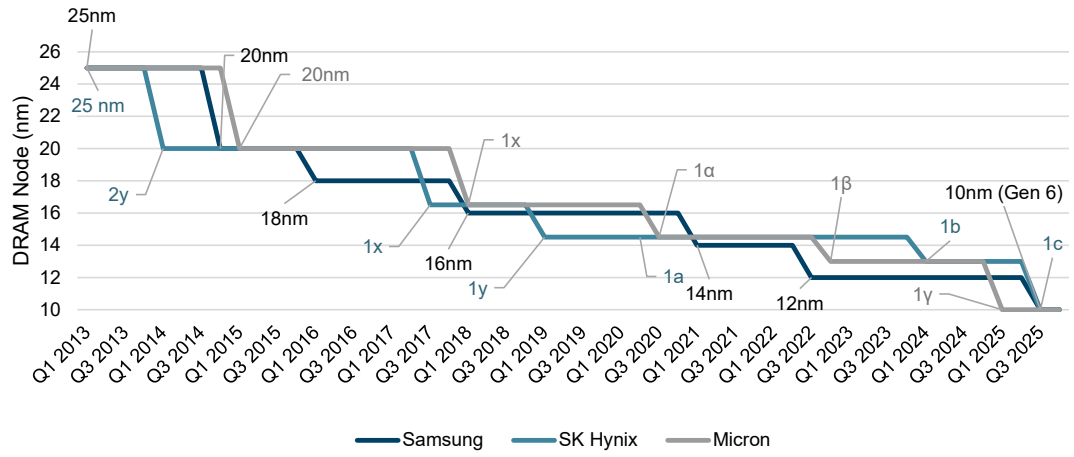
*Based on DDR3 (4 GB) spot prices & SK Hynix process node nomenclature
Sources: William Blair Equity Research, DRAMeXchange

DDR5 advantages

One main advancement of DDR5 compared to legacy generations is the redesign of the DIMM channel architecture. In DDR4, there is one channel on each DIMM; each channel represents a 64-bit data bus. A data bus is simply a set of wires that transmits data between the CPU and DIMM, essentially shuffling data between the memory and processor in bundles of 64 bits at a time.

In DDR5, however, the data bus is effectively split; instead of one 64-bit bus, there are now two 32-bit buses attached to each DIMM. While the cumulative number of bits that can be transferred at once remains the same, the effective bandwidth is improved as the memory controller can now issue requests in parallel to smaller segments of data. This approach leads to large improvements in latency for small data transfers.

Exhibit 10
Total Recall
DRAM Node Advancements Across Major Memory Vendors



Source: Company reports and William Blair Equity Research

HBM

High-bandwidth memory has grown increasingly prominent as the main memory system capable of scaling to the bandwidth needs of GPUs, AI accelerators, high-performance computing systems, and field-programmable gate arrays (FPGAs). Whereas traditional RAM systems are two-dimensional—stacked next to each other in a flat package off die—HBM DRAMs are stacked vertically, allowing proximity to or even direct placement on the accelerator die.

The DRAMs in the HBM stack are purpose-built and cannot be repurposed for standard DRAM since they require twice the die surface area relative to same-capacity conventional DRAM. This is because of the need for TSVs that vertically interconnect the dies into an HBM stack—TSVs consume more silicon area, forcing wider routing pitches, and reducing overall packaging density.

Thus, the DRAM dies are not interchangeable with DDR4 or DDR5 DRAM solutions. HBM has quickly become the fastest-growing memory product, driven by strong attach rates to GPUs and the ongoing buildout of increasingly larger AI clusters. See page 21 for a more in-depth analysis of HBM.

External Memory/Storage

External memory—colloquially called storage—sits at the base of the memory hierarchy, offering lower cost per bit, greater capacity, and greater data permanence. Because external memory has lower accessibility and higher latency than internal memory, it is typically used for data that will not be immediately required by the CPU.

This behavior is explained by the principle of "locality of reference," which is fundamental to the organization of the memory hierarchy. Locality of reference holds that memory access is not random; programs tend to repeatedly access the same data (temporal locality) or nearby data (spatial locality) over short periods. This predictable access pattern allows the processor to optimize where data is stored. Recently accessed data is retained in the fastest, most readily accessible storage—such as registers or caches—while data that has not been accessed, along with its neighboring addresses, is progressively demoted to slower, external memory.

When the memory controller of the CPU is in search of data, it first checks the most accessible and fastest memory: the registers, then the cache, in a process called "cache lookup." If the necessary bits are not located in the cache—i.e., a cache miss—the operating system (OS) will then check the next-most-accessible location, main memory. If the data is too large to be stored in RAM or has not been accessed in a long time, a page fault is triggered. The page fault signals the OS to intervene to locate the data from external memory (SSD/HDD).

Data in external memory are stored in pages. Once the OS finds the necessary page of data, it loads it into the RAM, swapping out an older page that is then sent to storage. Data stored in the main memory is now retrievable by the CPU, allowing the program to continue running. Loading pages from SSDs and HDDs into main memory takes microseconds and milliseconds, respectively; this is comparatively 1,000 to 1 million times slower than retrieving data from the main memory. Therefore, once the OS steps in to load pages, it locates any related data that it thinks the CPU might need in the near term and loads it into the main memory as well, attempting to minimize future retrievals from external memory.

SSDs and NAND flash

Across the client and mobile landscape (and increasingly in enterprises), SSDs are the predominant storage technology. These SSDs are built on NAND flash and integrate a flash controller to help manage communication with the CPU and execution of data reads/writes from the flash.

Upon receipt of data from the CPU, the controller of the SSD maps the address to the physical NAND pages, inscribing the bits to the NAND memory cells. The "flash" aspect refers to the ability to quickly erase old data in bulk, streamlining rewriting of data points. Flash memory's ability to erase and rewrite new data was a key advantage over its predecessor, erasable programmable read-only memory.

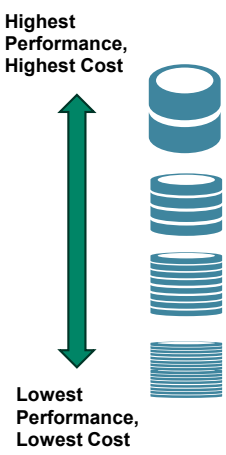
While the function of the NAND memory cell is the same as that of DRAM or SRAM cell—to store bits of data—it is built on a different silicon architecture made up of floating-gate MOSFETs. The floating gate traps the electrons in the memory cell, preventing charge erosion since the electrons are "gated." Once enough electrons are in the memory cell, the voltage threshold is reached, turning the transistor "on." If the voltage threshold is not reached, the memory cell is uncharged. For single-level cells, this concept is simple: if the memory cell is charged, it will hold a bit value of 1; if it is uncharged, it will hold a value of 0.

Different types of flash

NAND memory cells diverge from DRAM and SRAM cells in that they can hold more than one bit per cell. This works by introducing intermediate charge states; instead of just charged or uncharged, now the memory cells can hold different levels of charge.

For example, in a multi-level cell (MLC), which stores two bits per cell, the cell can represent four distinct charge levels instead of just two. In an MLC, those two extremes still exist (e.g., 00 for the lowest charge and 11 for the highest), but the voltage range between them is subdivided into two additional, precisely defined intermediate charge states that represent 01 and 10. By encoding multiple bits in a single physical cell through finer charge discrimination, MLC increases storage density without increasing the number of cells.

Exhibit 11
Total Recall
Different Types of NAND Flash



NAND Flash Type	Bits/ Memory Cell	Applications
Single-Level Cell	1	Mission-critical applications (i.e., industrial automation, aerospace & defense, ADAS) Fastest read & write speeds, highest reliability
Multi-Level Cell	2	High-performance SSDs (declining in use) Moderate density, reliability, and cost
Triple-Level Cell	3	Consumer SSDs, mobile, PC Mainstream in regular-grade enterprise market
Quad-Level Cell	4	Data centers, low-write devices, eSSD applications Low cost, read-intensive workloads, lower reliability

Sources: William Blair Equity Research, Samsung

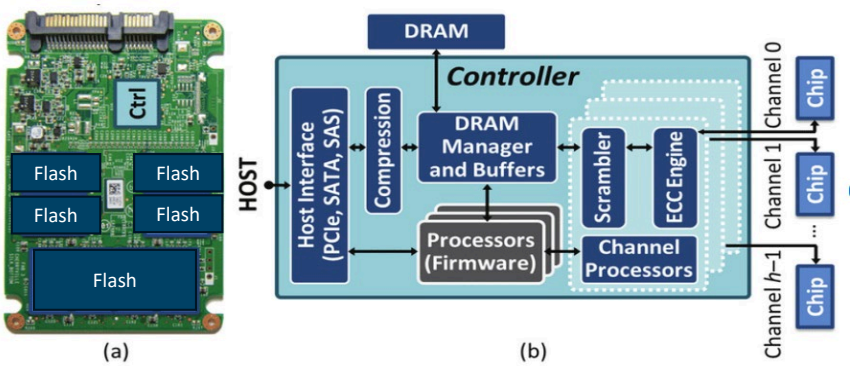
While single-level cells boast the best endurance and write-precision, they are the most expensive, and by nature, the lowest density. On the opposite end of the spectrum, quad-level cells (QLCs) have been gaining momentum in consumer and data-center applications, as they are particularly well suited to read-heavy workloads. Since QLCs store four bits of data per memory cell, they are significantly more cost effective but tend to have a higher rate of errors. This makes them a poor fit for mission-critical operations, but suitable for enterprise operations that can tolerate some error.

SSD controllers gain importance with QLC

As multi-level cell architectures become increasingly commonplace, especially in enterprise SSDs, the importance of NAND flash controllers also increases. The main function of the controller is to manage the interactions between the host system (typically via NVMe running on PCIe) and the raw NAND flash. For instance, the controller manages logical-to-physical mapping, allowing the OS to map bit addresses to the physical NAND pages as well as enabling parallelism between the numerous NAND dies and channels. The controller also ensures bit integrity by managing wear leveling, which involves spreading charges across NAND pages to limit bit degradation, erasing old blocks of data, and correcting errors (through technology called its error correction code).

In many ways the SSD controller is a gatekeeper to flash performance, efficiency, and durability. Phison—a key vendor of SSD controllers—estimates a TAM in excess of $1.4 billion for client SSD controllers in 2025. We project long-term expansion in this TAM given tailwinds such as: 1) the continued transition from legacy storage form factors (i.e., HDDs) to SSDs, 2) growing storage needs driven by robust demand for AI compute, 3) the progression into 300+ layer NAND, and 4) increased enterprise adoption of QLC.

Exhibit 12
Total Recall
What Is an SSD Controller?



a) SSD architecture showing controller (CTRL) and chips; b) detailed view of connections between controller components and chips

Sources: Marvell and William Blair Equity Research

The SSD value chain

The value chain of the storage industry begins with the most fundamental component of the SSD, raw NAND flash. The flash market is highly concentrated, with the four largest flash makers—Samsung, SK Hynix (including Solidigm), Kioxia, and Micron—accounting for more than 80% of global NAND revenues as of the third quarter of 2025, according to data from IDC.

Progressing to the next tier of the value chain, the SSD combines the flash memory with an SSD controller, essentially the brain of the system. SSD controllers are mainly produced by Marvell, Phison, and Silicon Motion. Some NAND flash makers—like Samsung, Micron, and SK Hynix/Solidigm—have in-house controller solutions. Nonetheless, design complexities and rising costs, as memory nodes continue to get smaller and more expensive to manufacture, have led some flash makers to refocus away from in-house controller projects toward merchant providers of turnkey solutions.

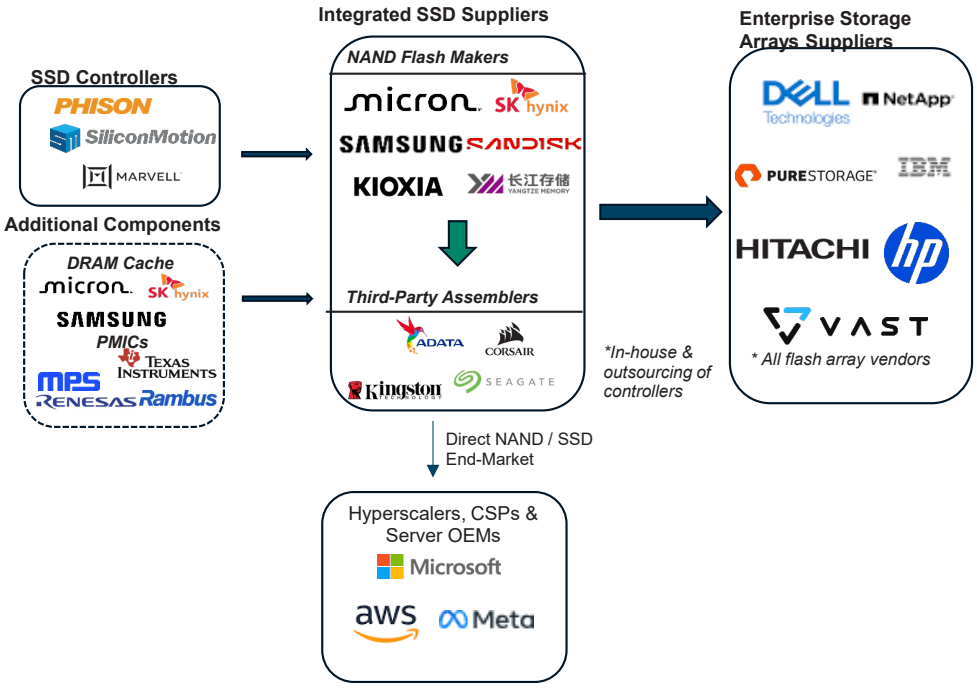
SSD firmware may operate on different interfaces. The interface determines how the SSD connects to the CPU via the storage fabric. Whereas in the early days of SSDs, the SATA interface (evolved from HDDs) was more heavily used to run the AHCI protocol, current deployments of SSDs primarily use PCIe to run the NVMe protocol.

NVMe is flash-optimized, supporting lower latencies and being the preferred interface for high-performance, enterprise, and data center SSDs. While most integrated SSDs are manufactured by vertically integrated flash makers, note that merchant assemblers (Kingston, Corsair, and ADATA) represent a smaller segment of the market, sourcing the controller and flash individually into consolidated SSD solutions.

At the downstream end of the value chain, SSDs are integrated into larger enterprise storage array systems that combine numerous SSDs with controllers, cache, interconnects, and software to deliver a consolidated and scalable storage solution. This end of the market is marked by significant competition, with differentiation increasingly dependent on software.

The major storage array vendors include Dell (leader of the enterprise SSD market), NetApp (dominant in file and all-flash storage), and Pure Storage. Pure Storage is a notable market share gainer in the enterprise storage array space at the expense of more traditional vendors (IBM, Dell, and Hitachi) that have historically been share donors.

Exhibit 13
Total Recall
SSD Value Chain



Source: William Blair Equity Research

Hard disk drives and legacy storage

Hard disk drives (HDDs) have long served as the backbone of digital storage, offering large capacity at relatively low cost per gigabyte. These legacy storage devices rely on spinning magnetic platters and mechanical read/write heads, which inherently limit their performance due to the physical movement.

While HDDs are still suitable for cold storage and archival purposes where speed is less critical, they suffer from high latency, lower IOPS (input/output operations per second), and greater power consumption compared to solid-state alternatives. The falling price of NAND over the last decade has made SSDs increasingly more price competitive with HDDs. This has resulted in a shift away from HDDs across the client and mobile market as well as increasingly within the hyperscaler and enterprise data center.

While SSD shipments are forecast to grow at a significantly higher compound annual rate (about 24% compared to 13% for HDDs, 2024-2029), IDC estimates that HDD storage composed 80% of the installed base of storage capacity in hyperscale data centers as of 2024. HDDs remain the preferred mechanism for low-latency mass data storage primarily due to the cost advantage, with QLC eSSDs garnering a 4-5x higher ASP compared to HDDs. As this price differential narrows, we expect that transitions to SSDs will accelerate.

Exhibit 14
Total Recall
Key Memory Technologies Summarized

Memory Type	Current Generation	Major Vendors	Primary Buyers	Upcoming Development Catalysts
DDR-SDRAM	DDR5-5600/ DDR5-6400 (DIMMs)	SK hynix, SAMSUNG, micron	Cloud providers, OEMs	Scaled DDR5 adoption in AI/ML models DDR6 expected in ~2026-2027
HBM	HBM3E,12-Hi	SK hynix, micron, SAMSUNG	Hyperscalers NVIDIA, AMD	HBM4 commercialization/ wide-scale production in 2026
NAND Flash	3D NAND (5th–7th gen), 176–232+ layers	SAMSUNG, SK hynix, KIOXIA, micron, 长江存储 YANGTZE MEMORY	OEMs, hyperscalers, third-party SSD manufacturers	Commercialization of 232-layer TLC & QLC; 300+ layer NAND expected in 2025-2026

Source: William Blair Equity Research

Redefining Memory: The Emergence of AI

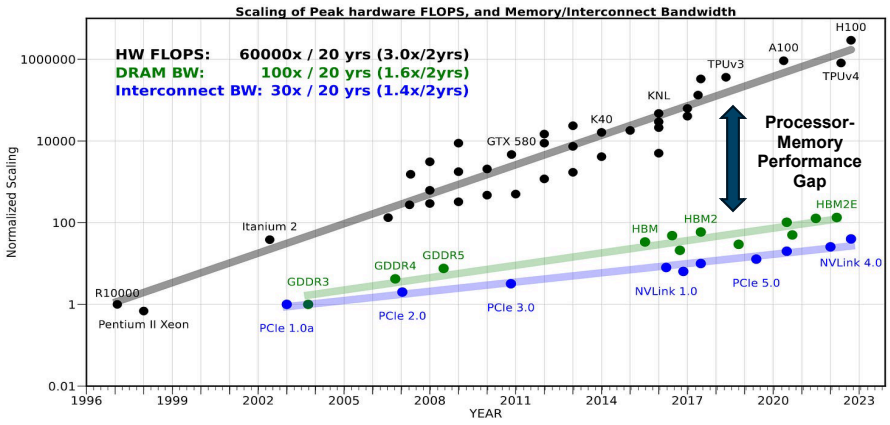
Hitting a Wall: Compute Performance Outpaces Memory Bandwidth
The rapid adoption of AI has reshaped modern computing. However, the disproportionate growth in compute capability relative to memory bandwidth and latency has exacerbated the memory wall, increasingly constraining system-level efficiency.

Scaling memory is not as straightforward as shrinking transistors and enabling parallel processing, as with compute. Memory bandwidth is hindered by physical limits that are harder to simply bypass through the engineering process. Over the past 20 years, processing performance has increased over 60,000 times in accordance with Moore's Law. Contrastingly, DRAM bandwidth has scaled only about 100 times, resulting in a 600:1 compute to memory disparity.

The divergence in processor performance versus memory performance originates from two fundamental areas. First, transferring data between the processor and RAM is slow and substantially more energetically costly than the actual computational FLOPS. This principle is described by the von Neumann bottleneck, an inherent limitation to memory speeds, where the physical separation between the processor and memory drives the latency.

The second contributor to the bottleneck is the sheer volume of data that processors have to consume and retrieve. Reasoning LLMs have exponentially higher memory bandwidth needs as they must not only store immediately relevant tokens, but subsequent tokens generated from interactions with the user. As inference becomes the dominant use-case for AI computing infrastructure, the demands on storage are only expected to increase.

Exhibit 15
Total Recall
Memory Wall: Compute-Memory Gap



Source: 2024, Gholami, AI & Memory Wall (arXiv:2403.14123), edited by William Blair Equity Research

Today, memory providers are able to drive up memory bandwidth using three main tactics:

1. *Pushing per-pin signaling speeds higher.* Increasing the per-pin throughput is primarily achieved by using advanced signaling (PAM4), improving I/O drivers, implementing stringent timing control, and using DDR transfers (i.e., DDR SDRAM).

2. *Adding more channels.* Newer generations add more pin groups by increasing the number of channels. Memory bandwidth gains in DDR5 are largely attributable to the doubled DIMM channels compared to DDR4, enabling a greater degree of parallel data transfer.

3. *Widening the data bus.* The data bus is simply a set of wire interconnects enabling data flow between the processor and memory. Intuitively, by widening the bus, more bits of data can be transferred at the same time—solutions like GDDR5 leverage a 256- to 384-bit bus versus the more traditional 64-bit bus.

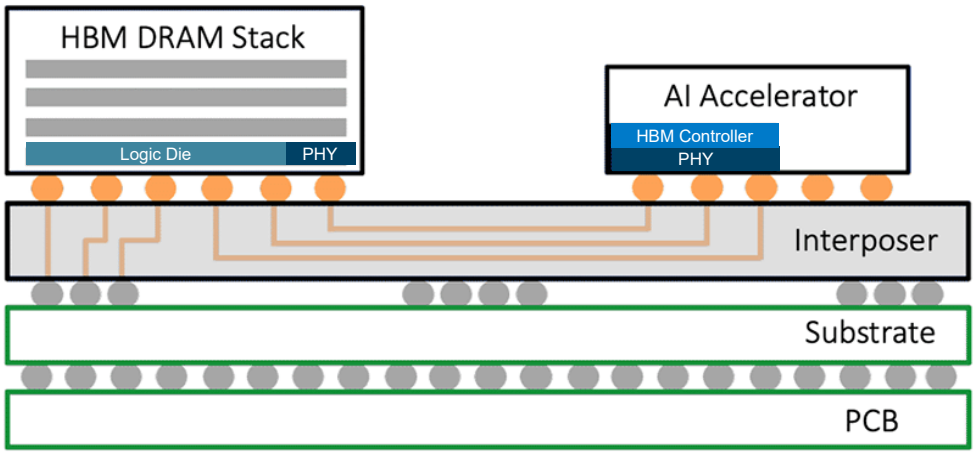
HBM emerges as solution to memory bandwidth limitations

First commercialized by SK Hynix in 2013, HBM has emerged as the premier main memory for AI accelerators. HBM has emerged as a key solution to the memory wall. HBM gets higher bandwidth primarily by massively widening the data bus and increasing parallelism, while keeping pin speeds low and distances extremely short—so utilization and power efficiency stay high. HBM is typically integrated into the GPU or AI accelerator package to enable a wide memory interface, with thousands of data wires running between the GPU and HBM. These short, tightly controlled connections make a 1024-bit (and soon 2048-bit) interface practical.

GPUs are designed to maximize throughput, not to minimize the response time of a single operation. Unlike CPUs, which often wait on the result of a single memory access to continue executing a dependent instruction stream, GPUs run tens of thousands of threads at once. When one group of threads stalls on a memory access, the GPU simply switches to another group that is ready to execute. This architectural choice fundamentally reduces the importance of raw memory latency and elevates the importance of sustained memory bandwidth.

To enable better memory capacity, DRAM dies are stacked together vertically, connected using TSVs. These are small copper wires that feed through the silicon, allowing the DRAM to communicate and function as a fully integrated unit of memory. HBM3E and HBM4 leverage 12 DRAM layers to build an HBM stack.

Exhibit 16
Total Recall
Cross-Sectional View of HBM Integrated With the Processor



Source: William Blair Equity Research and Rambus

This enhanced efficiency and higher bandwidth—with Nvidia calling for 11 GB/s for HBM4 as part of its Rubin architecture—comes at a significant cost premium versus more traditional DRAM. Our research indicates that HBM ASPs are roughly 5 times that of DDR5. This premium is attributable to 1) higher bandwidth; 2) increased silicon wafer requirements per bit, roughly 3-4x times traditional DRAM; and 3) a more-challenging, lower-yield manufacturing process, resulting from increased complexity of aligning multiple DRAM dies into a single form-factor solution.

HBM3E architecture

HBM3E is the current generation of HBM, integrated with the leading AI processors, including Nvidia's B200/B300 and AMD's MI350. The latest generation of custom ASICs are also leveraging HBM3E, including Google's TPU v6 and upcoming v7chips designed with Broadcom, as well as Amazon's Trainium2 UltraServers chip designed with Marvell.

HBM3E achieves over 1.2 TB/s memory bandwidth per stack, a 1.4x improvement from HBM3, as well as a 50% capacity improvement resulting from a shift to 12 DRAM layers versus only 8 prior. While both HBM3 and HBM3E support configurations of 8 or 12 DRAM dies per stack, the improved controller design of the HBM3E logic die enables better input/output (I/O) signaling at lower power and higher bandwidth. In addition, enhanced HBM3E packaging, using newer node technologies and mass reflow-molded underfill (MR-MUF), improves heat dissipation—critical at such high data transfer rates.

SK Hynix was the first large memory vendor to begin volume production of 8-layer HBM3E in March 2024, eventually introducing the first 12-layer stack in September 2024. Micron was the next to market with an 8-high configuration in the first half of 2024, achieving comparable bandwidths to SK Hynix, but boasting a 30% reduction in power.

Samsung's HBM3E solution (Shinebolt) had a considerably more tumultuous path to commercialization. Samsung first released the 8-layer solution in the second half of 2024. Nonetheless, it took almost one more year for Samsung to pass Nvidia's qualification testing for its 12-high HBM3E (which finally passed in September 2025). The delay is speculated to be due to lower performance and yields resulting from Samsung sticking with its 1α fabrication technology, rather than advancing to the 1ß process node (fifth generation, 10nm-class) like its competitors Micron and SK Hynix.

Driven by its first-to-market position and coveted status as the main supplier for Nvidia chips, SK Hynix controls the largest market share in HBM. Micron's 8-high HBM3E is designed into Nvidia's B200 and GB200; the 12-high solution is also integrated into Nvidia's GB300 systems as well as AMD's Instinct MI355X GPU platform and the AMD Instinct MI350 Series solutions. Micron has mentioned two other unidentified major customers, noting traction in ASICs platforms.

Micron expects an inflection from 8 to 12 layers in the first half of 2026, driven by the launch of Nvidia's Rubin GPU; Micron has noted that all of its HBM capacity is sold out for 2026, with the company only able to address half of its current customer demand. SK Hynix leads the HBM market with nearly 70% share in HBM3 and HBM3E, while Micron has gained momentum as the No. 2 player with roughly 21% share. Samsung is catching up with its HBM3E qualification with Nvidia (discussed above) and success with Broadcom ASIC programs (e.g., Meta/Google). Samsung has leveraged price to drive sales of its HBM inventories. We expect Samsung will gain increased traction for HBM4, particularly as yields improve and the company leverages significant TSV capacity. Meanwhile, we expect Micron should maintain its low-20% market share in HBM4, while SK Hynix sees its share decrease toward 60%, as tight capacity limits HBM bit growth.

HBM manufacturing processes

HBM manufacturing is significantly more complex than traditional DRAM fabrication. HBM requires advanced DRAM nodes, complex interposer packaging solutions, and 3D stacking capabilities. HBM fabrication requires an additional 19 engineering steps to the 700 process steps required for standard DRAM, according to Applied Materials, which results in lower yields and drives a cost per bit premium relative to DDR5.

HBM3E uses the same DRAM processing nodes as DDR5. Unlike CPUs and GPUs, the DRAM fabrication process is vertically integrated, with Samsung, Micron, and SK Hynix designing and manufacturing wafers at their own specialized fabs. The current DRAM nodes for HBM3E are 10nm, though the naming conventions differ slightly across vendors. Both Micron and SK Hynix use 1β (fifth generation, 10nm-class) process nodes for HBM3E—both plan to continue on these nodes into HBM4. Samsung has followed a slightly different roadmap, developing HBM3E on the legacy 1α node, though planning to bypass 1β straight to 1γ for HBM4.

Exhibit 17
Total Recall
HBM Roadmaps Across Major DRAM Memory Vendors

Vendor	Generation	DRAM Node	Base Die Node	Main Internal Node Fabs
SK Hynix	HBM3E	1β	1β	M16 (Icheon, South Korea) *Scheduled to begin production of HBM4 in February 2026*
	HBM4	1β	TSMC 12FFC+	
	HBM4E	1γ	TSMC N5/N3*	
Samsung	HBM3E	1a	1a	P3 and P4 (in Pyeongtaek (South Korea)
	HBM4	1c	SF4	
	HBM4E	1c	SF2*	
Micron	HBM3E	1β	1β	MTET and MTAT (Taiwan) and Hiroshima (Japan)
	HBM4	1β	TSMC 12FFC+ and N5	
	HBM4E	1γ	TSMC N5/N3*	

* Unconfirmed

Source: William Blair Equity Research

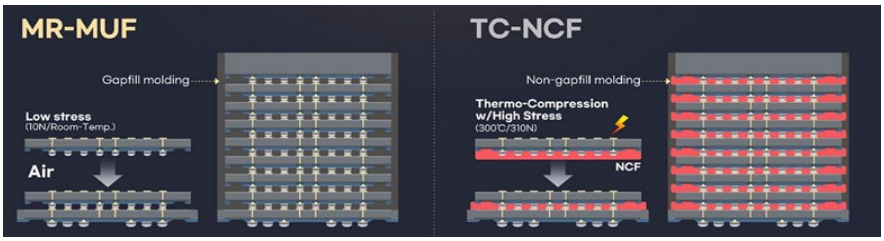
Formation of the through-silicon vias (TSVs) is the most critical, yet complex step of the HBM manufacturing process. The TSV is a vertical electrical interconnect that passes through the silicon die, allowing the DRAM dies to function as a cohesive memory unit. To create these interconnects, 1,024 holes are drilled and copper-filled using chemical mechanical polishing technology before solder micro-bump formation, wafer thinning, chip cutting, and physical stacking of the DRAM dies.

The stacking process only becomes more arduous as DRAM nodes shrink and HBM stacks scale past 12 layers. The TSV formation step is completed internally at Samsung and SK Hynix rather than by OSAT providers.

The approach to stacking and packaging technologies is a key point of differentiation across the major vendors. SK Hynix uses its new MR-MUF technology, while Micron and Samsung rely on legacy non-conductive film thermo compression (NCF TC) techniques. MR-MUF packaging heats and interconnects all DRAM dies in one step, in contrast to individually applying a film to each stacked chip as in NCF TC, resulting in less contamination between dies and ultimately roughly 20% higher yields.

MR-MUF also leverages EMC (epoxy molding compound), a proprietary thermosetting polymer, to increase the number of thermal bumps to achieve superior thermal dissipation. SK Hynix and Samsung complete the final packaging step on the silicon interposer in-house. Micron designs its own packaging but ultimately outsources to TSMC for chip-on-wafer-on-substrate packaging design.

Exhibit 18
Total Recall
HBM Fabrication Process



Source: SK Hynix

HBM4 and beyond

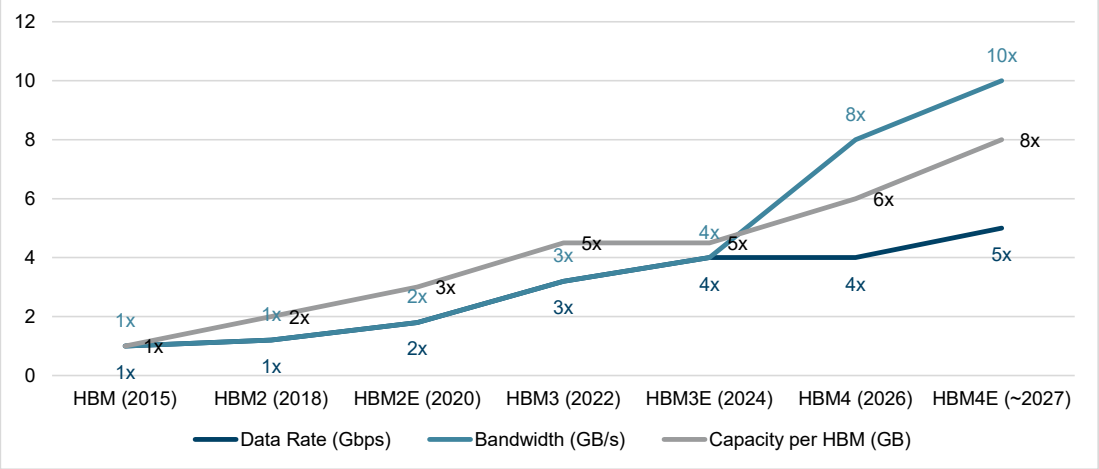
The three major memory vendors are targeting mass production of the fourth generation of HBM in 2026. HBM4 is anticipated to exceed 2 TB/s per stack, a 67% bandwidth improvement relative to HBM3E, largely attributable to a doubling of the I/O channels (to 2,048 bits). Micron has claimed speeds exceeding 2.8 TB/s, as Nvidia asks memory providers to push pin speeds to 11 Gbps.

HBM4 will support stacks of up to 16 DRAM dies with improved capacity per stack up to 64 GB. This major improvement in capacity results in higher power efficiency per bit. In addition, HBM4 technologies will be compatible with HBM3 controllers, facilitating integration into existing architectures. Volume production of HBM4 is in alignment with the anticipated 2026 launch of Nvidia's Rubin GPU as well as AMD's Series MI400 GPUs.

While hyperscalers—Google, Microsoft, and Amazon—are also key customers of HBM, note that ASICs tape-outs typically occur two to three years prior to launch, suggesting that next-generation custom chips (e.g., upcoming Google TPU, next-generation AWS Trainium, and Microsoft Maia) will operate on HBM3E rather than HBM4. In addition, given HBM supply constraints, it is becoming harder for hyperscalers to reserve supply in an Nvidia-dominated buying market.

Exhibit 19
Total Recall
Generation-Over-Generation HBM Improvements

HBM Generation	HBM (2015)	HBM2 (2018)	HBM2E (2020)	HBM3 (2022)	HBM3E (2024)	HBM4 (2026)	HBM4E (2027/2028)
Data Rate	2 Gbps	2.4 Gbps	3.6 Gbps	6.4 Gbps	8 Gbps	8 Gbps	10 Gbps
# of I/O	1024					2048	
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	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



Source: William Blair Equity Research and KAIST

Looking beyond HBM4, we expect HBM4e (the enhanced version of HBM4) to start shipping in 2027 alongside Nvidia's Rubin Ultra GPU systems. Memory makers will be able to customize base dies in HBM4e packages (extended caches, custom protocols, etc.), introducing another novel source of differentiation. The logic base die is the foundational silicon layer that sits at the bottom of the DRAM stack. In HBM3E and predecessor generations, the die served the main purpose of interfacing between the HBM stack and the processor, with limited logic capabilities. In HBM4, the base die will be customized to integrate a memory controller (formerly in the processor), support interfacing to other memory technologies (low-power DDR DRAM, LPDDR), and monitor power and thermal management (see page 30 for more discussion of the base die in HBM4).

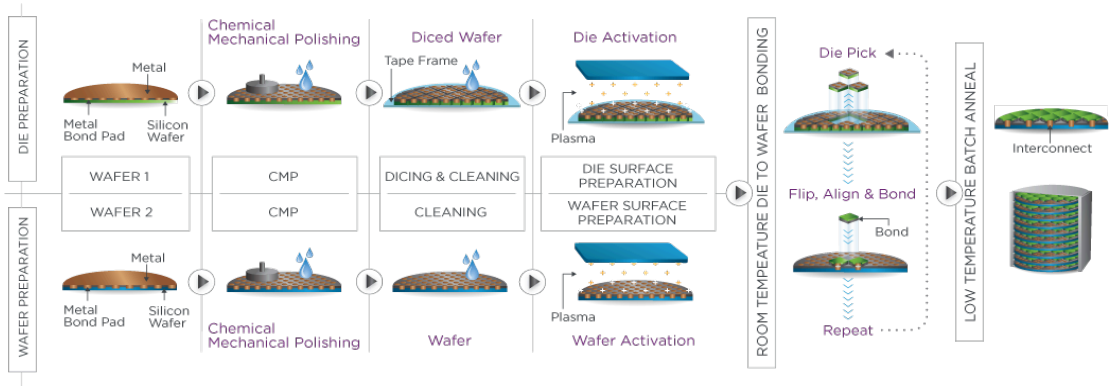
Hybrid bonding

While MR-MUF packaging enabled SK Hynix to gain the lead in HBM3E, the emergence and potential shift to hybrid bonding could be the next evolution in HBM manufacturing. Hybrid bonding enables direct copper-copper bonding between stacked and planarized dies using a low-temperature anneal process, eliminating the need to solder micro-bumps. Micro-bumps connect dies by clicking into divots on top of the die below. This process is highly susceptible to defects and creates spacing (about 40 μm) between the dies.

The implications of hybrid bonding are twofold, enabling more and tighter die connections between layers while reducing the overall stack height of HBM. This results in palpable improvements in thermal resistance and signal integrity compared to the current micro-bump process. Had JEDEC (the standards body for microelectronics) not relaxed the package height requirements for HBM4, hybrid bonding adoption would have been inevitable as stacks advance into 16 layers.

The main headwinds to vendor adoption are the roughly 2x cost premium relative to thermocompression bonding and the higher sanitation requirements, necessitating next-gen HBM packaging facilities. While dominance in hybrid bonding is a strategic focus, mass adoption is not anticipated until the prospective release of HBM5 20-high in 2028-2029, likely used in Nvidia's Feynman generation.

Exhibit 20
Total Recall
Hybrid Bonding Approach

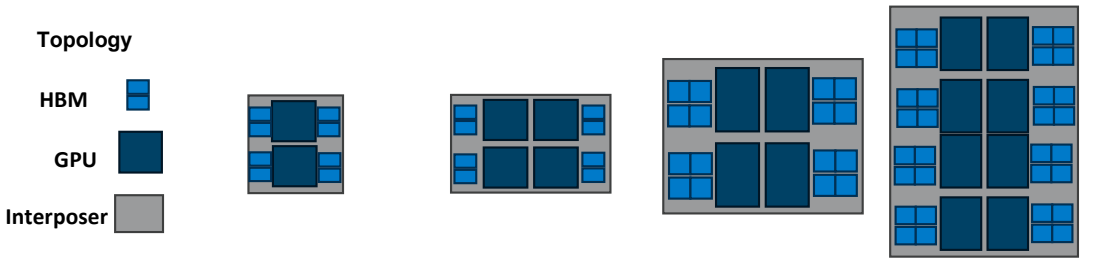


Source: Adeia

Adeia is the owner of foundational hybrid bonding IP (direct-bond interconnect, DBI/DBI Ultra), which was initially acquired by its former parent company, Xperi. Adeia's IP portfolio encompasses the entire bonding process—including the surface prep, cleaning process, and inspection steps to identify bonding layer defects. Adeia has already announced that Micron, YMTC, and Hamamatsu have entered licensing agreements for its hybrid bonding IP.

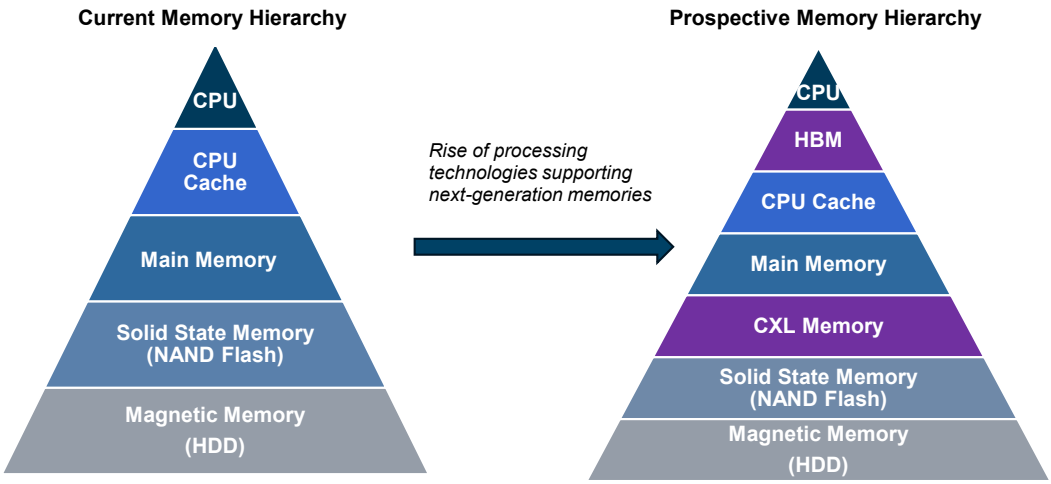
Exhibit 21
Total Recall
HBM Integration Across Nvidia GPU Architectures

GPU Architecture	Rubin (2026)	Feynman (2028)	Post-Feynman (2030)	Next-Gen Architecture (~2032)
GPU Die Size	728 mm²	750 mm²	700 mm²	600 mm²
GPU Power	800 W	900 W	1,000 W	1,200 W
# of GPU Dies	×2	×4	×4	×8
# of HBM Stacks	HBM4 ×8	HBM5 ×8	HBM6 ×16	HBM7 ×32
Interposer Die Size	2,194 mm²(46.2 mm × 48.5 mm)	4,788 mm²(85.2 mm × 56.2 mm)	6,014 mm²(102.8 mm × 58.5 mm)	9,245 mm²(96.4 mm × 95.9 mm)
Total Bandwidth	16 / 32 TB/s	48 TB/s	128 / 256 TB/s	1,024 TB/s
Total HBM Capacity	288 / 384 GB	400 / 500 GB	1,536 / 1,920 GB	5,120 / 6,144 GB
Total Power	2,200 W	4,400 W	5,920 W	15,360 W



Source: William Blair Equity Research and KAIST

Exhibit 22
Total Recall
Revised Memory Hierarchy

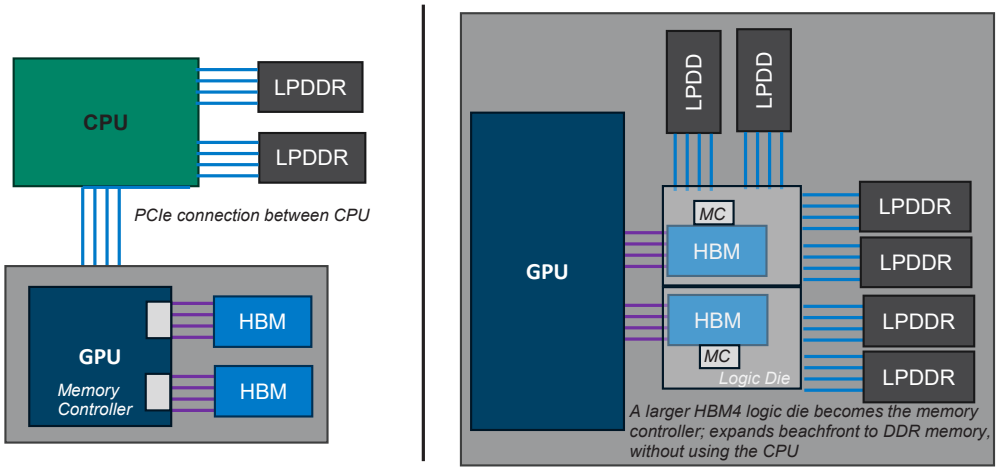


Sources: William Blair Equity Research

Memory Content in AI Racks

While HBM has become a core driver of memory demand and growth, AI is pulling through demand across the memory hierarchy to expand the capacity of AI servers and racks to hold larger models and more context. In this section, we analyze the content of different types of memory in today's leading AI rack solutions.

Exhibit 23
Total Recall
HBM4 Architectural Redesign Expands Logic Die Capabilities and Memory Beachfront



Source: William Blair Equity Research

With HBM3E, AI server architectures leverage a CPU-centric approach. Using Nvidia's architecture as a reference, the memory controller of the CPU is directly connected to LPDDR main memory and to the B200 GPU using NVLink 5 (or PCIe 5.0 if using x86 or customer architectures). B300 has direct interconnects to eight HBM3E 12-high stacks, providing 36 GB of high-bandwidth local memory for a total of 288 GB per chip.

A bottleneck emerges when HBM capacity is maxed out and the GPU must retrieve data from the higher-capacity LPDDR modules attached to the CPU. The bandwidth of PCIe/NVLink is much lower than the HBM I/O channels, resulting in reduced GPU utilization while awaiting the data transfer.

In the HBM4 architecture, with the logic die being redesigned for a 2084-bit bus it becomes significantly larger with added responsibility and complexity. The base die must absorb more routing, power delivery, and signal conditioning. HBM4 also offloads more intelligence and memory controller functionality to the base logic die. Even partial offload—such as advanced error correction code, reliability features, or local scheduling—changes how large GPU-memory complexes can scale.

In addition, HBM4 opens the door to custom HBM, touted by ASIC designers like Marvell and Broadcom. These customer base logic interfaces can be used to manage memory more centrally. One proposed approach is to connect LPDDR directly to HBM base logic die, bypassing the slower GPU-to-CPU connection.

While HBM is indisputably the premier memory technology within the data center, it comes at a significant cost premium. Cutting corners at HBM would compromise the compute integrity of an already expensive infrastructure systems. With GPUs scaling to accommodate more HBM stacks, coupled with expanding die counts per HBM stack (moving into 16-high solutions with HBM4), simple algebra demonstrates the exponential increase in near-term demand.

Exhibit 24
Total Recall
Next-Generation GPUs Sizing Up HBM Demand

GPU Architecture	HBM Generation	Number of HBM Stacks	Total HBM Capacity (GB)	Est. ASP/ stack ($)	Est. Total HBM Dollar Content
H200	HBM3E 8-Hi	6	141	$375	$2,250
B200	HBM3E 8-Hi	8	192	$375	$3,000
B300	HBM3E 12-Hi	8	288	$475	$3,800
Rubin	HBM4 12-Hi	8	288	$600	$4,800
Rubin Ultra	HBM4E 16-Hi	16	1,000 (1 TB)	$750	$12,000

Sources: William Blair Equity Research, Nomad Semi

The solution to the power-hungry and cost drawbacks of HBM has trended toward creating "mixed memory" solutions, or rather, supplementing HBM with alternative types of DRAM. DDR5 is the current AI server main memory standard, though LPDDR is being used increasingly in AI data centers.

Nvidia's GB300 superchip (one Grace CPU paired with two B300 GPUs) presently boasts up to 480 GB of LPDDR5X memory attached to the CPU. Nvidia plans to continue using LPDDR into successive generations, including 218 TB of LPDDR in its roadmap for Rubin Ultra (expected in the second half of 2027), signaling that LPDDR growth should continue to see strong AI-related tailwinds. The main advantages of LPDDR are its low power consumption, cost advantage over traditional DDR5, superior heat dissipation, and the flexibility of its BGA packaging—allowing easy integration into the HBM base die.

Exhibit 25
Total Recall
Different DRAMs for AI Applications

DRAM	Latency	Capacity	Power Efficiency	Cost
DDR-SDRAM				
LPDDR				
GDDR				
HBM				

Sources: William Blair Equity Research, Semiconductor Engineering

Advanced AI data center racks have begun the transition from the current standard DDR5 RDIMM to DDR5 MRDIMM. MRDIMM is essentially an intermediate between DDR5 and DDR6, doubling bandwidth by using two RDIMMs at elevated power and costs.

To keep pace with technological advancements as well as to promote supply chain diversification and competitive pricing, hyperscalers increasingly opt for a multivendor sourcing approach. Recent years have seen an increase in direct procurement from memory vendors as hyperscalers leverage existing DRAM relationships for their HBM needs, a sentiment noted by custom chip makers.

Storage

The current nonvolatile storage standard within the data center is PCIe Gen5 NVMe SSDs (peripheral component interconnect express fifth-generation nonvolatile memory express). These solutions offer approximately 60 TB of memory, with 120 TB and 128 TB models for eSSD applications in development by Silicon Motion and Innodisk. PCIe is a short-range (<0.5 to 1 meter) high-speed interconnect technology capable of connecting the CPU to AI accelerators as well as interfacing with NVMe SSDs.

Nonvolatile memory express (NVMe) refers to the specific transfer protocol for NAND flash over the PCIe connection interface. NVMe is a significant improvement over older protocols (such as SATA and ACHI), enabling lower latency, parallelism of data queues, and overall superior performance. The PCIe Gen5 interconnect, while still in the adoption phase, is expected to scale to over 50% of enterprise SSD shipments by the fourth quarter of 2025. While adoption continues to ramp up, PCIe Gen5 represents significant technological improvements over Gen4, achieving substantially faster read and write speeds, double the bandwidth, and improved power efficiency. IDC expects approximately 21.3 million units are forecast to ship in 2026.

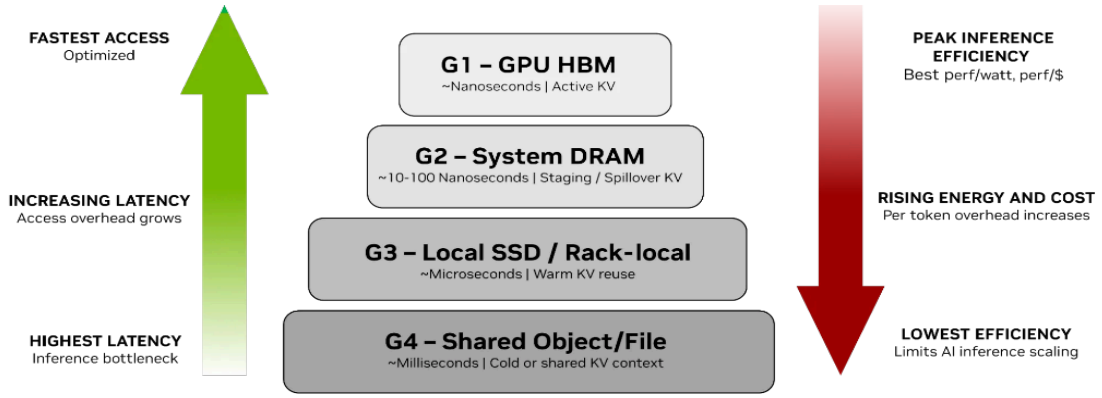
Nvidia's inference context memory storage platform

Nvidia announced a new storage architecture at CES in January 2026 that addresses a growing bottleneck in LLM inference—namely, the explosive growth of key-value (KV) cache as context lengths and agentic workloads expand. Transformer models build up KV cache to avoid recomputing attention history, but this cache quickly outstrips the limited capacity of GPU HBM and system DRAM at large scales, forcing either costly recomputation or inefficient use of slow general-purpose storage.

To solve this, Nvidia has introduced the Inference Context Memory Storage (ICMS) Platform, powered by the BlueField-4 data processor and built on a new context-aware storage tier that extends GPU memory with a cluster-level shared KV cache designed specifically for inference workloads. This platform sits between the fast but small HBM and the slow but large object/file storage tiers, creating a high-bandwidth, purpose-built context memory tier that holds transient KV data and enables GPUs to pre-stage and access it efficiently across nodes in a rack-scale system.

The architecture integrates tightly with Nvidia's broader software, including Dynamo (a disaggregated inference framework), DOCA (the DPU acceleration framework), NIXL libraries, and Spectrum-X Ethernet networking to move KV cache data with high throughput and low latency. By treating KV cache as first-class data and offloading its storage and movement to BlueField-4-powered context memory, this platform claims up to 5 times improvement in tokens-per-second and power efficiency versus traditional storage approaches, while reducing redundant recomputation (prefill) overhead and enabling persistent, reusable context for multi-turn or agentic interactions.

Exhibit 26
Total Recall
Memory Hierarchy Introduced by Nvidia

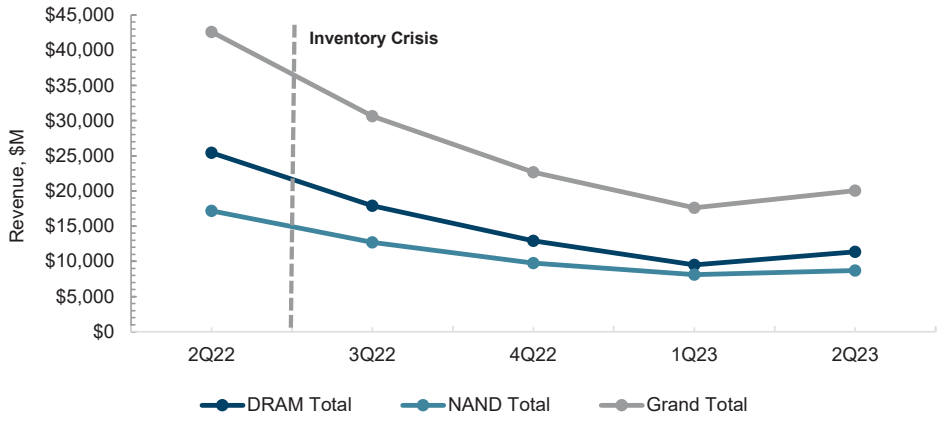


Source: Nvidia

Memory Total Addressable Market

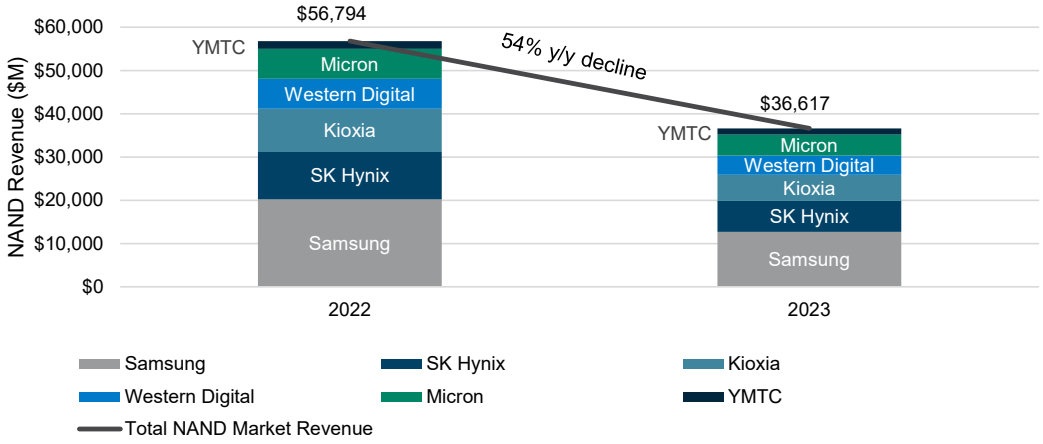
Historically, the semiconductor memory market has been highly cyclical. The most severe correction in the recent history of the sector commenced in the third quarter of 2022, with DRAM and NAND flash ASPs falling in excess of 40% and 50%, respectively. To further contextualize this downturn, SK Hynix, the long-time memory giant, saw a roughly 35% year-over-year dip in NAND revenues in 2023. While primarily ascribed to a large-scale inventory buildup after a period of tight supply coming out of COVID, the downcycle was also driven by weakening consumer electronics demand, macroeconomics headwinds, and supply chain disruptions. In response to this acute market disruption, large memory vendors strategically and rapidly reduced DRAM and NAND production to restrict supply, ultimately stabilizing pricing by the third quarter of 2023.

Exhibit 27
Total Recall
Inventory Crisis: DRAM & NAND Revenues



Sources: William Blair Equity Research; IDC Worldwide Memory Market Shares, 3Q25

Exhibit 28
Total Recall
Inventory Crisis: NAND Revenues Contract Across Major Flash Vendors



Sources: William Blair Equity Research; IDC Worldwide Memory Market Shares, 3Q25

HBM

HBM, the core driver of memory market growth, has a present approximated total addressable market (TAM) of $35 billion, representing 7% of the overall DRAM market, a number rapidly gaining in share. HBM3E demands about three times the amount of silicon per bit relative to traditional DRAMs (referred to as the trade ratio), which in addition to the added complexity of TSV integration, base-die design, and lower yields results in an ASP that is 3-4x that of traditional DRAM solutions.

HBM demand remains strong despite this premium, with major hyperscaler customers contracting more than one year in advance. The largest customer of HBM remains Nvidia. HBM is also pertinent to AMD, Intel, and custom chip makers (e.g., Marvell and Broadcom). Hyperscalers directly contracting HBM from memory vendors has also been a noticeable trend in the custom market.

SK Hynix has been the dominant HBM3E vendor benefitting from its first-to-market positioning with its 12-die, 36 GB HBM3E offering—it expects a doubling of HBM revenue year-over-year in 2025. Micron, joining the HBM market later in the HBM2E era, estimates multibillion dollars of HBM revenue for 2025. Micron projects its peak penetration of the HBM market to be around 23%-24%, the same as its traditional DRAM share.

Samsung, however, has had a less linear trajectory in the HBM market. With original 100% market share in HBM2 and 70% in HBM2E, scaling back investment into HBM in 2019 proved a strategic failure for Samsung. After ceding market share in HBM3E, Samsung is focused on regaining share in HBM4. The transition from 8- to 12-die stacks remains strong, with near-term focus on the launch of HBM. With all three major vendors projecting a large-scale launch in the second half of 2026, the competitive landscape remains an open question, ultimately to be answered by the contracting choices of Nvidia.

DRAM

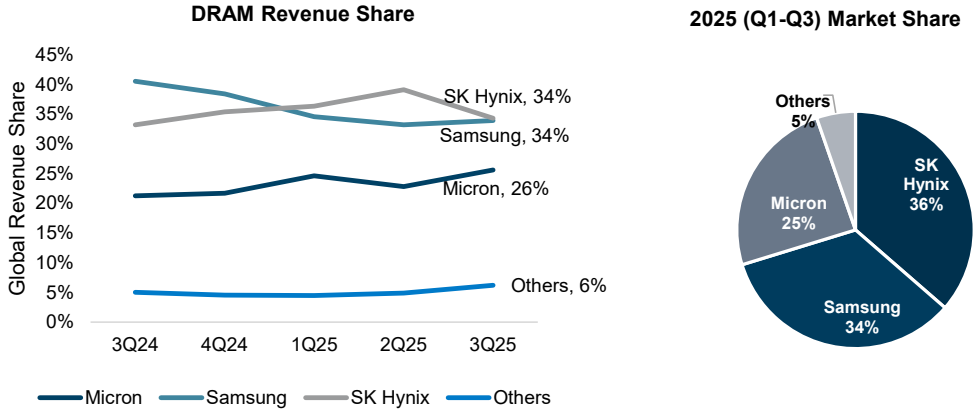
Within the non-HBM DRAM market, DDR5 adoption is accelerating. The shift from DDR4 is driven by 1) the rise of DDR5-compatible CPUs, and 2) discontinuation of DDR3 and DDR4 by major manufacturers as they redirect investments to higher-value products. This strategic transition

away from older generations of DDR is also a response to pricing pressure from Chinese vendors, Changxin Memory Technology (CXMT) and Fujian Jinhua, which have expanded into low-cost DDR4 and LPDDR4.

In DDR5, SK Hynix leads the market with its 16 GB technology. Nonetheless, the rise in AI infrastructure has refocused manufacturers on higher-capacity solutions—as 96 GB DRAM modules become standard among reasoning AI models. The DRAM AI server segment is dominated by Samsung's 128 GB and 256 GB modules. The mobile and PC subsectors are the primary procurers of the non-AI DRAM market. With mobile expected to grow mid- to low single digits in 2025 and increased corporate demand driving the PC market, growth in non-AI applications appears to be normalizing. We do worry about the tightening supply for non-AI memory resulting in demand destruction and slower growth in more traditional parts of the DRAM market, particularly in consumer, mobile, and PC devices.

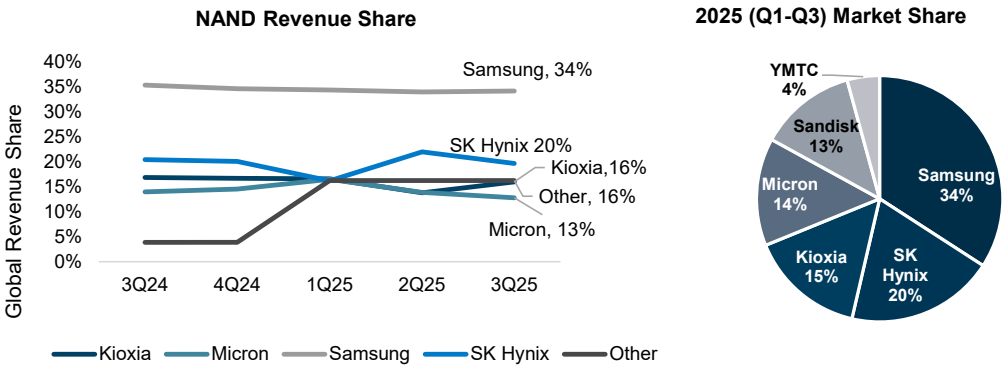
Exhibit 29
Total Recall
DRAM and NAND Market Share

DRAM



**Other primarily includes: Winbond, PowerChip, Nanya, JHICC, CXMT*

NAND



**Other primarily includes: Intel, SanDisk, Toshiba, YMTC*

SK Hynix includes Solidigm;

Sources: William Blair Equity Research; IDC Worldwide NAND Flash Demand and Supply 3Q25, IDC Worldwide DRAM Demand and Supply 3Q25

NAND

While NAND flash usage first emerged in smartphones and mobile devices, as prices of NAND have come down, flash has made strong headway into the enterprise storage market. The enterprise SSD (eSSD) market is poised to considerably outpace the presently dominant smartphone segment.

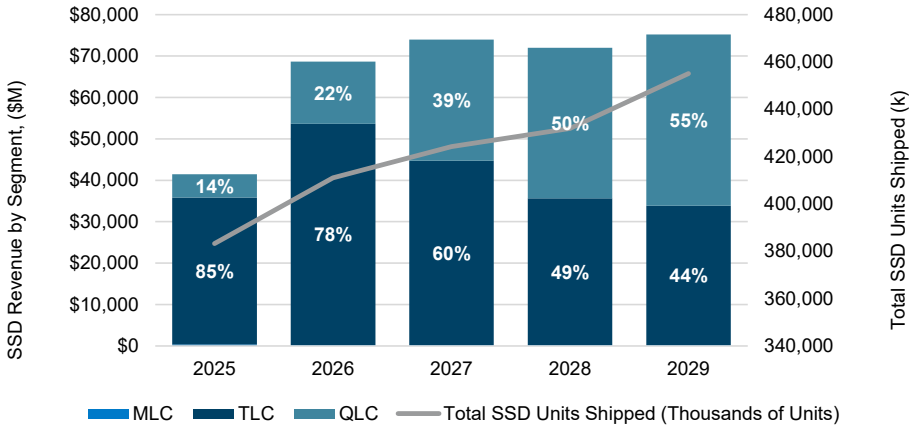
Samsung is the largest producer of NAND flash memory, accounting for 34% share of global volume production in the first three quarters of 2025; Kioxia and SK Hynix follow as the next-largest producers. Market leaders are continuing the transition to eighth generation NAND across applications, with Sandisk and Kioxia expecting a full year 2026 inflection to BiCS8—an eighth-generation, 218-layer, 3D flash technology.

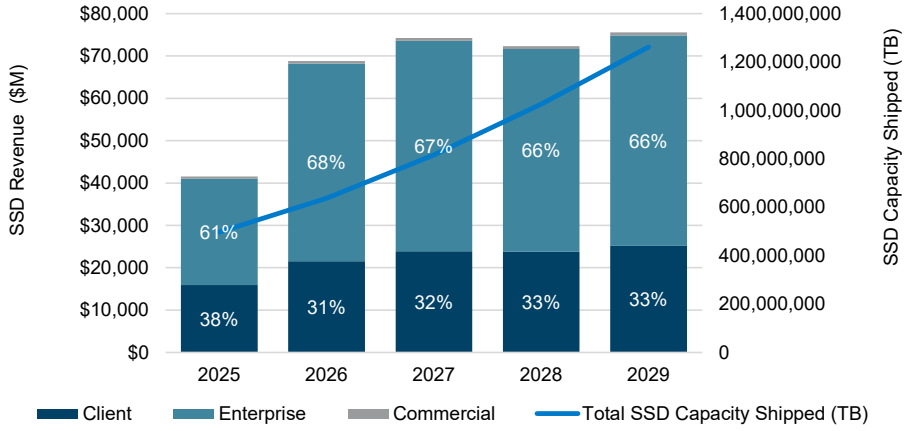
Kioxia and Sandisk have pioneered the transition to 3D NAND, recently unveiling a 10th-generation, 332-layer flash boasting a 33% interface speed improvement compared to BiCS8. The progression of NAND layers continues to be a formidable trend in the flash market, with 3D NAND presenting a cost-effective and capacity optimized solution relative to 2D NAND; industry projections forecast the technology will reach in excess of 400 layers in the next few years.

Strong demand in the enterprise segment is advancing the rise of high-capacity quad-level cell (QLC) eSSD products. QLC is a promising long-term growth driver of broader flash use, given the strong balance between capacity and costs. Samsung began mass production of its 64 TB QLC ninth-generation V-NAND in third quarter 2024, achieving bit density that is 86% higher than the previous generation.

Despite the recent uptick in QLC, TLC (triple-level cell) NAND composed roughly 85% of SSD revenue in 2025. QLC will likely continue to garner incremental share of the SSD market off of a growing base as the premium to legacy HDDs declines, with IDC estimating QLC to rise to approximately 55% of global SSD revenue in 2029.

Exhibit 30
Total Recall
Global SSD Shipments by Technology and End-Market





Sources: William Blair Equity Research; IDCWorldwide Solid State Drive Forecast Pivot Update, 2025–2029

Exhibit 31
Micron Technology, Inc.
William Blair Memory Market Forecast

($ in millions)

	2022A Year	2023A Year	2024A Year	2025E Mar-25A	2025E Jun-25A	2025E Sep-25A	2025E Dec-25E	2025E Year	2026E Mar-26E	2026E Jun-26E	2026E Sep-26E	2026E Dec-26E	2026E Year	2027E Mar-27E	2027E Jun-27E	2027E Sep-27E	2027E Dec-27E	2027E Year	2028E Year	2029E Year	2030E Year
Total DRAM Revenues	**80,772**	**52,574**	**98,488**	**26,340**	**31,025**	**39,717**	**53,896**	**150,978**	**58,881**	**66,065**	**70,062**	**73,565**	**268,572**	**77,243**	**85,925**	**95,463**	**95,224**	**353,854**	**325,118**	**320,346**	**347,381**
% of Total Memory Revenues	*58.3%*	*58.5%*	*59.8%*	*68.0%*	*67.1%*	*69.4%*	*72.7%*	*69.8%*	*74.3%*	*75.4%*	*74.9%*	*74.7%*	*74.8%*	*76.3%*	*78.1%*	*79.9%*	*79.5%*	*78.6%*	*77.0%*	*77.1%*	*77.3%*
Sequential Change %	-	-	-	*(10.0)%*	*17.8%*	*28.0%*	*35.7%*	-	*9.2%*	*12.2%*	*6.1%*	*5.0%*	-	*5.0%*	*11.2%*	*11.1%*	*(0.2)%*	-	-	-	-
Year over Year Change %	-	*(34.9)%*	*87.3%*	*40.9%*	*32.0%*	*47.0%*	*84.1%*	*53.3%*	*123.5%*	*112.9%*	*76.4%*	*36.5%*	*77.9%*	*31.2%*	*30.1%*	*36.3%*	*29.4%*	*31.8%*	*(8.1)%*	*(1.5)%*	*8.4%*
Total Bit Shipments (millions GB)	22,976	27,191	32,430	8,062	9,564	10,873	12,504	41,002	11,878	12,116	12,237	12,237	48,469	12,237	13,216	14,538	15,265	55,256	62,902	69,587	76,411
Sequential Change %	-	-	-	*(3.1)%*	*18.6%*	*13.7%*	*15.0%*	-	*(5.0)%*	*2.0%*	*1.0%*	*0.0%*	-	*0.0%*	*8.0%*	*10.0%*	*5.0%*	-	-	-	-
Year over Year Change %	-	*18.3%*	*19.3%*	*5.2%*	*17.4%*	*31.0%*	*50.3%*	*26.4%*	*47.3%*	*26.7%*	*12.5%*	*(2.1)%*	*18.2%*	*3.0%*	*9.1%*	*18.8%*	*24.7%*	*14.0%*	*13.8%*	*10.6%*	*9.8%*
Average DRAM ASPs ($/GB)	3.52	1.93	3.04	3.27	3.24	3.65	4.31	3.68	4.96	5.45	5.73	6.01	5.54	6.31	6.50	6.57	6.24	6.40	5.17	4.60	4.55
Sequential Change %	-	*(45.0)%*	-	*(7.1)%*	*(0.7)%*	*12.6%*	*18.0%*		*15.0%*	*10.0%*	*5.0%*	*5.0%*		*5.0%*	*3.0%*	*1.0%*	*(5.0)%*		-	-	-
Year over Year Change %	-	*(45.0)%*	*57.1%*	*34.0%*	*12.5%*	*12.2%*	*22.5%*	*21.2%*	*51.7%*	*68.1%*	*56.7%*	*39.5%*	*50.5%*	*3.0%*	*19.2%*	*14.7%*	*3.8%*	*15.6%*	*(19.3)%*	*(10.9)%*	*(1.2)%*
Non-HBM DRAM (DDR, LPDDR. GDDR) Revenue	**78,697**	**49,761**	**80,525**	**19,946**	**23,255**	**30,414**	**44,174**	**117,789**	**49,548**	**56,559**	**58,974**	**60,881**	**225,962**	**63,270**	**71,114**	**78,578**	**75,807**	**288,770**	**246,256**	**236,892**	**262,509**
% of DRAM revenues	*97.4%*	*94.6%*	*81.8%*	*75.7%*	*75.0%*	*76.6%*	*82.0%*	*78.0%*	*84.1%*	*85.6%*	*84.2%*	*82.8%*	*84.1%*	*81.9%*	*82.8%*	*82.3%*	*79.6%*	*81.6%*	*75.7%*	*73.9%*	*75.6%*
Sequential Change %	-	-	-	*(3.9)%*	*16.6%*	*30.8%*	*45.2%*	-	*12.2%*	*14.1%*	*4.3%*	*3.2%*	-	*3.9%*	*12.4%*	*10.5%*	*(3.5)%*	-	-	-	-
Year over Year Change %	-	*(36.8)%*	*61.8%*	*17.5%*	*12.8%*	*37.2%*	*112.8%*	*46.3%*	*148.4%*	*143.2%*	*93.9%*	*37.8%*	*91.8%*	*27.7%*	*25.7%*	*33.2%*	*24.5%*	*27.8%*	*(14.7)%*	*(3.8)%*	*10.8%*
Total Shipments (millions GB), non-HBM	22,747	26,879	30,933	7,558	8,991	10,096	11,649	38,295	11,024	11,219	11,268	11,172	44,683	11,086	11,996	13,074	13,581	49,738	55,416	61,101	67,094
Sequential Change %	-	-	-	*(0.9)%*	*19.0%*	*12.3%*	*15.4%*	-	*(5.4)%*	*1.8%*	*0.4%*	*(0.9)%*	-	*(0.8)%*	*8.2%*	*9.0%*	*3.9%*	-	-	-	-
Year over Year Change %	-	-	*(268.1)%*	*0.6%*	*13.8%*	*27.9%*	*52.8%*	*(25.2)%*	*45.9%*	*24.8%*	*11.6%*	*(4.1)%*	*98.5%*	*0.6%*	*6.9%*	*16.0%*	*21.6%*	*(69.7)%*	*(153.0)%*	*(74.2)%*	*(384.4)%*
Average ASPs ($/GB). Non-HBM DRAM	3.46	1.83	2.46	2.64	2.59	3.01	3.79	3.08	4.49	5.04	5.23	5.45	5.06	5.71	5.93	6.01	5.58	5.81	4.44	3.88	3.91
Sequential Change %	-	-	-	*(3.1)%*	*(2.0)%*	*16.5%*	*25.9%*	-	*18.5%*	*12.2%*	*3.8%*	*4.1%*	-	*4.7%*	*3.9%*	*1.4%*	*(7.1)%*	-	-	-	-
Year over Year Change %	-	*(47.1)%*	*34.2%*	*16.8%*	*(0.9)%*	*7.2%*	*39.3%*	*25.2%*	*70.3%*	*94.9%*	*73.7%*	*43.7%*	*64.4%*	*27.0%*	*17.6%*	*14.8%*	*2.4%*	*14.8%*	*(23.5)%*	*(12.8)%*	*0.9%*
HBM Revenue	**2,075**	**2,813**	**17,962**	**6,395**	**7,770**	**9,303**	**9,722**	**33,190**	**9,333**	**9,506**	**11,087**	**12,684**	**42,609**	**13,972**	**14,811**	**16,884**	**19,417**	**65,085**	**78,862**	**83,454**	**84,872**
% of DRAM revenues	*2.6%*	*5.4%*	*18.2%*	*24.3%*	*25.0%*	*23.4%*	*18.0%*	*22.0%*	*15.9%*	*14.4%*	*15.8%*	*17.2%*	*15.9%*	*18.1%*	*17.2%*	*17.7%*	*20.4%*	*18.4%*	*24.3%*	*26.1%*	*24.4%*
Sequential Change %	-	-	-	*(24.9)%*	*21.5%*	*19.7%*	*4.5%*	-	*(4.0)%*	*1.9%*	*16.6%*	*14.4%*	-	*10.2%*	*6.0%*	*14.0%*	*15.0%*	-	-	-	-
Year over Year Change %	-	*35.6%*	*538.6%*	*272.7%*	*169.5%*	*91.6%*	*14.2%*	*84.8%*	*45.9%*	*22.3%*	*19.2%*	*30.5%*	*28.4%*	*49.7%*	*55.8%*	*52.3%*	*53.1%*	*52.7%*	*21.2%*	*5.8%*	*1.7%*
Total HBM Bit Shipments	**230**	**312**	**1,497**	504	573	777	854	2,707	854	897	969	1,066	3,785	1,151	1,220	1,464	1,683	5,518	7,487	8,485	9,317
% of DRAM bits	*1.0%*	*1.1%*	*4.6%*	*6.2%*	*6.0%*	*7.1%*	*6.8%*	*6.6%*	*7.2%*	*7.4%*	*7.9%*	*8.7%*	*7.8%*	*9.4%*	*9.2%*	*10.1%*	*11.0%*	*10.0%*	*11.9%*	*12.2%*	*12.2%*
Sequential Change %	-	-	-	*(27.3)%*	*13.7%*	*35.6%*	*10.0%*		*0.0%*	*5.0%*	*8.0%*	*10.0%*		*8.0%*	*6.0%*	*20.0%*	*15.0%*	-	-	-	-
Year over Year Change %	-	*35.8%*	*379.6%*	*230.8%*	*133.2%*	*91.2%*	*23.3%*	*80.9%*	*69.6%*	*56.6%*	*24.7%*	*24.7%*	*39.8%*	*34.7%*	*36.0%*	*51.1%*	*58.0%*	*45.8%*	*35.7%*	*13.3%*	*9.8%*
ASPs (GB), HBM	**9.03**	**8.89**	**12.00**	12.70	13.57	11.98	11.38	12.26	10.93	10.60	11.45	11.90	11.26	12.14	12.14	11.53	11.53	11.80	10.53	9.84	9.11
Sequential Change %	-	-	-	*3.4%*	*6.9%*	*(11.7)%*	*(5.0)%*		*(4.0)%*	*(3.0)%*	*8.0%*	*4.0%*	-	*2.0%*	*0.0%*	*(5.0)%*	*0.0%*	-	-	-	-
Year over Year Change %	-	*(1.5)%*	*34.9%*	*12.7%*	*15.6%*	*0.2%*	*(7.3)%*	*2.2%*	*(13.9)%*	*(21.9)%*	*(4.5)%*	*4.6%*	*(8.2)%*	*11.1%*	*14.6%*	*0.8%*	*(3.1)%*	*4.8%*	*(10.7)%*	*(6.6)%*	*(7.4)%*
NAND Revenues	**57,714**	**37,268**	**66,264**	**12,389**	**15,181**	**17,496**	**20,208**	**65,273**	**20,370**	**21,608**	**23,473**	**24,893**	**90,344**	**24,047**	**24,025**	**24,003**	**24,483**	**96,558**	**97,370**	**95,219**	**101,754**
% of Total Memory Revenues	*41.7%*	*41.5%*	*40.2%*	*32.0%*	*32.9%*	*30.6%*	*27.3%*	*30.2%*	*25.7%*	*24.6%*	*25.1%*	*25.3%*	*25.2%*	*23.7%*	*21.9%*	*20.1%*	*20.5%*	*21.4%*	*23.0%*	*22.9%*	*22.7%*
Sequential Change %	-	-	-	*(26.4)%*	*22.5%*	*15.3%*	*15.5%*	-	*0.8%*	*6.1%*	*8.6%*	*6.1%*	-	*(3.4)%*	*(0.1)%*	*(0.1)%*	*2.0%*	-	-	-	-
Year over Year Change %	-	*(35.4)%*	*77.8%*	*(13.4)%*	*(10.5)%*	*(3.6)%*	*20.0%*	*(1.5)%*	*64.4%*	*42.3%*	*34.2%*	*23.2%*	*38.4%*	*18.1%*	*11.2%*	*2.3%*	*(1.6)%*	*6.9%*	*0.8%*	*(2.2)%*	*6.9%*
Total Bit Shipments	644,110	729,699	821,108	180,624	235,329	267,465	280,839	964,256	269,605	274,997	292,872	307,515	1,144,989	322,891	332,578	342,555	349,406	1,347,431	1,569,381	1,808,256	2,047,726
Sequential Change %	-	-	-	*(11.9)%*	*30.3%*	*13.7%*	*5.0%*	-	*(4.0)%*	*2.0%*	*6.5%*	*5.0%*	-	*5.0%*	*3.0%*	*3.0%*	*2.0%*				
Year over Year Change %	-	*13.3%*	*12.5%*	*(12.2)%*	*14.8%*	*30.3%*	*36.9%*	*17.4%*	*49.3%*	*16.9%*	*9.5%*	*9.5%*	*18.7%*	*19.8%*	*20.9%*	*17.0%*	*13.6%*	*17.7%*	*16.5%*	*15.2%*	*13.2%*
Average ASPs ($/GB)	0.09	0.05	0.08	0.07	0.06	0.07	0.07	0.07	0.076	0.079	0.080	0.081	0.08	0.074	0.072	0.070	0.070	0.07	0.06	0.05	0.05
Sequential Change %	-	-	-	*(16.4)%*	*(5.9)%*	*1.4%*	*10.0%*	-	*5.0%*	*4.0%*	*2.0%*	*1.0%*	-	*(8.0)%*	*(3.0)%*	*(3.0)%*	*0.0%*	-	-	-	-
Year over Year Change %	-	*(43.0)%*	*58.0%*	*(1.4)%*	*(22.0)%*	*(26.0)%*	*(12.3)%*	*(16.1)%*	*10.2%*	*21.8%*	*22.5%*	*12.5%*	*16.6%*	*(1.4)%*	*(8.1)%*	*(12.6)%*	*(13.4)%*	*(9.2)%*	*(13.4)%*	*(15.1)%*	*(5.6)%*
Total Revenue, DRAM & NAND	**138,486**	**89,842**	**164,752**	**38,729**	**46,206**	**57,213**	**74,104**	**216,252**	**79,251**	**87,673**	**93,534**	**98,458**	**358,915**	**101,289**	**109,950**	**119,466**	**119,707**	**450,413**	**422,488**	**415,565**	**449,135**
Sequential Change %	-	-	-	*(16.0)%*	*19.3%*	*23.8%*	*29.5%*	-	*6.9%*	*10.6%*	*6.7%*	*5.3%*	-	*2.9%*	*8.6%*	*8.7%*	*0.2%*	-	-	-	-
Year over Year Change %	-	*(35.1)%*	*7.0%*	*17.4%*	*14.2%*	*26.6%*	*60.7%*	*31.3%*	*104.6%*	*89.7%*	*63.5%*	*32.9%*	*66.0%*	*27.8%*	*25.4%*	*27.7%*	*21.6%*	*25.5%*	*(6.2)%*	*(1.6)%*	*8.1%*

Source: Company reports and William Blair Equity Research

Next-Gen Technologies: Further Expanding Access to Memory

While HBM has enabled record bandwidth, it is architecturally incapable of solving the von Neumann bottleneck. HBM enables better processor-memory proximity than traditional DRAM, yet data retrieval is still limited to a finite memory bus. Various other memory technologies are being developed to help further address the memory wall.

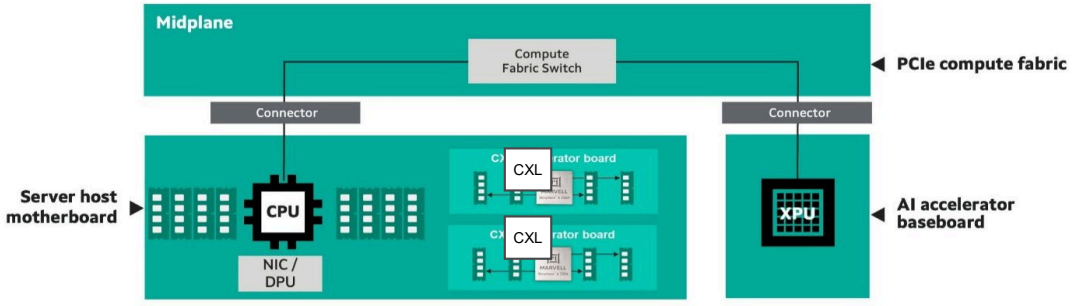
CXL

Compute Express Link (CXL) is an emerging memory interconnect that was developed by Intel, AMD, Nvidia, and Micron, among other large tech players, to support memory pooling and disaggregation in AI/ML workloads. CXL, while in the early innings of development, enables multiple processors and accelerators to access a shared memory pool, reducing data movement and, ideally, total cost of ownership.

CXL builds on its PCIe predecessor (the current standard of scale-up in AI data centers) to streamline data sharing across AI workloads. CXL is directly built on top of the PCIe layer, adding three additional protocol layers (CXL.io, CXL.cache, and CXL.mem) to enable cache-coherency and, in turn, memory pooling. Cache coherency is essential to the CXL mechanism and differentiation from non-coherent PCIe as it unifies memory across the CPU, GPU, and memory hierarchy to reduce latency and errors.

In standard PCIe, software must manually copy data back and forth between caches. In CXL, however, all compute and external memory is shared, so any updates in the memory of one of the systems is automatically visible across all other devices in the rack. The latest CXL specification, CXL 3.2, released in December 2024, is built on the PCIe 6.0 interface; adoption and deployment, however, are based on the preceding CXL 3.1 specification. CXL's market potential is large, with Yole Research estimating a $16 billion market in 2028.

Exhibit 32
Total Recall
CXL Flowchart



Source: Marvell

While CXL technologies presently exist, the market remains immature, and implementation is contingent on the commercialization of compatible CPUs, such as Intel's Granite Rapids and AMD's Turin (processors released in 2025, supporting the CXL 2.0 specification). Astera Labs is also a key player, expanding into CXL with its Leo Memory controller product line, which was first announced in November 2021. Though a lack of compatible CPUs has limited adoption, the more consequential pain point has been CXL's suboptimal bandwidth for AI workloads, even when built onto PCIe 6.0 and PCIe 7.0 interfaces.

Although nascent, the CXL market is fiercely competitive, especially given the lack of incumbency. Other key CXL solutions have been developed by Samsung (512 GB DDR5-based expansion module), Micron (CZ120 CXL memory expansion module), Marvell (Structera CXL memory expansion controllers), Rambus (CXL 3.0 and 2.0 controllers), Montage Technology (CXL expander controller), and SK Hynix (DDR5-based expansion module).

Compute-in-Memory Chips

Compute-in-memory chips (CIM, or processing-in-memory) fundamentally depart from the traditional von Neumann memory architecture, as the memory is designed to process operations. Enabling computations to occur directly on the memory bit lines essentially eliminates the data bus, permitting parallel processing and improved energy efficiency. There are three main emerging categories of CIM, differing in their computational methods: digital CIM (DCIM), analog CIM (ACIM), and hybrid CIM.

Relative to DCIM, ACIM has lower latencies and improved energy efficiencies at the expense of precision and noise susceptibility. Analogous to QLCs, ACIM applications are optimized for high data workloads that can sacrifice some precision for speed, such as edge AI use-cases. ACIMs, a concept that has been in academic development since the 1990s, did not garner commercial interest until the rise of AI inference workloads more recently.

ACIMs remain in the prototype phase, primarily being developed by memory-tech start-ups. DCIMs have outpaced ACIMs in commercialization due to their enhanced scalability and compatibility with existing data infrastructure; large memory vendors such as Samsung and SK Hynix have developed working DCIMs. Industry sentiment is skeptical on the large-scale implementation of CIMs outside academia.

High-Bandwidth Flash (HBF)

High-bandwidth flash (HBF) is an exploratory memory concept that rethinks NAND flash to behave more like a wide-I/O, stacked memory rather than a block device. The core idea is to pair dense flash arrays with a base logic die and very wide parallel interfaces, enabling bandwidth far beyond PCIe/NVMe SSDs. In effect, HBF tries to occupy the gap between DRAM/HBM (fast but expensive and capacity-limited) and traditional flash (cheap and dense but bandwidth- and latency-constrained).

Architecturally, HBF borrows from HBM, in that it contains multiple memory dies stacked vertically, short interconnects, and a logic layer that handles scheduling, error correction code, wear management, and possibly near-memory compute. Access would be burst-oriented and highly parallel, optimized for large sequential transfers rather than fine-grained random reads. The goal is not to match DRAM latency, which is fundamentally unrealistic for flash, but to deliver exponentially higher sustained bandwidth at flash-like cost per bit.

Large model checkpoints, embeddings, feature stores, and cold-but-frequently-streamed data stress SSD bandwidth long before capacity. HBF proposes a tier where data can be streamed into accelerators much faster than NVMe, reducing pressure on expensive HBM capacity while avoiding a full round trip to storage.

HBF technology is still in development. Flash physics impose microsecond-scale access times, and flash management (garbage collection, wear leveling) complicates predictable memory semantics. Power density, thermals in stacked NAND, and software interfaces (neither POSIX files nor simple load/store) are additional hurdles. Research and early disclosures from companies like Samsung and SK Hynix have shown promise for HBF, with early adoption expected in 2028 or 2029 (at the same time as HBM6 deployments).

ReRAM

Another emerging memory solution is resistive RAM (ReRAM or RRAM). ReRAM uses a metal oxide layer between two electrodes, enabling a process called resistive switching. In short, ReRAM technologies register the binary code based on the resistance level inside the metal oxide layer. Since the resistance level is maintained in the absence of power, ReRAM is nonvolatile, presenting a potential low power replacement to existing flash technologies. This mechanism enables expedited read and write speeds, low power consumption, and high bit density. ReRAM technologies mostly remain in the pilot stage, with the main development use-case being focused on embedded nonvolatile memory.

Weebit Nano is a pioneer in ReRAM, leveraging its advanced oxygen vacancy-based ReRAM (OxRAM) technology that is directly integrated with existing CMOS fabrication methods, allowing a cost-effective transition from current manufacturing processes. The company is in the early stages of commercialization of its ReRAM portfolio, partnering with foundries including Onsemi, DB HiTek, and SkyWater Technology. TSMC's shift from MRAM (magneto resistive RAM) to ReRAM in its roadmap could be a helpful proving point on the long-term viability of ReRAM. Challenges with endurance, scalability, and cost competitiveness necessitate significant development improvements before ReRAM can effectively take its place in the next-generation memory hierarchy, competing with incumbent nonvolatile NAND flash.

Neuromorphic

Taking inspiration from the morphology of the human brain, neuromorphic computing seeks to simulate neurological processing. Neuromorphic systems process data using spiking neural networks (SNNs) and artificial ionic channels, mimicking the neurobiological activation and subsequent firing at the human synapse. Neuromorphic technologies seek to emulate brain plasticity to improve computational adaptability and permit parallel processing through the asynchronous structure of the SNNs.

Due to the low power requirements and real-time processing capabilities, neuromorphic chips have potential in edge AI applications, such as autonomous vehicles, robotics, and sensory processing.

Exhibit 33
Total Recall
Private Next-Generation Memory Companies

Company	Next-Generation Memory Technology	Description
EnchargeAI	ACIM	ACIM EN100 AI Accelerator for on-device, edge computing.
Anaflash	ACIM	Nonvolatile edge AI micocontroller leveraging time-based compute-in-memory.
MythicAI	Analog Matrix Processors (AMPS) ACIM	M1076 AMP flagship product, leverages Mythic ACE compute architecture for low-power, edge computing.
AxeleraAI	DCIM	Titania AI DCIM chiplet, Metis AI platform enabled by quad-core AIPU PCIe AI accelerator card.
Intrinsic Semi	ReRAM	Silicon oxide-based ReRAM portfolio.
RainAI	Neuromorphic	IP licensing for DCIM tile and software-stack for on-device workloads; proprietary chip in development.
Innatera	Neuromorphic	Pulsar microcontroller, leverages SNN engine with RISC-V core; first mass-market neuromorphic processor for sensor edge.
MemComputing	CIM	Neuromorphic MemCPU platform, including: MEMAI chips, MEMCPU chips, MEM5G chips. MemAI platform in development.
Crossbar	ReRAM	Patented ReRAM technology for terabyte-scale non-volatile storage and secure embedded memory solutions.
SynSense	Neuromorphic	Dedicated mixed-signal and digital neuromorphic processors for low-power, low-latency sensory inference.
Syntiant	ReRAM	Patented ReRAM technology for terabyte-scale non-volatile storage and secure embedded memory solutions.
Upmem	Processing in Memory	Integrates hundreds of programmable DPUs directly into DRAM to accelerate data-intensive applications like genomics and analytics.

Source: William Blair Equity Research

Conclusion

We remain confident that the AI era is here to stay, and as the industry inflects toward memory-intensive inferencing workloads, the demand for higher-performance memory will only escalate. Demand for HBM has driven significant TAM expansion and better profitability across a largely commoditized memory market, and we expect that growing memory content and GPU shipments will continue to benefit leading manufacturers (SK Hynix, Samsung, and Micron). Increasing technological complexity as AI accelerates the pace of innovation in memory is also opening up a broader opportunity for many other downstream memory vendors (e.g., vendors working on DIMM interface technologies and SSDs). Though the memory market is likely to return to a cyclical cadence at some point, stratospheric AI demand is likely to underpin a memory supercycle over the coming years.

The prices (1/20) of the common stock of other public companies mentioned in this report follow:

Adeia Inc.	$19.15
Alphabet Inc. (Outperform)	$322.16
Amazon.com, Inc. (Outperform)	$231.00
Ambiq Micro, Inc.	$33.02
Advanced Micro Devices, Inc.	$231.92
Applied Materials, Inc. (Market Perform)	$318.23
Astera Labs, Inc. (Outperform)	$183.75
Broadcom Inc. (Outperform)	$332.60
Corsair Gaming, Inc.	$5.43
Credo Technology Group Holding Ltd (Outperform)	$153.22
Dell Technologies Inc.	$111.07
International Business Machines Corporation	$291.35
Intel Corporation	$48.56
Marvell Technology, Inc.	$79.80
Meta Platforms, Inc. (Outperform)	$604.12
Micron Technology, Inc. (Outperform)	$365.00
Microsoft Corporation (Outperform)	$454.52
Monolithic Power Systems, Inc. (Outperform)	$1,034.49
NetApp (Market Perform)	$94.11
Nokia Oyj	$6.41
NVIDIA Corporation (Outperform)	$178.07
ON Semiconductor Corporation (Market Perform)	$60.06
Pure Storage, Inc. (Outperform)	$70.06
Rambus (Outperform)	$110.10
SanDisk Corporation	$453.12
Seagate Technology Holdings plc	$325.99
Silicon Motion Technology Corporation (Outpeform)	$113.00
SkyWater Technology Inc.	$34.78
Texas Instruments Incorporated	$189.59
Taiwan Semiconductor Manufacturing Co., Ltd.	$327.16
Western Digital Corporation	$222.97
Xperi Inc.	$6.10

IMPORTANT DISCLOSURES

William Blair or an affiliate is a market maker in the security of Arm Holdings plc, Astera Labs, Inc., Broadcom Inc., Credo Technology Group Holding Ltd, Marvell Technology, Inc., Micron Technology, Inc., Monolithic Power Systems, Inc., NVIDIA Corporation, Rambus Inc., Silicon Motion Technology Corporation, Arista Networks, Inc., Cisco Systems, Inc., F5, Inc. and Oracle Corporation.

William Blair or an affiliate expects to receive or intends to seek compensation for investment banking services from Arm Holdings plc, Astera Labs, Inc., Broadcom Inc., Credo Technology Group Holding Ltd, Marvell Technology, Inc., Micron Technology, Inc., Monolithic Power Systems, Inc., NVIDIA Corporation, Rambus Inc., Silicon Motion Technology Corporation, Arista Networks, Inc., Cisco Systems, Inc., F5, Inc. and Oracle Corporation or an affiliate within the next three months.

Officers and employees of William Blair or its affiliates (other than research analysts) may have a financial interest in the securities of Arm Holdings plc, Astera Labs, Inc., Broadcom Inc., Credo Technology Group Holding Ltd, Marvell Technology, Inc., Micron Technology, Inc., Monolithic Power Systems, Inc., NVIDIA Corporation, Rambus Inc., Silicon Motion Technology Corporation, Arista Networks, Inc., Cisco Systems, Inc., F5, Inc. and Oracle Corporation.

This report is available in electronic form to registered users via R*Docs™ at https://williamblairlibrary.bluematrix.com or www.williamblair.com.

Please contact us at +1 312 236 1600 or consult https://www.williamblair.com/equity-research/coverage for all disclosures.

Sebastien Naji attests that 1) all of the views expressed in this research report accurately reflect his/her personal views about any and all of the securities and companies covered by this report, and 2) no part of his/her compensation was, is, or will be related, directly or indirectly, to the specific recommendations or views expressed by him/her in this report. We seek to update our research as appropriate. Other than certain periodical industry reports, the majority of reports are published at irregular intervals as deemed appropriate by the research analyst.

DOW JONES: 48488.60
S&P 500: 6796.86
NASDAQ: 22954.30

Additional information is available upon request.

Current Rating Distribution (as of January 22, 2026):

Coverage Universe	Percent	Inv. Banking Relationships *	Percent
Outperform (Buy)	72	Outperform (Buy)	11
Market Perform (Hold)	28	Market Perform (Hold)	3
Underperform (Sell)	1	Underperform (Sell)	0

*Percentage of companies in each rating category that are investment banking clients, defined as companies for which William Blair has received compensation for investment banking services within the past 12 months.

The compensation of the research analyst is based on a variety of factors, including performance of his or her stock recommendations; contributions to all of the firm's departments, including asset management, corporate finance, institutional sales, and retail brokerage; firm profitability; and competitive factors.

OTHER IMPORTANT DISCLOSURES

Stock ratings and valuation methodologies: William Blair & Company, L.L.C. uses a three-point system to rate stocks. Individual ratings reflect the expected performance of the stock relative to the broader market (generally the S&P 500, unless otherwise indicated) over the next 12 months. The assessment of expected performance is a function of near-, intermediate-, and long-term company fundamentals, industry outlook, confidence in earnings estimates, valuation (and our valuation methodology), and other factors. Outperform (O) - stock expected to outperform the broader market over the next 12 months; Market Perform (M) - stock expected to perform approximately in line with the broader market over the next 12 months; Underperform (U) - stock expected to underperform the broader market over the next 12 months; not rated (NR) - the stock is not currently rated. The valuation methodologies include (but are not limited to) price-to-earnings multiple (P/E), relative P/E (compared with the relevant market), P/E-to-growth-rate (PEG) ratio, market capitalization/revenue multiple, enterprise value/EBITDA ratio, discounted cash flow, and others. Stock ratings and valuation methodologies should not be used or relied upon as investment advice. Past performance is not necessarily a guide to future performance.

The ratings and valuation methodologies reflect the opinion of the individual analyst and are subject to change at any time.

Our salespeople, traders, and other professionals may provide oral or written market commentary, short-term trade ideas, or trading strategies-to our clients, prospective clients, and our trading desks-that are contrary to opinions expressed in this research report. Certain outstanding research reports may contain discussions or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Investing in securities involves risks. This report does not contain all the material information necessary for an investment decision. Always refer to the most recent report on a company or issuer. Our asset management and trading desks may make investment decisions that are inconsistent with recommendations or views expressed in this report. We will from time to time have long or short positions in, act as principal in, and buy or sell the securities referred to in this report. Our research is disseminated primarily electronically, and in some instances in printed form. Research is simultaneously available to all clients. This research report is for our clients only. No part of this material may be copied or duplicated in any form by any means or redistributed without the prior written consent of William Blair & Company, L.L.C.

This is not in any sense an offer or solicitation for the purchase or sale of a security or financial instrument.

The factual statements herein have been taken from sources we believe to be reliable, but such statements are made without any representation as to accuracy or completeness or otherwise, except with respect to any disclosures relative to William Blair or its research analysts. Opinions expressed are our own unless otherwise stated and are subject to change without notice. Prices shown are approximate.

This report or any portion hereof may not be copied, reprinted, sold, or redistributed or disclosed by the recipient to any third party, by content scraping or extraction, automated processing, or any other form or means, without the prior written consent of William Blair. Any unauthorized use is prohibited.

If the recipient received this research report pursuant to terms of service for, or a contract with William Blair for, the provision of research services for a separate fee, and in connection with the delivery of such research services we may be deemed to be acting as an investment adviser, then such investment adviser status relates, if at all, only to the recipient with whom we have contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing). If such recipient uses these research services in connection with the sale or purchase of a security referred to herein, William Blair may act as principal for our own account or as riskless principal or agent for another party. William Blair is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities referred to herein.

For important disclosures, please visit our website at williamblair.com.

This material is distributed in the United Kingdom and the European Economic Area (EEA) by William Blair International, Ltd., authorised and regulated by the Financial Conduct Authority (FCA). William Blair International, Limited is a limited liability company registered in England and Wales with company number 03619027. This material is only directed and issued to persons regarded as Professional investors or equivalent in their home jurisdiction, or persons falling within articles 19 (5), 38, 47, and 49 of the Financial Services and Markets Act of 2000 (Financial Promotion) Order 2005 (all such persons being referred to as "relevant persons"). This document must not be acted on or relied on by persons who are not "relevant persons."

This report is being furnished in Brazil on a confidential basis and is addressed to the addressee personally, and for its sole benefit. This does not constitute an offer or solicitation for the purchase or sale of a security by any means that would constitute a public offering in Brazil under the regulations of the Brazilian Securities and Exchange Commission (*Comissão de Valores Mobiliários*) or an unauthorized distribution under Brazilian laws and regulations. The securities are authorized for trading on non-Brazilian securities markets, and this report and all the information herein is intended solely for professional investors (as defined by the applicable Brazilian regulation) who may only acquire these securities through a non-Brazilian account, with settlement outside Brazil in a non-Brazilian currency.

Any statements in this report that are attributable to IDC Research, Inc. ("IDC") represent William Blair's interpretation of data, research opinion or viewpoints published as part of a syndicated subscription service by IDC and have not been reviewed by IDC. IDC's research is current as of the date IDC published it, not the date that William Blair's reports are published. Further, IDC's research contains IDC's opinion, not representations of fact, and are subject to change without notice.

William Blair & Company, L.L.C. licenses and applies the SASB Materiality Map® and SICSTM in our work.

William Blair

Equity Research Directory

John Kreger, Partner Director of Research +1 312 364 8612
Kyle Harris, CFA, Partner Operations Manager +1 312 364 8230

Scott Hansen, Partner Associate Director of Research +1 212 245 6526

CONSUMER

Sharon Zackfia, CFA, Partner +1 312 364 5386
Group Head–Consumer
Lifestyle and Leisure Brands, Restaurants, Automotive/E-commerce

Jon Andersen, CFA, Partner +1 312 364 8697
Consumer Products

Phillip Blee, CPA +1 312 801 7874
Home and Outdoor, Automotive Parts and Services, Discount and Convenience

Dylan Carden +1 312 801 7857
E-commerce, Specialty Retail

ECONOMICS

Richard de Chazal, CFA +44 20 7868 4489

ENERGY AND POWER TECHNOLOGIES

Jed Dorsheimer +1 617 235 7555
Group Head–Energy and Power Technologies
Generation, Efficiency, Storage

Neal Dingmann +1 312 801 7835
Oil and Gas, Rare Earth

Tim Mulrooney, Partner +1 312 364 8123
Energy and Environmental Services

FINANCIAL SERVICES AND TECHNOLOGY

Adam Klauber, CFA, Partner +1 312 364 8232
Group Head–Financial Services and Technology
Financial Analytic Service Providers, Insurance Brokers, Property & Casualty Insurance

Andrew W. Jeffrey, CFA +1 415 796 6896
Fintech

Cristopher Kennedy, CFA +1 312 364 8596
Fintech, Specialty Finance

Jeff Schmitt +1 312 364 8106
Wealthtech, Wealth Management, Capital Markets Technology

GLOBAL SERVICES

Tim Mulrooney, Partner +1 312 364 8123
Group Head–Global Services
Commercial and Residential Services

Andrew Nicholas, CPA +1 312 364 8689
Consulting, HR Technology, Information Services

Trevor Romeo, CFA +1 312 801 7854
Staffing, Waste and Recycling

HEALTHCARE

Biotechnology

Matt Phipps, Ph.D., Partner +1 312 364 8602
Group Head–Biotechnology

Sami Corwin, Ph.D. +1 312 801 7783

Lachlan Hanbury-Brown +1 312 364 8125

Andy T. Hsieh, Ph.D., Partner +1 312 364 5051

Myles R. Minter, Ph.D., Partner +1 617 235 7534

Healthcare Technology and Services

Ryan S. Daniels, CFA, Partner +1 312 364 8418
Group Head–Healthcare Technology and Services
Healthcare Technology, Healthcare Services

Brandon Vazquez, CFA +1 212 237 2776
Dental, Animal Health, Medical Technology

Life Sciences

Matt Larew, Partner +1 312 801 7795
Life Science Tools, Bioprocessing, Healthcare Delivery

Andrew F. Brackmann, CFA +1 312 364 8776
Diagnostics

Max Smock, CFA +1 312 364 8336
Pharmaceutical Outsourcing and Services

INDUSTRIALS

Brian Drab, CFA, Partner +1 312 364 8280
Co-Group Head–Industrials
Advanced Manufacturing, Industrial Technology

Ryan Merkel, CFA , Partner +1 312 364 8603
Co-Group Head–Industrials
Building Products, Specialty Distribution

Louie DiPalma, CFA +1 312 364 5437
Aerospace and Defense, Smart Infrastructure

Ross Sparenblek +1 312 364 8361
Diversified Industrials, Robotics, and Automation

TECHNOLOGY, MEDIA, AND COMMUNICATIONS

Jason Ader, CFA, Partner +1 617 235 7519
Co-Group Head–Technology, Media, and Communications
Infrastructure Software

Arjun Bhatia, Partner +1 312 364 5696
Co-Group Head–Technology, Media, and Communications
Software

Dylan Becker, CFA +1 312 364 8938
Software

Louie DiPalma, CFA +1 312 364 5437
Government Technology

Jonathan Ho, Partner +1 312 364 8276
Cybersecurity, Security Technology

Sebastien Naji +1 212 245 6508
Infrastructure Software, Semiconductor and Infrastructure Systems

Maggie Nolan, CPA, Partner +1 312 364 5090
IT Services

Jake Roberge +1 312 364 8056
Software

Ralph Schackart III, CFA, Partner +1 312 364 8753
Internet and Digital Media

Stephen Sheldon, CFA, CPA, Partner +1 312 364 5167
Vertical Technology – Real Estate, Education, Restaurant/Hospitality

EDITORIAL AND SUPERVISORY ANALYSTS

Steve Goldsmith, Head Editor and SA +1 312 364 8540
Katie Anderson, Editor and SA +44 20 7868 4451
Audrey Majors, Editor and SA +1 312 364 8992
Beth Pekol Porto, Editor and SA +1 312 364 8924
Lisa Zurcher, Editor and SA +44 20 7868 4549