on 31-Aug-2026

BERNSTEIN
SOCIETE GENERALE GROUP

Global Memory

Global Memory: An AI memory primer



Mark Li
+852 2123 2645
mark.li@bernsteinsg.com



Mark C. Newman
+1 212 845 7822
mark.newman@bernsteinsg.com



Edward Hou, CFA
+852 2123 2623
edward.hou@bernsteinsg.com



Phoebe Sun
+1 917 344 8481
phoebe.sun@bernsteinsg.com

*AI presents a great challenge to memory, and it seems that every day there are new innovations attempting to surmount the challenge. HBF, Z-NAND, CXL, CMX, ZAM, HBC, etc. - supply chain are coming up with new ideas & acronyms at an unprecedented pace. It has been a struggle for investors to understand them, and let alone put them into an overarching framework & then to assess their relative positions, potential and investment implications. This primer aims to address that & includes many photos & diagrams to help investors visualize these technologies. Our **prior primer** covers memory technologies in general and in this one we will focus on memories used in AI. We will zoom into LLMs based on the transformer architecture, and will start with the needs for different memories at the different parts of the workloads. We then summarize all these innovations into two hierarchies, from the perspective of architecture & hardware, to establish the overall framework.*

Different AI workloads need different memory. Training is the most demanding task, and puts a difficult constraint on all tiers of memory. Inference can be divided into prefill and decode, with prefill "compute-bound" and XPUs & HBM inside XPUs more important. Decode is on the other hand "memory-bound", esp. with KV cache size compounding quickly with both token count & concurrent user count. Expanding HBM capacity at an affordable cost, offloading to conventional DRAM & NAND, and/or creating additional memory tiers among them are becoming a key focus for innovations. Creating an RAG database requires large-capacity SSD/HDD & DRAM. Searching the database relies on system DRAM more. Finally, agentic workflow first triggers many non-AI tasks for CPU & system memory. The results of agentic works then need to be understood by AI & that in turn calls for more compute & memory needs.

Innovations in the "architectural" hierarchy. We define this hierarchy with first memory (usually SRAM) inside XPU on the top, followed by system memory, local SSD & finally by shared storage (usually SSD/HDD/tape) at the bottom. HBM adds a new tier right beneath the memory inside XPU. CXL & "Storage Next" try to insert tiers between system memory & local SSD. And CMX is about adding a tier between local SSD & shared storage.

Innovations in the "hardware" hierarchy. LPU resorts to SRAM to minimize latency. zHBM promises to outperform HBM by putting HBM on the top of XPU. NVHBM achieves the same goal but compromises memory suppliers' position with a base die designed by NVIDIA. HBF is designed to complement HBM with the significantly higher capacity & lower cost per bit of NAND, thus increasing overall bandwidth to XPU for AI inference. HBC from Qualcomm has the same objective, but with LPDDR. XBM & ZAM from Intel are very different, though also aim to complement/replace HBM. MRDIMM, SOCAMM2 & LPCAMM2 innovate on DRAM modules for high-performance needs. China is accelerating $4F^2$ & 3D DRAM R&D for DRAM in general. And the industry has been trying to create SCM between DRAM & NAND for a while, with for example Z-NAND & XL-FLASL based on "enhanced" NAND, & MRAM, PCM (including XPoint), RRAM based on mechanisms fundamentally different from DRAM or NAND. PIM mixes logic and memory in a piece of silicon, & is both an architectural and hardware change. Adoption of PIM has been very limited as it differs from the current computer systems & supply chain significantly. zNAND-O targets on-device AI by stacking NAND dies with TSV.

See the Disclosure Appendix of this report for required disclosures, analyst certifications and other important information. Alternatively, visit our Global Research Disclosure Website.

First Published: 28 Aug 2026 14:43 UTC Completion Date: 28 Aug 2026 14:43 UTC

on 31-Aug-2026

BERNSTEIN TICKER TABLE

Ticker	Rating	Cur	28 Aug 2026 Closing Price	Price Target	TTM Rel. Perf.	Cur	Reported EPS 2025A	Reported EPS 2026E	Reported EPS 2027E	Reported P/E (x) 2025A	Reported P/E (x) 2026E	Reported P/E (x) 2027E
005930.KS (SEC- Samsung)	O	KRW	256,500	440,000	236.5%	KRW	6,611.53	48,393	77,273	38.8	5.3	3.3
005935.KS (SEC-Pref - Samsung)	O	KRW	186,800	374,000	197.4%	KRW	6,611.53	48,393	77,273	28.3	3.9	2.4
SMSN.LI (Samsung)	O	USD	4,702.00	7,350.00	241.4%	USD	116.15	812.39	1,290.80	40.5	5.8	3.6
000660.KS (SK hynix)	O	KRW	1,658,000	3,300,000	485.5%	KRW	60,341	395,677	568,862	27.5	4.2	2.9
MU (Micron)	O	USD	935.39	1,300.00	647.4%	USD	8.29	67.39	158.99	112.9	13.9	5.9
285A.JP (KIOXIA)	U	JPY	47,900	40,000	1806.5%	JPY	1,014.00	10,013	9,656.84	47.2	4.8	5.0
SNDK (SanDisk)	O	USD	1,484.95	3,000.00	2799.8%	USD	70.08	238.36	272.09	21.2	6.2	5.5
STX (Seagate)	O	USD	847.20	1,350.00	372.2%	USD	15.53	40.61	64.40	54.6	20.9	13.2
WDC (Western Digital)	O	USD	462.00	770.00	443.9%	USD	10.19	23.39	36.58	45.3	19.8	12.6
ASIAX			1,975.06									
EM			1,837.35									
SPX			7,730.99									

O - Outperform, M - Market-Perform, U - Underperform, NR - Not Rated, CS - Coverage Suspended

MU, SNDK, STX, WDC estimate is Adjusted EPS; MU, SNDK, STX, WDC valuation is Adjusted P/E (x); SNDK, STX, WDC base year is 2026;

Source: Bloomberg, Bernstein estimates and analysis.

INVESTMENT IMPLICATIONS

Samsung Electronics: We rate Samsung Electronics Outperform with price target of KRW 440,000.

SK hynix : We rate SK hynix Outperform with price target of KRW 3,300,000.

Micron: We rate Micron Outperform with price target of US$1,300.00.

KIOXIA: We rate KIOXIA Underperform with price target of JPY 40,000.

Sandisk: We rate SNDK Outperform with a price target of $3,000.00

Seagate: We rate STX Outperform with a price target of $1350.00

Western Digital: We rate WDC Outperform with a price target of $770.

on 31-Aug-2026

Table Of Contents

Different workloads & hence different memory requirements in AI 3
Training 3
Inference - Prefill 4
Inference - Decode 4
RAG (retrieval-augmented generation) 4
Agentic workflow 5
"Architectural" hierarchy & emerging tiers in AI 5
G0 - Memory inside processor 6
G1 - HBM (high bandwidth memory) 6
G2 - System memory 6
G2.5 - CXL (compute express link) memory 6
G2.6 - "Storage Next" memory 6
G3 - Local SSD 6
G3.5 - CMX (context storage memory) memory & STX (storage technology extension) 6
G4 - Shared storage 7
"Hardware" hierarchy & innovations for AI 13
SRAM (static random access memory) 13
DRAM (dynamic random access memory) 14
SCM (storage class memory) 16
NAND flash 17
HDD (hard disk drive) 18
TAPE 18

DETAILS

DIFFERENT WORKLOADS & HENCE DIFFERENT MEMORY REQUIREMENTS IN AI

Before we look at different memories, we first summarize how AI workloads differ at different phases and the characteristics of the corresponding memory usage. We focus on LLM (large language model) based on the transformer architecture as its dominance in AI deployments now and divide the workloads into **training**, **inference-prefill**, **inference-decode**, **RAG** (retrieval - augmented generation) & **agentic workflow**.

TRAINING

- This phase is very compute intensive, and memory bandwidth is often of particular importance too, as bandwidth often dictates the absolute upper limit of model size & compute speed. HBM hence plays a critical role in the training phase. In addition, massive training dataset has to be prepared and stored in advance, likely in some slower but cheaper memory. Along the training process, the data needs to be cached and staged for pipelines in some faster memory before it is copied into HBM. Moreover, checkpoints need to be inserted to take a snapshot of the intermediate result, as training can take months, and can't afford restart from the beginning in case that any hardware/software failures take place. This means in

on 31-Aug-2026

addition to HBM, system DRAM, local SSD, and networked storage, basically memory of all other tiers, are also essential to carry out training large models successfully in practice.

INFERENCE - PREFILL

- In a transformer model, inference is executed over two stages - “prefill” & “decode”. The first stage “prefill” processes prompt and generates the first token. Prefill is usually compute-bound, as it involves large matrix-matrix operations that will keep GPU highly utilized. Memory relatively will be less an issue, as the heavy operations inside GPUs take more time than data transfer. HBM is thus relatively more important type of memory in this stage. Once “prefill” completes, the first token will be then cached in pairs of key and value (hence KV cache) as the input for the next stage - “decode”.

- As prefill is "compute-bound", the key performance index of this stage is usually "time to first token" (TTFT) (i.e. latency). Some studies find TTFT of 0.5 second or lower ideal &1 second acceptable for chatbots. TTFT exceeding 2 seconds is probably tolerable for large tasks (e.g. digesting a 100-page pdf) especially with a "thinking" indicator, but often results in user frustration or abandoning the prompt.

INFERENCE - DECODE

- In the decode stage, tokens are generated “auto-regressively” - i.e. a model generates tokens one by one, and all prior tokens will be part of the input necessary to generate the next token (Exhibit 1). And to avoid recomputing the prior tokens, modern LLMs usually store the KV pairs of the prior tokens in a cache. This makes the memory requirement of KV cache grow linearly with the number of tokens. Further though weights in the model are still needed to calculate every new token and will take up some of HBM capacity, they are however “static” (i.e. read-only). Data in KV cache is on the other hand “dynamic” as it gets updated with every new token. And while weights can be shared over many users, KV cache differs across different users and grow linearly with the number of concurrent users. Growing linearly with the compound of token count & concurrent user count makes the decode stage "memory-bound". It is also why in a large AI deployment KV cache very possibly will require even more memory than the model weights, and become the dominant factor that determines the maximal concurrent users and context window sizes. And given memory is a bottleneck now, that further means how to deal with KV cache efficiently is a key factor in determining the deployment & revenue size of an AI model.

- The key performance index of the decode phase is usually "time per output token" (TPOT). The range of TPOT that users can accept varies by applications, possibly as long as 50ms (i.e. 20 tokens per second) for text chat, and less than 10ms for high-end use cases such as real-time voice conversation, coding and agentic works.

RAG (RETRIEVAL-AUGMENTED GENERATION)

- RAG is a technology that is widely used to argument AI models with data retrieved from databases. The database can for example contain domain-specific & private data to complement models trained on general & public data. The database can also store recent news/updates as the cost of training & keeping models updated is often prohibitively high. RAG is also very helpful in grounding AI models on facts & avoiding hallucination, or providing links to exact sources of information.

- The workloads of RAG can be thus divided into two parts - database generation and search. The generation part is to digest massive volumes of unstructured data (e.g. pdf, software code, web page) and organize them into a vector database with searchable indices. It requires first large capacity to store the unstructured data, ideally in SSDs over HDDs in order to reduce access time, and then large amounts of system DRAM to generate the vector database & indices. The use of HBM is relatively lighter. Most importantly the generation part is carries out before deployments of the database, and the workload is therefore offline and independent of the number of users or tokens. I.e. the RAG database generation is a "one-off" cost.

- The search part first converts a user query into a vector, typically with HBM, very quickly and then spends most time finding the nearest match in the RAG database, usually with system DRAM. System DRAM thus usually plays a bigger role than HBM & SSD.

- After the search finds & retrieves the nearest match, the data will be merged with user prompt, then go through the prefill & decode stage of inference to “generate” output tokens. The memory requirement hence will be as what we discuss above for inference.

on 31-Aug-2026

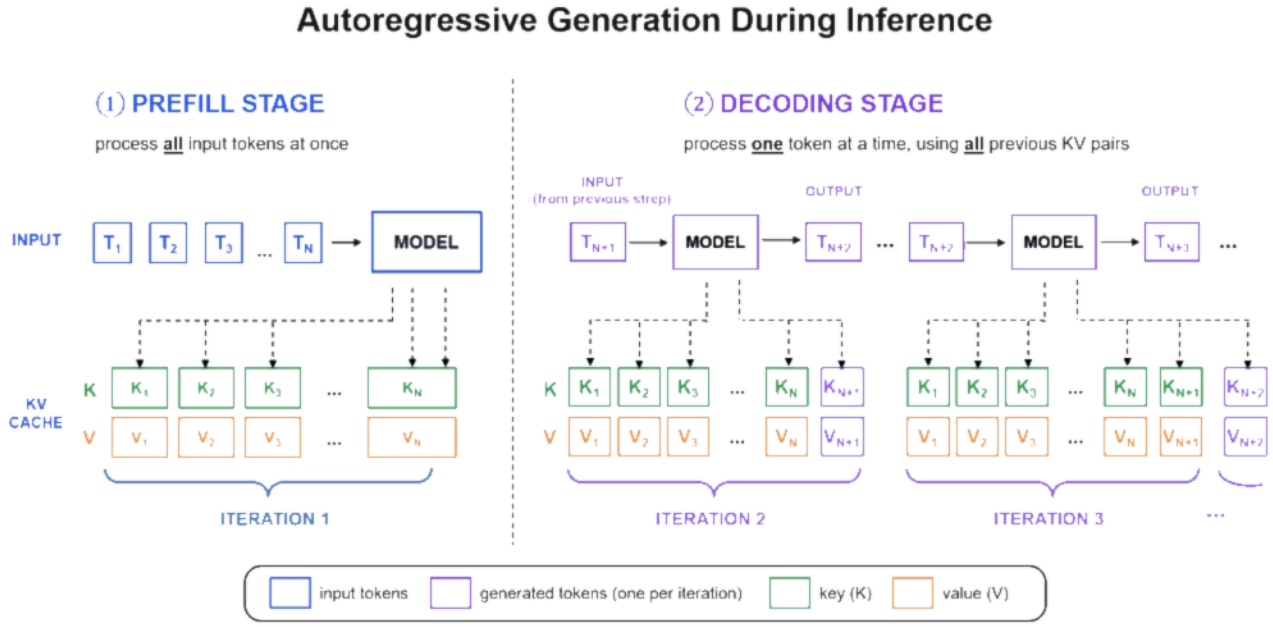
AGENTIC WORKFLOW

- Instead of a single step of "prompt → model → output", an agentic workflow shifts the paradigm to an autonomous, multi-step iterations. An agentic workflow will decompose a high-level objective into a plan of multiple smaller steps, interact with other models, or tools & environment external to the AI model to execute these steps, evaluate the results of these steps & make corrections if appropriate, until the objective is met or the process hits some limits.

- To support an agentic workflow, there is additional compute and memory need to oversee the entire workflow, decide on branch points and remember intermediate results in case the process needs to fall back to a prior state. The workloads & memory considerations of other AI models is as discussed above, but the interactions with external tools & environment result in a lot more demand for CPU & conventional memory from traditional (i.e. non-AI) servers.

- More importantly the output of an agent will become part of the prompt to the next agent, and the need of inference-prefill, inference-decode & the associate KV cache demand adds up very quickly when many agents are spawn, and compounds the memory challenges mentioned above.

EXHIBIT 1: **To generate a new token, LLMs need to consider all prior tokens. To avoid recomputing all the prior tokens, LLMs now usually store all the prior tokens in a KV cache.**



Source: Towards AI, Bernstein analysis

"ARCHITECTURAL" HIERARCHY & EMERGING TIERS IN AI

Below we first overview the "architectural" hierarchy of memory in a modern AI system. Please note it is based on the architectural positions & purposes of different memories, and thus will be similar to but different from another memory hierarchy that is based on hardware considerations. There is no official definition for this "architectural" hierarchy, but NVIDIA and Solidigm help us define some of the basic hierarchy below.

More importantly, NVIDIA added HBM as G1 & "CMX" (Content Storage Memory) as G3.5 tier below. And with our own analysis & judgments, we additionally map HBF (High Bandwidth Flash), CXL (Compute Express Link) & "Storage Next" memory into the hierarchy too, as they're emerging tiers to architecturally address the "memory wall" in high performance computing. These additional tiers are underscored below and also in Exhibit 2so that investors can differentiate them from the tiers that are commonly found in a usual "non-AI" server.

on 31-Aug-2026

G0 - MEMORY INSIDE PROCESSOR

- Extremely low latency, very small capacity, used for immediate active computations, usually through SRAM embedded inside the same logic dies of the accelerator (Exhibit 3). G0 memory is therefore manufactured by foundries like TSMC along with the entire logic dies.

G1 - HBM (HIGH BANDWIDTH MEMORY)

- Employing CoWoS to co-package stacked DRAM dies with accelerator together (Exhibit 4) provides high bandwidth & low latency. Hot, latency-critical data from model weights, KV cache, etc. used in active generation will be stored here.

G2 - SYSTEM MEMORY

- Often also called DDR DRAM, these DRAMs are predominately based on the DDR (double data rate) standard and are usually put right next to CPU in a module format (e.g. SOCAMM2 used in NVIDIA's Grace & Vera CPU (Exhibit 8, Exhibit 9), DIMM (dual inline memory module, Exhibit 27)). They have higher capacity than HBM but also longer latency. These DRAMs are managed by CPU but GPUs from NVIDIA & AMD however can bypass CPU and access these DRAMs directly.

G2.5 - CXL (COMPUTE EXPRESS LINK) MEMORY

- Originally developed by Intel, CXL is now widely supported. It is initially meant for DRAM, with the aim to combine physically separate DRAM together in a logically-common pool to avoid waste. For example through CXL, CPU can access HBM held by accelerator and accelerator can access the G2 system memory managed by CPU. G3 memories that are away from CPU and are connected with PCIe (PCI Express) sockets on the motherboard (often called memory expander, Exhibit 10, Exhibit 11) can be shared too if they're CXL-compliant. CXL is later expanded to support NAND as long as these NAND-based products are enhanced to be more like DRAM and become CXL-compliant. A common approach is to pair NAND with DRAM or SRAM and use DRAM/SRAM as cache to shorten the latency of the products. Samsung's CMM-H (CXL Memory Module-Hybrid) is an example as its a hybrid of DRAM and NAND together (Exhibit 12).

G2.6 - "STORAGE NEXT" MEMORY

- This is an ecosystem initiative led by NVIDIA and well-supported by basically all memory suppliers and many other industry participants (Exhibit 13). "Storage Next" aims to 1) turn the management of memory from CPU-centric to GPU-centric and more importantly 2) make NAND more like DRAM in order to enjoy the low cost and large capacity of NAND, but also deliver low latency (i.e. high IOPS (input/out operations per second)) and small data access granularity that is close to what DRAM can support. This is why NVIDIA puts "Storage Next" as an additional tier between DDR DRAM and local SSD (Exhibit 14). We define this as G2.6 tier, as it is very similar to the CXL-compliant NAND products mentioned in G2.5 tier. These products are often a hybrid of DRAM and NAND, and often use SLC (single-level cell) NAND that we will elaborate below. KIOXIA's GP-series SSD therefore offers "super high IOPS" and is a good example of "Storage Next" memory products (Exhibit 15).

G3 - LOCAL SSD

- This is NAND-based memory follows the standard of PCIe (a hardware standard (e.g. how highway should be constructed)) and NVMe (Non-Volatile Memory Express, a logic standard (e.g. the traffic rule for vehicles on highway)). These SSDs are usually on the edge of motherboard but stay localized in a compute node (i.e. a tray that has GPUs & possibly pairing CPUs). It is hence sometimes called "Compute SSD". This tier extends capacity beyond DRAM but is not shareable across the pod. The access to local SSD is managed by CPU but some accelerators (e.g. from NVIDIA & AMD) can bypass CPU and access the local SSD directly.

G3.5 - CMX (CONTEXT STORAGE MEMORY) MEMORY & STX (STORAGE TECHNOLOGY EXTENSION)

- This is called ICMS (Inference Context Memory Storage) before & is an additional tier specifically added by NVIDIA & optimized for KV cache. CMX-compliant SSDs sit inside data nodes (i.e. trays that are designed to store data), but are accessible across multiple compute nodes in the entire pod, possibly for shared, evolving content for many agents simultaneously. The communication among the compute and data nodes is via Ethernet, with a Spectrum-X switch chip

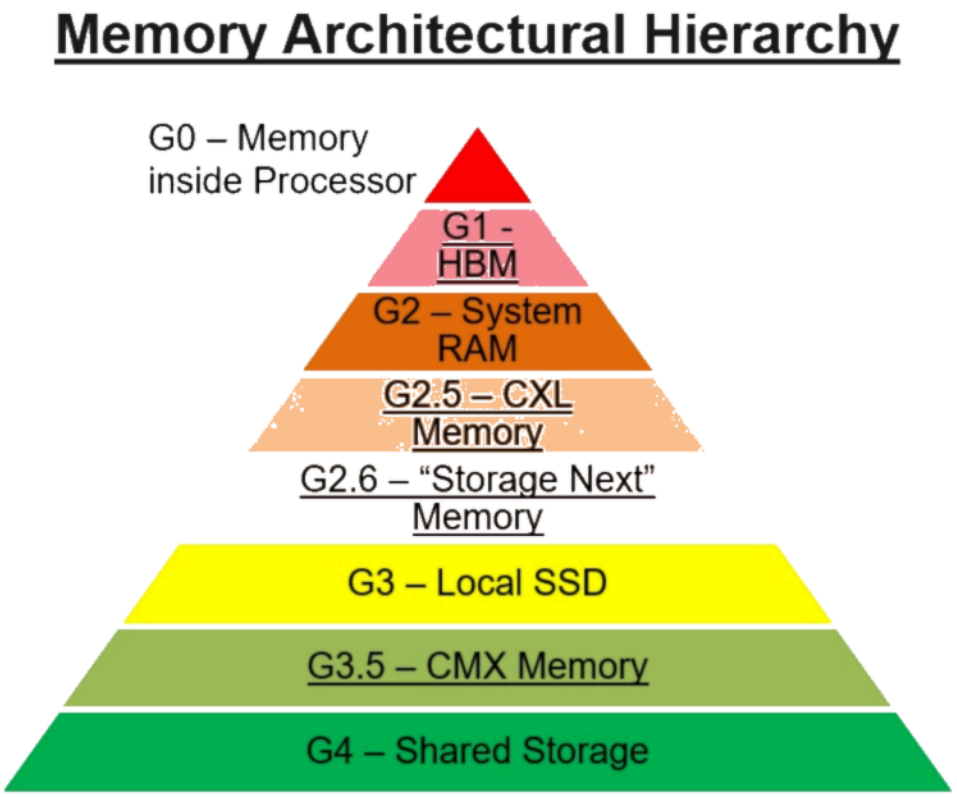
on 31-Aug-2026

in between, and a BlueField-4 DPU in the compute node & another talking to SSD via PCIe in the data node (Exhibit 16). The access to CMX memory can be done by GPUs directly without CPUs involved. NVIDIA further introduced STX as the reference architecture to integrate solutions & ensure compatibility from a wide array of storage & ODM partners to facilitate the adoption of CMX memory.

G4 - SHARED STORAGE

- This can be NAND-based SSD or HDD connected with PCIe or through some networks for long-term storage, and for data that is durable but is not on the immediate critical path. When with compatible networking & software, GPU possibly can bypass CPU and access G4 storage directly too.

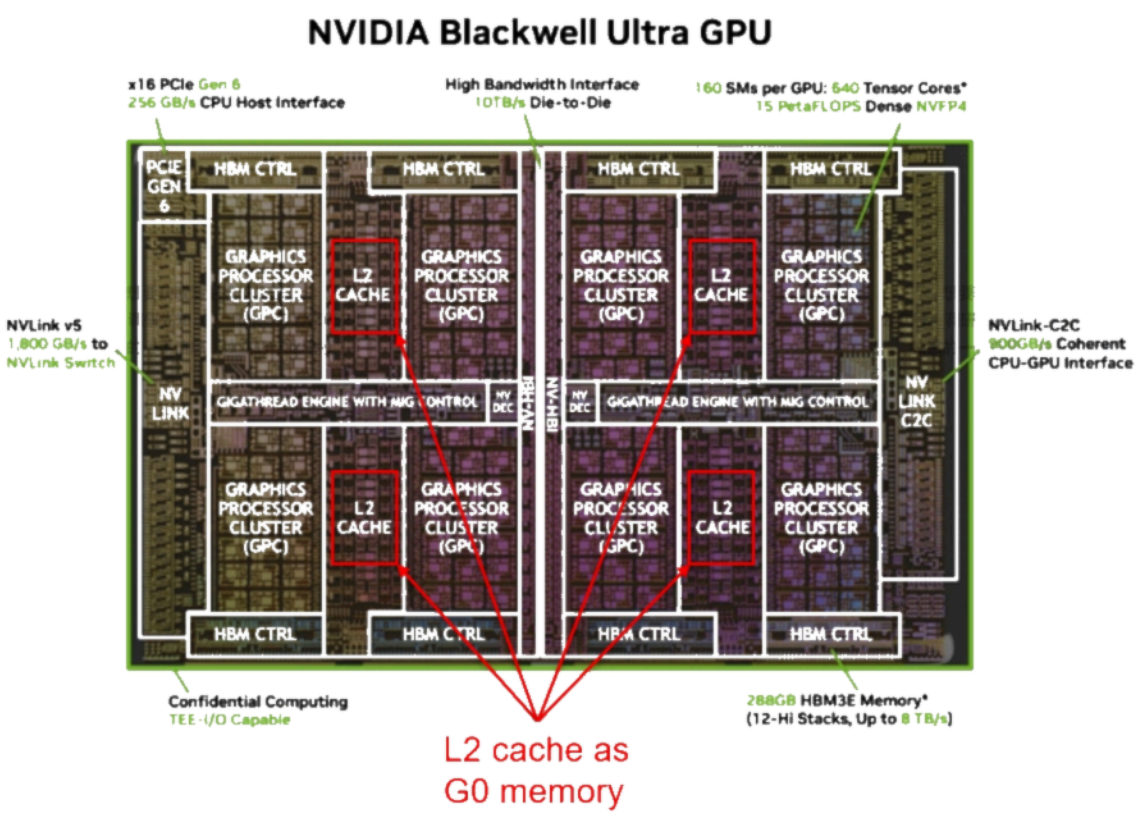
EXHIBIT 2: **We define an "architectural" memory hierarchy to summarize memories at different tiers in a high performance or AI system.**



Source: NVIDIA, Solidigm & Bernstein analysis

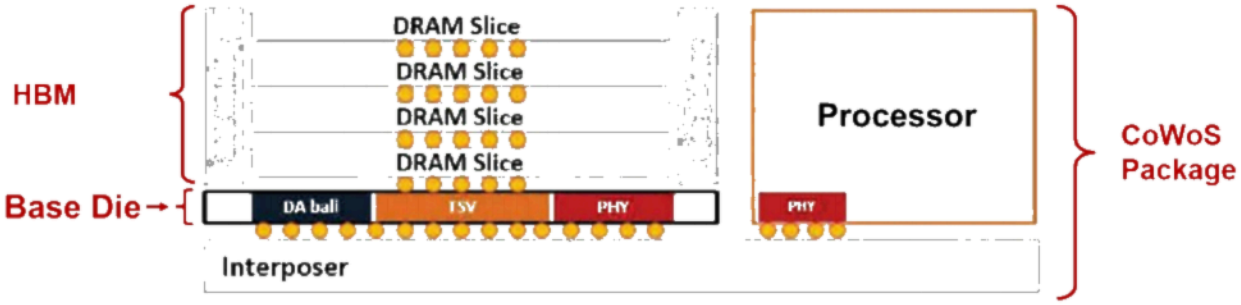
on 31-Aug-2026

EXHIBIT 3: **NVIDIA defines the memory (usually SRAM) embedded in the same logic XPU die "G0 memory".**



Source: NVIDIA and Bernstein analysis

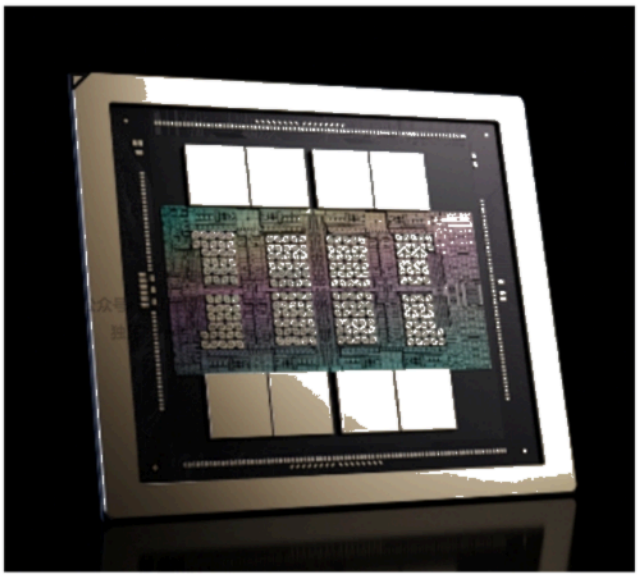
EXHIBIT 4: **HBM usually sits right next to the accelerator in a single CoWoS package to enable high bandwidth and low latency.**



Source: SK hynix and Bernstein analysis

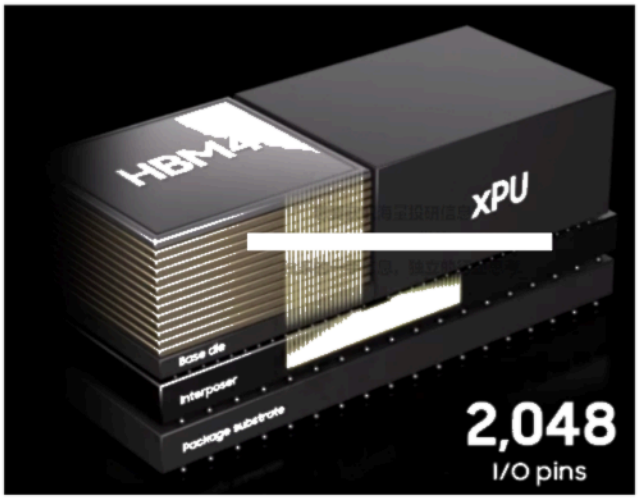
on 31-Aug-2026

EXHIBIT 5: **The exhibit below shows 8 HBM stacks are packaged in a Rubin GPU package.**



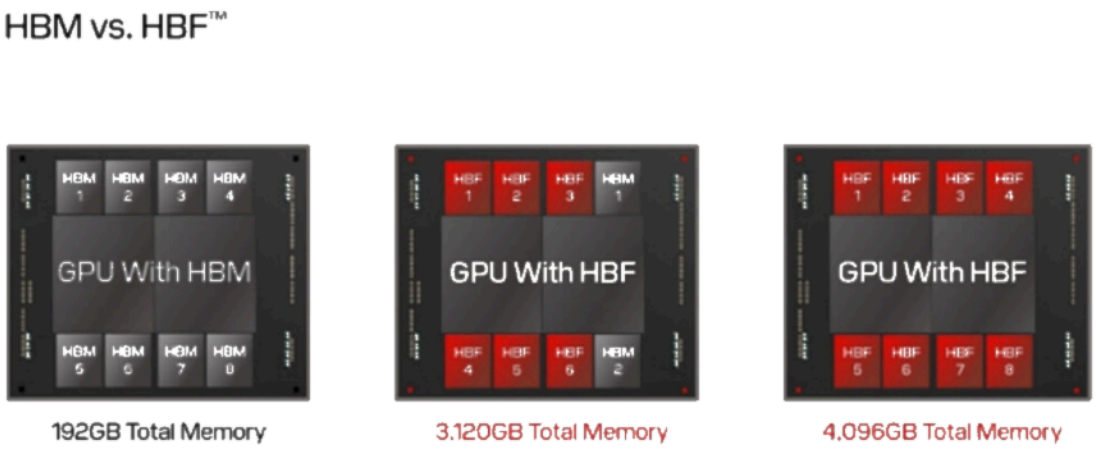
Source: NVIDIA

EXHIBIT 6: **Each HBM stack has multiple DRAM dies & a base die bonded together vertically, & is then packaged with processor with CoWoS-equivalent inside a single package.**



Source: Samsung

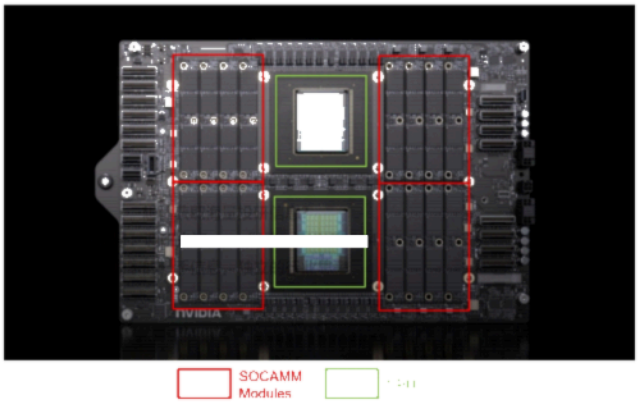
EXHIBIT 7: **Similar to DRAM dies stacked in HBM, NAND dies are stacked in HBF. HBF stacks then are mixed with HBM stacks and are also packaged with XPU with CoWoS-equivalent packaging technology.**



Source: Sandisk

on 31-Aug-2026

EXHIBIT 8: **A Vera CPU is paired with 8 SOCAMM2 modules (4 a side, and 2 sides). The exhibit below shows 2 Vera CPUs and 16 SOCAMM2 modules in total in a mother board.**



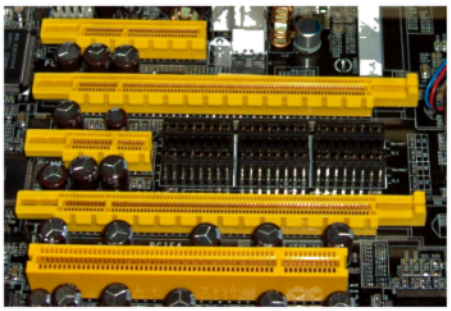
Source: NVIDIA and Bernstein analysis

EXHIBIT 9: **SOCAMM2 module uses LPDDR (low-power DDR) dies but packaged them in a module format different from the usual DIMM.**



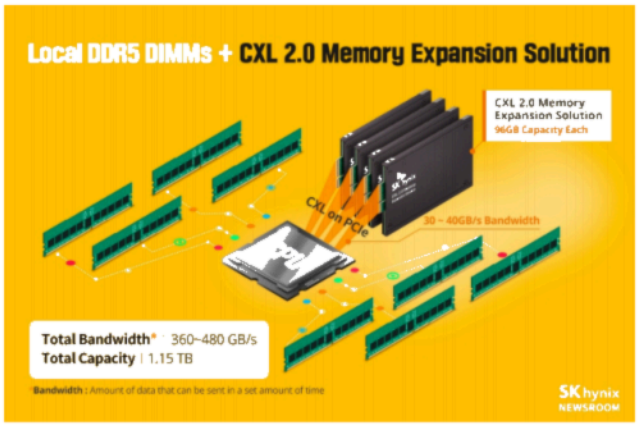
Source: Samsung

EXHIBIT 10: **Example of PCIe sockets on a motherboard**



Source: Wikipedia

EXHIBIT 11: **CXL combines the memores connected with PCIe with system DRAM into a same common memory pool.**



Source: SK hynix

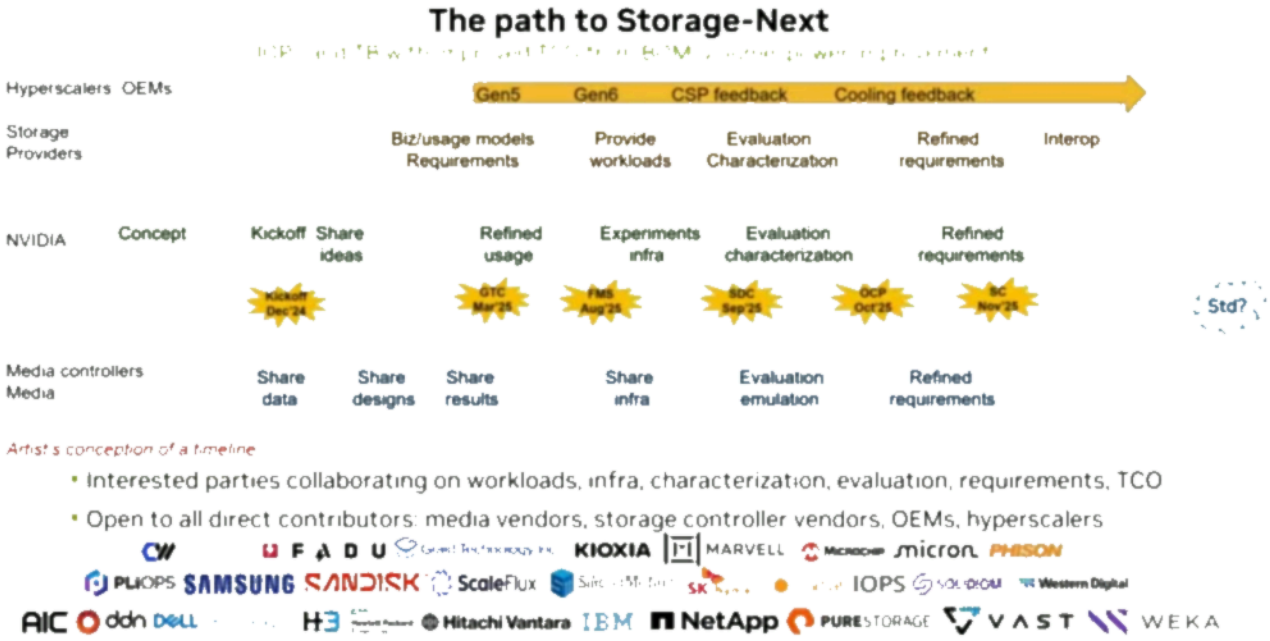
on 31-Aug-2026

EXHIBIT 12: **A CXL-compliant SSD from Samsung. CXL is mainly meant for DRAM but can support NAND-based SSD as long as it is enhanced (e.g. through leveraging DRAM as the cache) to be "DRAM-like".**



Source: Samsung

EXHIBIT 13: **NVIDIA is working with Samsung, SK hynix, Micron, KIOXIA, Sandisk and many other companies to prompt "Storage Next".**



Source: NVIDIA

on 31-Aug-2026

EXHIBIT 14: **NVIDIA is defining "Storage Next" memory as an additional tier between DDR DRAM & Compute SSD (i.e. local SSD).**



Source: ScaleFlux, NVIDIA

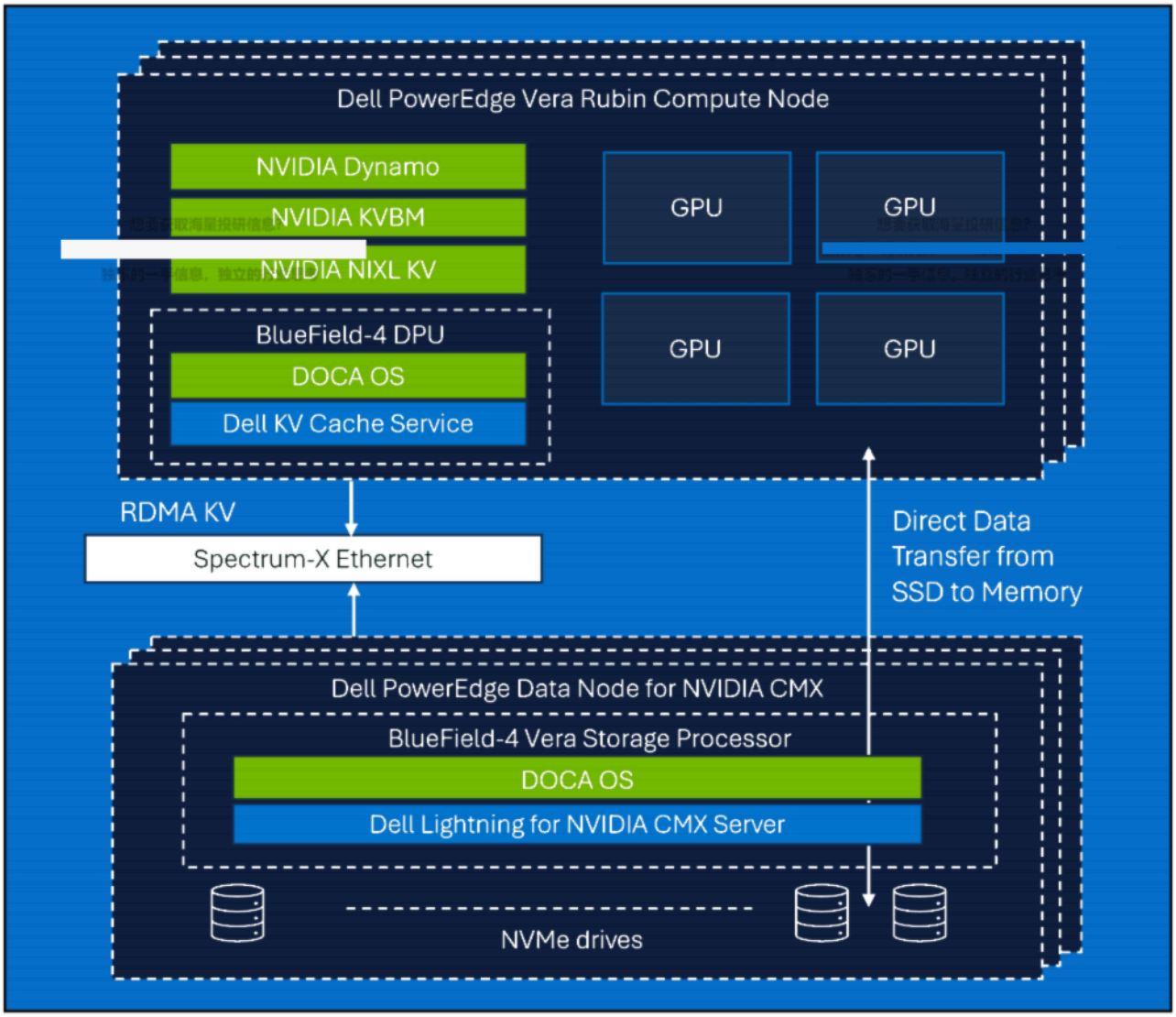
EXHIBIT 15: **In response to NVIDIA's "Storage Next" program, KIOXIA takes advantage of XL-Flash to deliver GP-series SSDs with "super high IOPS".**



Source: KIOXIA

on 31-Aug-2026

EXHIBIT 16: **As shown in a diagram from Dell below, CMX memory sits in a data node & communicates with compute nodes through Ethernet, with the help of BlueField-4 DPU on both ends & a Spectrum-X switch chip in between.**



Source: Dell

"HARDWARE" HIERARCHY & INNOVATIONS FOR AI

In addition to the "architectural" hierarchy above, we summarize another hierarchy below, based on the different cell structures & storage mechanisms these memories employ (Exhibit 17). The tiers in this "hardware" hierarchy correspond to those in the "architectural" hierarchy, but only approximately. In addition, different packaging and module technologies will be added onto different tiers of the "hardware" hierarchy. Mixing & matching them together offers the supply chain even more flexibility to derive solutions optimal for the needs.

SRAM (STATIC RANDOM ACCESS MEMORY)

- SRAM is faster than DRAM, but to build a cell & store a bit SRAM usually takes 6 transistors ("6T") & more silicon area than DRAM ("1T1C", 1 transistor & 1 capacitor). Both DRAM & NAND are volatile (i.e. data will be lost when power is off), but DRAM is "dynamic" & requires supporting circuit to constant refresh & recharge the capacitor in the cell, whereas SRAM is "static" and does not refresh circuits.

on 31-Aug-2026

- SRAM in a dedicate die & package can be still found in niche segments, but most of SRAM is now "embedded" - i.e. they are part of logic dies and are fabricated together by logic foundries like TSMC. The short access time (i.e. low latency) makes SRAM ideal for the on-chip cache and correspond to the G0 tier in the "architectural" hierarchy above. Cache can be further divided into different levels too, with level 1 (L1) being closest the computation circuits but smallest in capacity. L2 & L3 cache will have larger capacities, but will be slightly less responsive as they, even inside the same piece of logic die, are a bit away from the computation circuits.

- **LPU**

  - An innovation that we are seeing because of AI is stacking an SRAM die on top of an AI accelerator. For example LP30, the current version of Grog's LPU (language processing unit) (that NVIDIA licenses), employs on-chip SRAM to provide very short access time to latency-critical part of inference demand, but there are reports that the next version LP40 will have an SRAM die on top of a logic processor die, & bond them together with hybrid bonding, to compromise on access time only slightly but deliver more memory capacity at an affordable cost. When used this way, SRAM is effectively the G2 tier in the "architectural" hierarchy.

- **Embedded DRAM**

  - Finally "embedded DRAM" is DRAM embedded in logic dies but differs from embedded SRAM as it is "dynamic" (i.e. requiring refresh). Embedded DRAM hence should be considered in the G0 tier and will compete with embedded SRAM as on-chip cache, especially when cache capacity is larger (e.g. L4) & the lower per-cell area of DRAM structure justifies the overhead of the supporting refresh circuits. Vs. embedded SRAM, embedded DRAM however has a very small market share, and is only in for example IBM POWER CPU & PlayStation processor.

DRAM (DYNAMIC RANDOM ACCESS MEMORY)

- NAND is also a common memory technology, but DRAM is even more important when we consider the portion of revenue & profits from DRAM vs. NAND for the industry (Exhibit 18). DRAM is therefore better understood by investors, but we still highlight a few points below.

- HBM is a kind of DRAM, & shares the same wafer capacity as conventional DRAM. DRAM thus maps into largely the G1 & G2 tier above. As DRAM can be used in CXL memory expander, DRAM can be found in the G2.5 tier too. In fact, as many NAND-based products use DRAM as cache to store pre-fetched data to hide the long latency of NAND from users, DRAM can be found in other tiers, but just as a complement and thus with much lower volume.

- **$4F^2$ DRAM & 3D DRAM**

  - DRAM's technology evolution is still mostly driven by Moore's Law (i.e. lateral scaling or "node") (Exhibit 19). This is strategically important as this implies EUV is necessary to advance DRAM technology, and the lack of EUV will make China, or specifically the Chinese DRAM company CXMT, eventually not competitive in DRAM market. To overcome the lack of EUV, China is researching $4F^2$ DRAM. With F denoting the smallest dimension that lithography can pattern, the current approach of shrinking cell size & increasing bit density is employing more advanced lithographic equipment (e.g. EUV) to make F smaller. $4F^2$ attempts to achieve the same objective but differently by moving peripheral circuits from the side of the capacitor to below or even to another piece of wafer (Exhibit 20). Currently it takes $6F^2$ to create the "1T1C" structure & $4F^2$ builds "1T1C" too but with a different shape & hence a small areal size (Exhibit 21). That said, similar to adopting FinFET in logic, $4F^2$ DRAM offers only one-time benefit, and further technology advancements still require lateral shrink. China is also researching 3D DRAM which would like to follow 3D NAND and aims to put multiple cells vertically in a piece of wafer (Exhibit 20). 3D DRAM hence amortizes the same areal footprint over many cells and make the average cell area below $4F^2$. By building more & more cells vertically, 3D DRAM isn't just a "one-time" change and promises scalability (i.e. longer technology runway) that $4F^2$ DRAM can't. 3D DRAM is however more technologically

on 31-Aug-2026

challenging & commercialization of 3D DRAM will be even more years away.

- **zHBM**
  - As the connection between HBM stacks and XPU now is through an interposer beneath them, the packaging approach is hence "2.5D". With zHBM, Samsung aims to take it further to "3D" by putting HBM stacks on the top of XPU (Exhibit 22). zHBM hence shortens the communication between HBM and XPU further, but we wonder how the DRAM dies in zHBM can withstand the heat dissipated from XPU in the bottom and keep its operations within the designed specifications. Probably to improve heat dissipation, Samsung plans to use wafer-to-wafer hybrid bonding to bond the DRAM dies together. We understand that hybrid bonding indeed requires no gaps in between dies and can dissipate heat faster, but also wonder if the yield & thus cost of hybrid bonding is ready for commercial production.
- **NVHBM**
  - This is basically HBM customized with base die designed by NVIDIA & manufactured by TSMC. By moving "memory controller" circuits from XPU to the base die, NVHBM should reduce the die size & cost of XPU (or keep die size/cost the same but give more areas to compute circuits), and enhance bandwidth & lower power consumption. Though the base die is designed by NVIDIA, NVIDIA tries to make NVHBM "open" by 1) standardizing NVHBM implementation so that multiple memory suppliers can provide NVHBM, & 2) expanding NVLink Fusion with NVHBM so that XPUs from others can adopt NVHBM too. Amazon's Annapurna Labs for example has announced to work with NVIDIA on NVHBM.
  - Compared to standard HBM, NVHBM reduces the value provided by memory suppliers as the base die currently designed by memory suppliers will be designed by NVIDIA. Some memory suppliers (e.g. Samsung & Micron) manufacture the base die in the standard HBM, but the manufacturing value will be captured by foundries, most likely TSMC, in NVHBM. Memory suppliers currently differentiate on base die but the room of differentiation will be "standardized" away by the base die from NVIDIA.
- **XBM**
  - Intel unveiled XBM (cross-batch memory) in patents (here). The patent refers to using "backend" and thin-film transistor (TFT), and many thus link it together with another patent (here) describing a similar transistor structure. We consider XBM "emerging memory" too. Though the transistor described is also DRAM, but it builds cells in "back-end of line (BOEL)[1]" among metal wires (Exhibit 23, Exhibit 24), whereas the usual DRAM builds cells in "front-end of line (FOEL)[2]" and keeps BOEL with purely metal wires. This is an important difference, and as a result, it is possible that in a single piece of silicon these backend transistors of DRAM can co-exist with logic circuits who build transistors in FOEL. Without details, we're not sure manufacturing compatibility between XBM and DRAM vs. logic chip and hence who, memory suppliers, Intel, or others, will be more suited to produce XBM in the future. The possible co-existence of DRAM & logic circuits also makes XBM a candidate of PIM that we'll elaborate below.
- **ZAM**
  - In addition to lateral scaling, packaging & module is also being innovated on DRAM too. HBM is a proven example of using DRAM does & packaging them differently to deliver higher performance, and ZAM (Z-angle memory) is a similar attempt tried by Intel now. The main distinction of ZAM is reorienting DRAM dies in a stack by 90 degrees so that they become side-by-side horizontally (Exhibit 25). Heat then can dissipate more quickly through the vertical gaps among them. In an interview, Intel's partner SAIMEMORY (a 100% subsidiary of Softbank) suggested that these DRAM dies can be still

[1] Back-end of line (BOEL) refers to the later part of wafer fabrication process flow, usually to create metal wires & connect them together transistors.

[2] Front-end of line (FOEL) refers to the early part of wafer fabrication process flow, usually to create transistors.

on 31-Aug-2026

sourced from memory suppliers, implying that Intel & SAIMEMORY may focus only on the unique packaging & the base die. In the same interview SAIMEMORY said that reorientation makes it difficult to physically connect DRAM dies & the base die, and how to wirelessly transmit data among them is a challenge of ZAM, which is aimed to achieve practical use in FY2029. ZAM should correspond to the G1 tier.

- **HBC**

  - Qualcomm's HBC (high bandwidth compute) is another attempt of differentiating DRAM dies with a different package. While ZAM resorts to the unusual orientation to improve performance, HBC decides to compromise on performance in order to avoid the high cost of CoWoS. Using conventional packaging, HBC won't have as many wires connecting memory with accelerators, but avoids the cost of interposer and 2.5D organic substrate (Exhibit 26). Additionally, HBC will use LPDDR (low-power DDR) die, vs. HBM using DDR dies, to lower power consumption. Qualcomm targets to ship HBC gen-1 in its fiscal year 2027 and gen-2 in 2028. HBC is also considered as the G1 tier.

- **MRDIMM, SOCAMM2, LPCAMM2**

  - MRDIMM (multiplexed rank DIMM) (Exhibit 27) & SOCAMM2[3] (small outline compression attached memory module) (Exhibit 8, Exhibit 9) are examples of module innovations for DRAM. Compared to SOCAMM2, MRDIMM is a smaller change from the usual RDIMM (registered DIMM) used in servers as both MRDIMM & RDIMM share the same physical dimensions but MRDIMM puts more interface chips on the module to enable higher bandwidth. SOCAMM2 is a bigger change and has the shape & dimensions of the module different from RDIMM's. By lowering its power voltage to support LPDDR, widening the data bus width (128 vs. 64 bits) & lowering vertical profile to avoid obstruction, SOCAMM2 vs. DIMM can reduce power consumption of memory module, increase bandwidth per module, and improve air flow & thus thermal dissipation of AI server. SOCAMM2 is first supported by NVIDIA's CPU Vera, followed by AMD's CPU Verano. LPCAMM2[4] (Exhibit 28) similarly is a new module format, but is effectively a counterpart of SOCAMM2 for some of recent AI-focus notebook PCs. MRDIMM, SOCAMM2 & LPCAMM2 are all in the G2 tier above.

- **PIM (processing-in-memory)**

  - PIM is sometimes called CIM (compute-in-memory). We put PIM under DRAM though technically PIM puts a combination of logic & memory, mostly DRAM & possibly also NAND, circuits inside the same piece of silicon, so that processing carried out by logic circuits happens inside a memory chip. As traffic between logic and memory is often the bottleneck of computers, PIM's value proposition is to reduce traffic and remove the bottleneck. PIM however differs fundamentally from the computer architecture that was proposed by the mathematician & scientist John von Neumann in mid-1900s, and is now widely followed by nearly all computers. The Von Neumann architecture separates processing from memory whereas PIM mixes them together. Memory suppliers have been exploring PIM for a while, & proposing to realize PIM in different implementations (Exhibit 29), but the adoption has been very limited as it requires changes to processor, software and network, etc. as basically the entire system needs to be redesigned to realize the full potential of PIM. PIM also differs from the current supply chain, which after decades of development, has matured into two manufacturing bases dedicating for logic and memory separately with very large scales and very good efficiency.

SCM (STORAGE CLASS MEMORY)

- Unlike SRAM, DRAM or NAND that refer to a particular memory technology, SCM refers to generally all memory technologies whose properties are in between DRAM & NAND. Some of these technologies can be "emerging memory" as they differ from DRAM or NAND in the fundamental structure and mechanism, or can be NAND after enhancements. SCM broadly maps into

[3] The first version of SOCAMM is a proprietary trial of NVIDIA. SOCAMM2 is formalized as an industry standard released by JEDEC.

[4] LPCAMM is the early prototype led by Samsung, and LPCAMM2 becomes an industry standard formalized by JEDEC.

on 31-Aug-2026

the range of the low-end of the G2 tier to the high-end of the G3 tier.

- **MRAM, PCM (including XPoint), RRAM**

  - We classify MRAM (magnetic random access memory), PCM (phase change memory), & RRAM (resistive random access memory) as "emerging memory" as their underlying mechanism of retaining information differs from DRAM and NAND fundamentally. Please note the 3D XPoint memory that Intel once offered is a favor of PCM. We discussed these "emerging memory" technologies in details in an earlier report (Weekend Tech Byte - Global Memory Series (Part 1): a Memory Technology Primer) and Exhibit 30 compares these "emerging memories" with SRAM, DRAM, NAND & HDD on key characteristics.

- **Z-NAND, XL-FLASH**

  - Some NAND products achieve higher IOPS, lower latency & small data granularity through pairing NAND with some DRAM or SRAM. They can be viewed in the G2.5 tier (when CXL-compliant), or G2.6 (when supporting the "Storage Next" initiative), or even high-end segment of G3 (when neither CXL nor "Storage Next" is supported). Also considered as good technologies for the G2.6 tier, Z-NAND & XL-FLASH go further by using SLC or MLC to make NAND behave even more like DRAM. KIOXIA's XL-FLASH for example resorts to SLC (single-level cell, i.e. 1 bit per cell) & later MLC (multi-level cell, i.e. 2 bits per cell) & sacrifices on capacity & cost benefit of TLC (triple-level cell, i.e. 3 bits per cell) to make its GP-series SSD "super high IOPS" (Exhibit 31), as high as 10M IOPS this year & 100M+ in 2027. Samsung productizes Z-NAND into Z-SSD to achieve "ultra low latency" (Exhibit 32). They both claim to be SCM, and are set to compete again Intel's PCM-based 3D XPoint earlier, and now to meet the requirements of some particular IO workloads. Beyond SLC or MLC, pSLC (pseudo-SLC) is also used by some SSDs. While the NAND cell is capable of storing multiple bits (TLC in most cases), pSLC will set aside part of the cells and store only a bit in them. As TLC is the most common multi-level technology currently and has the best economy of scale, pSLC allows suppliers to take advantage of that, but offers the high endurance & fast speed of SLC. Micron's XTR SSD is an example of pSLC.

  - Overall how NAND can be optimized for specific segments of AI inference need is summarized well in a slide from KIOXIA (Exhibit 33). XL-FLASH leverages SLC/MLC & enables SCM-class products for NVIDIA's "Storage Next" initiative (i.e. G2.6 tier). TLC is used in CMX-compliant products (i.e. G3.5 tier) thanks its right balance between performance and cost. QLC (quad-level cell, i.e. 4 bits per cell) is deployed in products that have large enough capacity to compete against HDD (i.e. G4 tier).

NAND FLASH

- **HBF**

  - Led by Sandisk and SK hynix, HBF (high-bandwidth flash) aims to match the bandwidth of HBM but with lower cost & more capacity. Should we map it to the architectural hierarchy, HBF likely can be viewed as G1.5 tier. As NAND is 1 to 2 tiers lower than DRAM in the hardware hierarchy and HBM is at the top end of DRAM, it requires a big leap over multiple tiers to make HBF performant enough to complement HBM and be G1.5. The technological hurdle is thus high.

- **zNAND-O**

  - While technologies above are meant to address the needs of AI in data centers, Samsung's zNAND-O targets to run AI on devices (edge devices). zNAND-O uses TLC to offer large capacity and thus differs from Z-NAND which resorts to SLC or MLC to shorten latency. The packaging of zNAND-O is similar to that of HBM - multiple NAND dies will be first bonded in a stack, and then one or multiple stacks will be connected with the processor with advanced, likely wafer-based, packaging technology (Exhibit 34).

on 31-Aug-2026

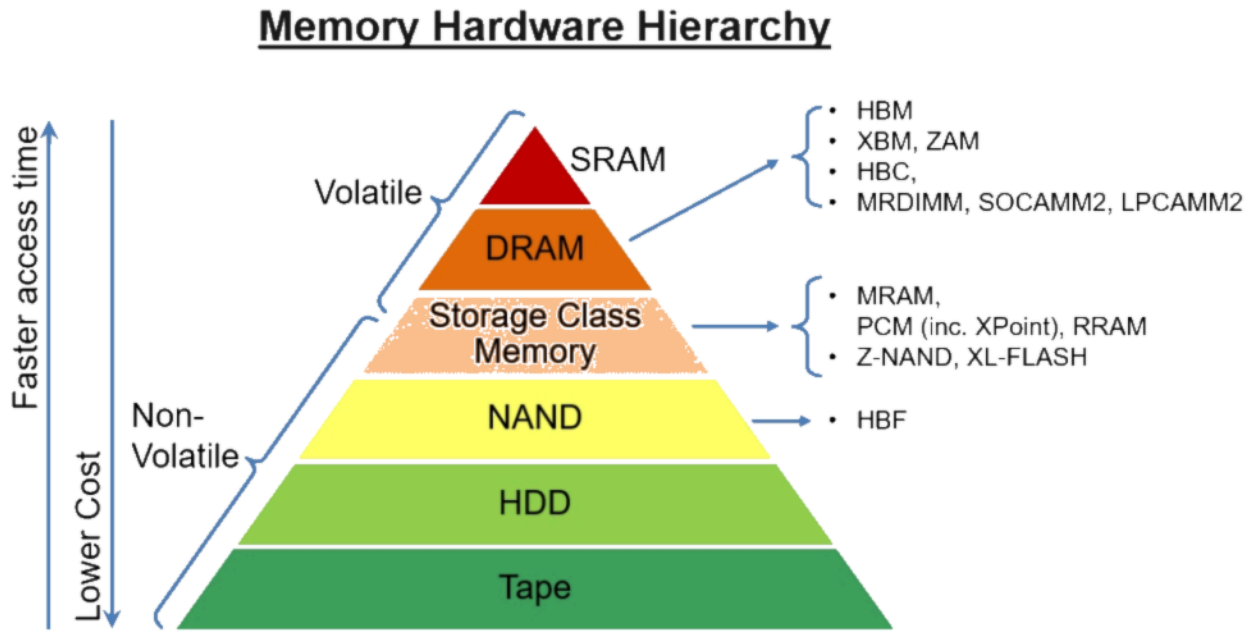
HDD (HARD DISK DRIVE)

HDD is the second-bottom tier in our memory hardware hierarchy, and still represents a good storage choice for "cold data" which requires large low-cost capacity but is infrequently accessed. HDD is in the G4 tier above and is benefiting from booming data storage needs including KV cache overflow as the least critical data trickles down.

TAPE

Right at the bottom of the hardware hierarchy sits tape. Tape is traditionally only used for data archival & corresponds to also the G4 tier above, but below HDD. Due to the huge demand for data storage and KV cache needs exploding, we have heard that even tape needs are growing as there is not enough NAND or HDDs to fulfill current demand.

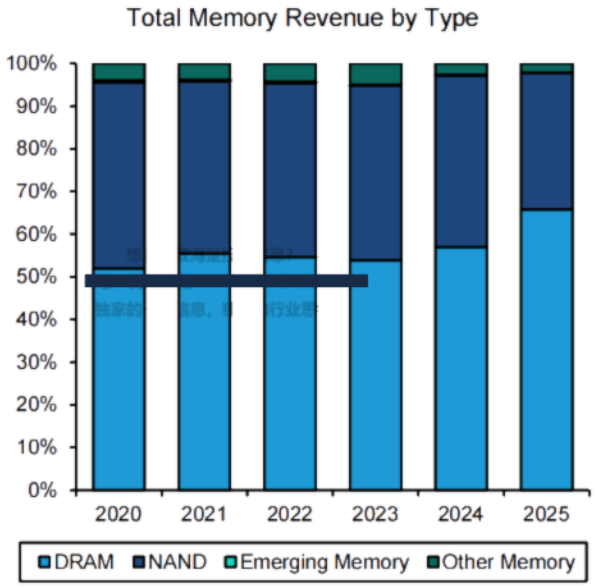
EXHIBIT 17: **Memory can also be classified based on different cell structures and storage mechanisms.**



Source: Bernstein analysis

on 31-Aug-2026

EXHIBIT 18: **DRAM accounts for the majority of overall memory revenue.**



Source: Company reports, Gartner and Bernstein analysis

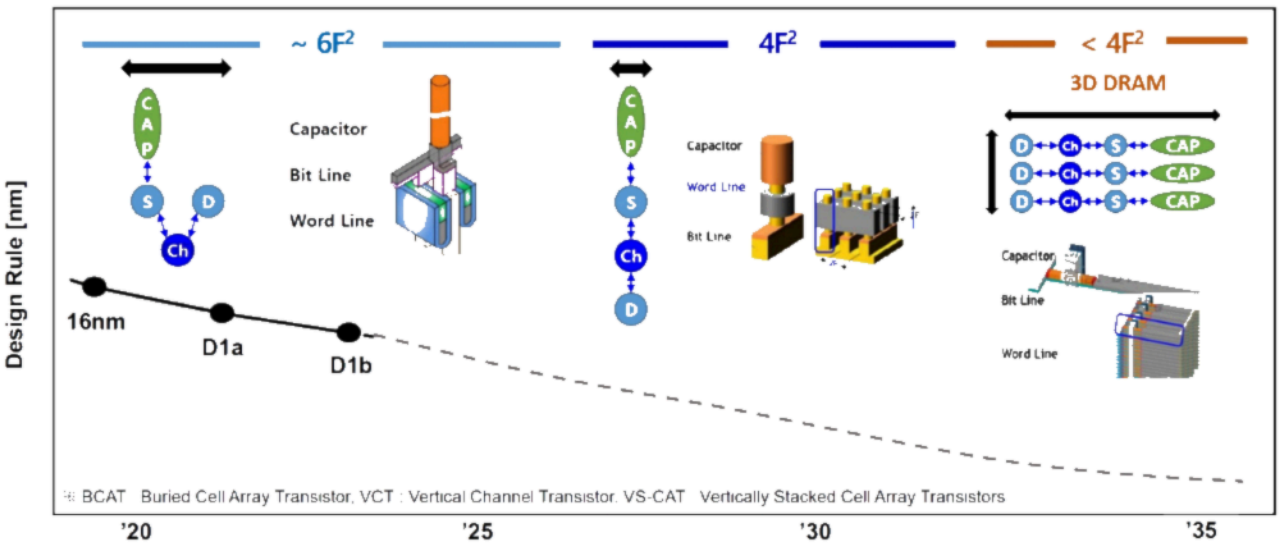
EXHIBIT 19: **DRAM's technology evolution is still mostly driven by Moore's Law.**

	2025 MS	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026E	2027E	2028E
Samsung	34%	28 nm	25/23nm	20nm	18 nm (1x)	17 nm (1y)				15/16 nm (1z)*		13/14 nm (1a)		12 nm (1b)	1c nm		1d nm
SK hynix	36%	29 nm		25 nm	21 nm	18 nm (1x)		17 nm (1y)		15/16 nm (1z)		14 nm (1a)*		12 nm (1b)	1c nm		1d nm
Micron	22%	30 nm		25 nm	20 nm	17 nm (1x)		17nm (1y		15/16 nm (1z)		14 nm (1α)		13 nm (1β)	1γ nm*		1δ nm
Nanya	1%	30 nm					2x nm					1A nm			1B nm		1C nm
CXMT	6%						G1 (2y nm)					G3 (1x-1ynm)			G4 (1z-1anm)		G5

*: Beginning of EUV adoption in mass production; Samsung adopted EUV in 1x & SK hynix in 1y very limitedly

Source: Company release, and Bernstein estimates and analysis.

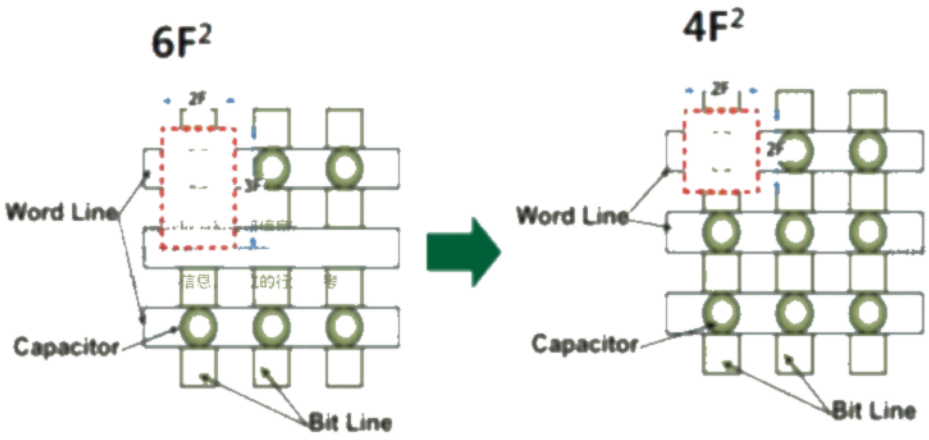
EXHIBIT 20: **6F2 is the prevailing structure of DRAM cell, but the industry, including CXMT, is trying to advance to 4F2 structure & to 3D DRAM.**



Source: Samsung, Bernstein analysis

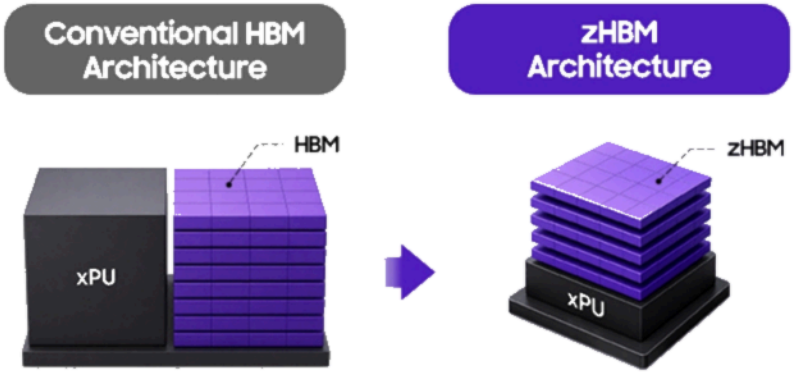
on 31-Aug-2026

EXHIBIT 21: **A top view of DRAM cell arrays to illustrate the chagne from 6F2 to 4F2. By moving peripheral circuits below capacitor in the same wafer, or even to a different wafer, 4F2 architecture reduces the area of a DRAM cell from 6F2.**



Source: Samsung

EXHIBIT 22: **zHBM moves HBM stacks from the sides of XPU to the top. Samsung also aims to use W2W hybrid bonding to improve the interconnections and thermal dissipation across the DRAM dies.**

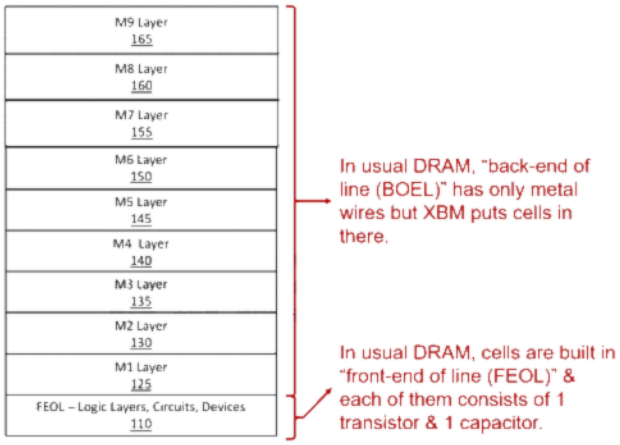


Source: Samsung

人在草木间

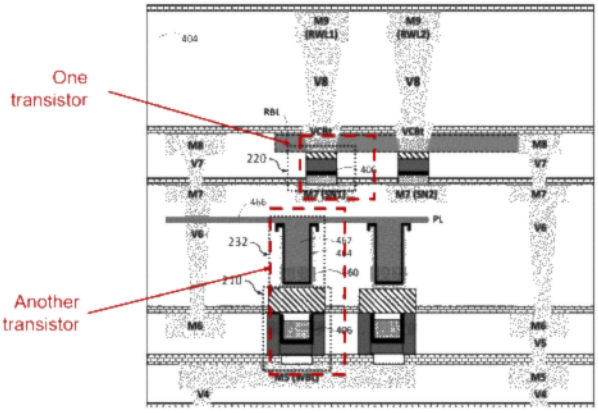
on 31-Aug-2026

EXHIBIT 23: **Usual DRAM builds cells in FEOL but XBM builds cells in BOEL.**



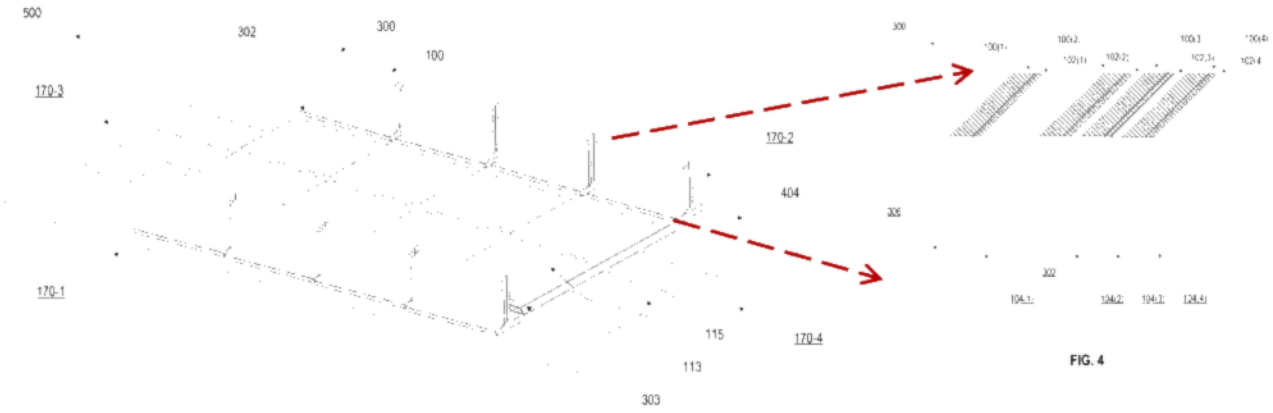
Source: Intel

EXHIBIT 24: **In a patent, Intel proposes to use 2 transistors to build a DRAM cell in BOEL.**



Source: Intel

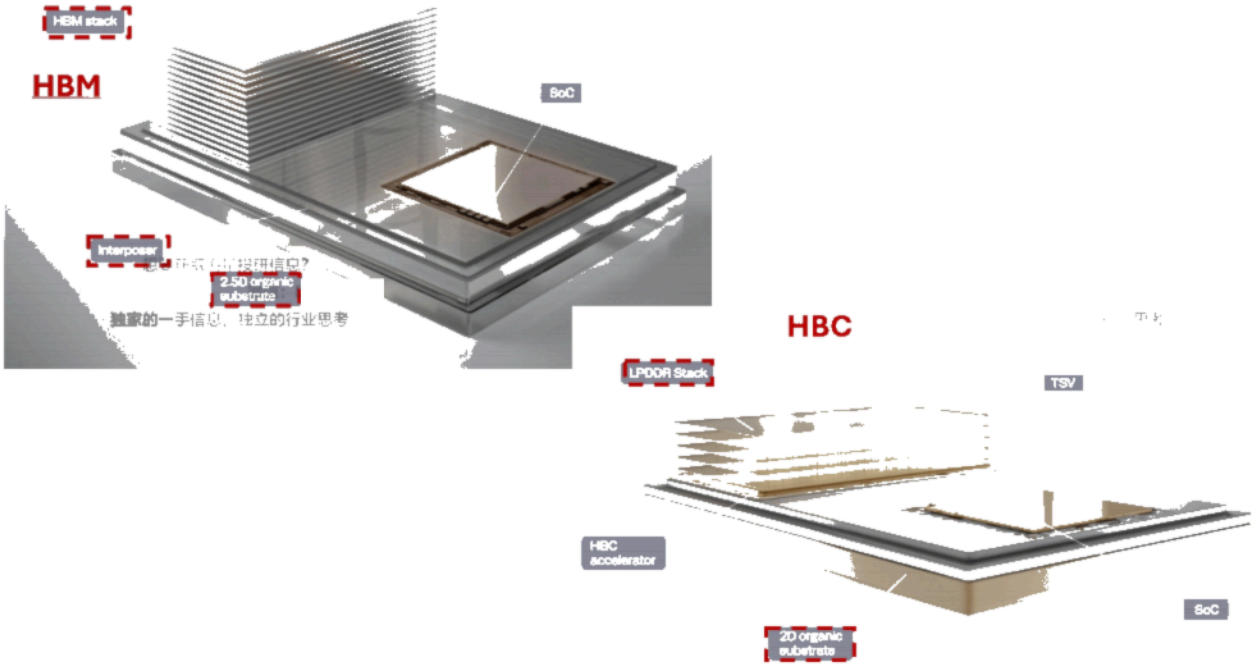
EXHIBIT 25: **Intel's ZAM reorients DRAM dies in a stack by 90 degrees to improve thermal dissipation.**



Source: Intel, Bernstein analysis

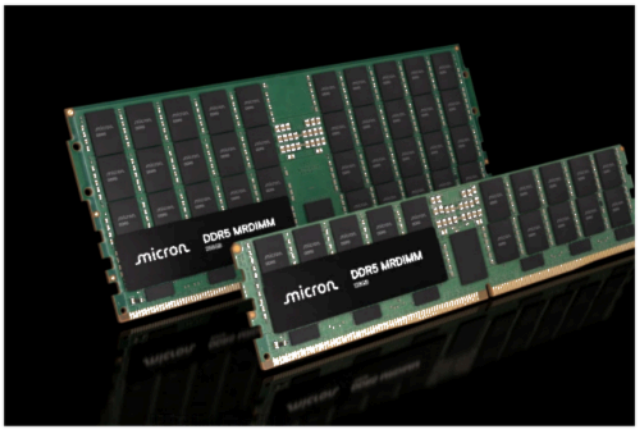
on 31-Aug-2026

EXHIBIT 26: **Using conventional packaging instead of CoWoS, Qualcomm's HBC compromises on performance but avoids the cost of interposer & 2.5D organic substrate.**



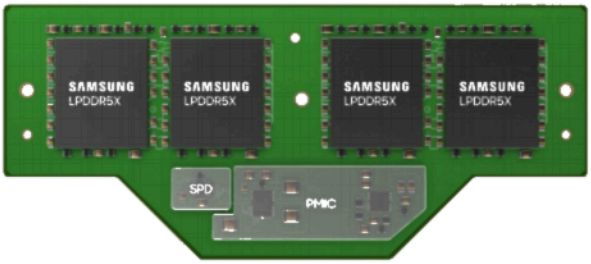
Source: Qualcomm

EXHIBIT 27: **RDIMM (Registered DIMM) & MRDIMM (Multiplexed Rank DIMM) are high-end variants of DIMM and are used in servers.**



Source: Micron

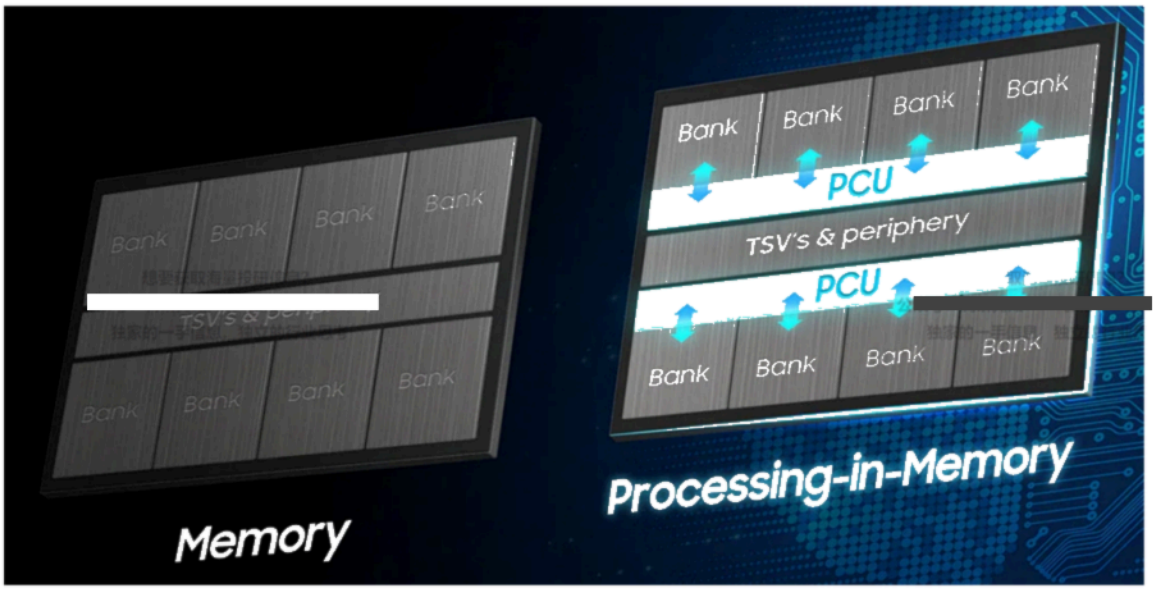
EXHIBIT 28: **LPCAMM2 puts LPDDR dies into a new module format.**



Source: Samsung

on 31-Aug-2026

EXHIBIT 29: **Samsung proposed a few implementations of PIM. In the case of combining PIM with HBM, Samsung proposed adding logic circuits (called PCU - programmable computing unit) inside the memory die.**



Source: Samsung

EXHIBIT 30: **Different memory technologies have different costs, access time (i.e. latency), etc.**

	SRAM (Static Random Access Memory)	DRAM	MRAM (Magnetic Random Access Memory)	PCM (Phase Change Memory; e.g. 3D XPoint)	RRAM (Resistive Random Access Memory)	NAND	HDD
Access Time	<10ns	~35ns	~10-50ns	>100ns	~100ns	~25-100us	~3ms
Volatile/Non-Volatile	Volatile	Volatile	Non-Volatile	Non-Volatile	Non-Volatile	Non-Volatile	Non-Volatile
Power Consumption	Varying	High	Medium/Low	Medium	Low	Low	High
Endurance (Cycles)	10^{15}	10^{15}	10^{9}-10^{12}	10^{6}-10^{7}	10^{5}-10^{6}	10^{5}	
Cell Size (F^2)	160-280	6	6-30	4/4 layers	6-30	4/hundreds of layers	

Source: Applied Materials, Toshiba, Everspin Yole, and Bernstein analysis

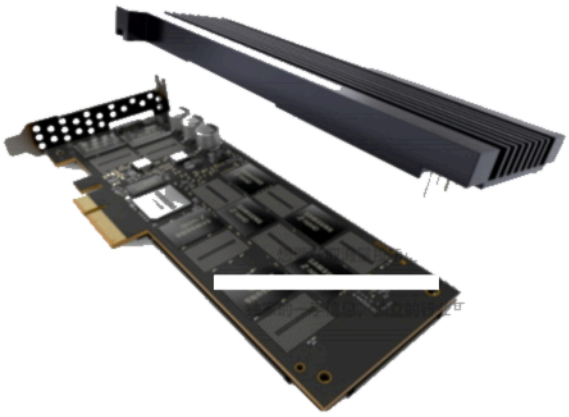
on 31-Aug-2026

EXHIBIT 31: **KIOXIA claims its GP-series SSDs achieving super high IOPS through leveraging XL-FLASH technology.**



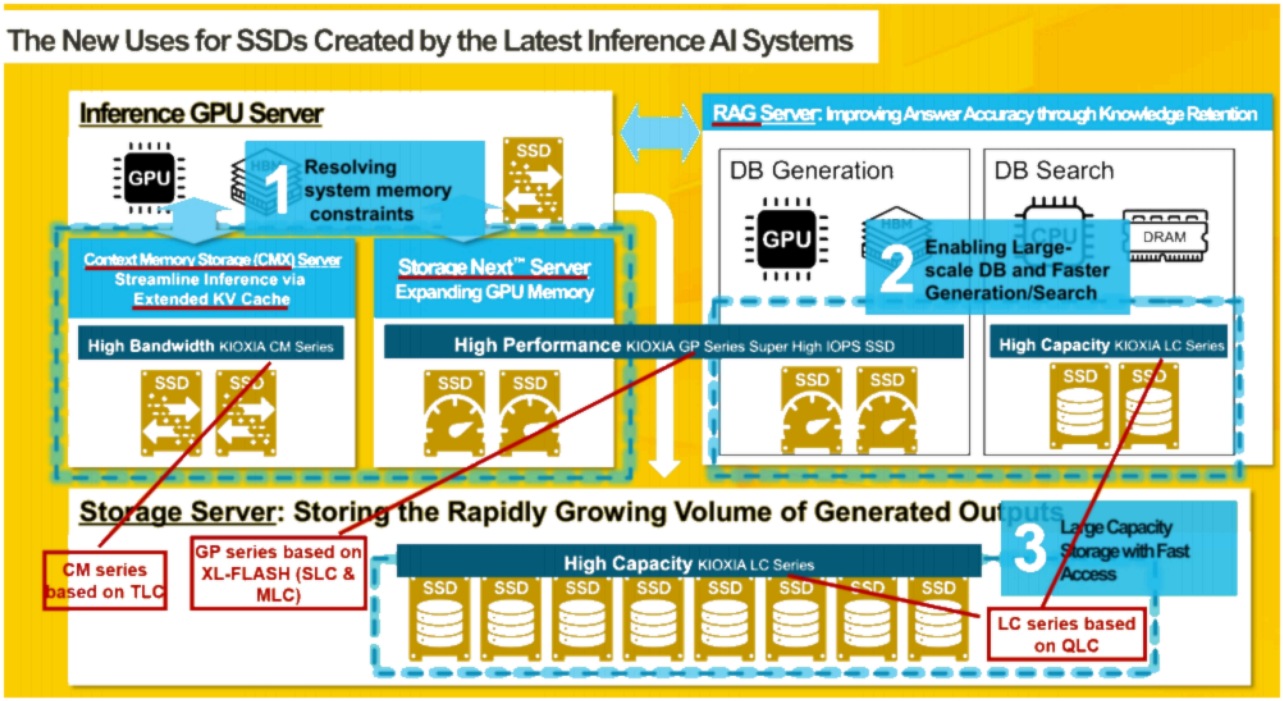
Source: KIOXIA

EXHIBIT 32: **Samsung takes advantage of Z-NAND in its Z-SSDs to achieve "ultra low latency".**



Source: Samsung

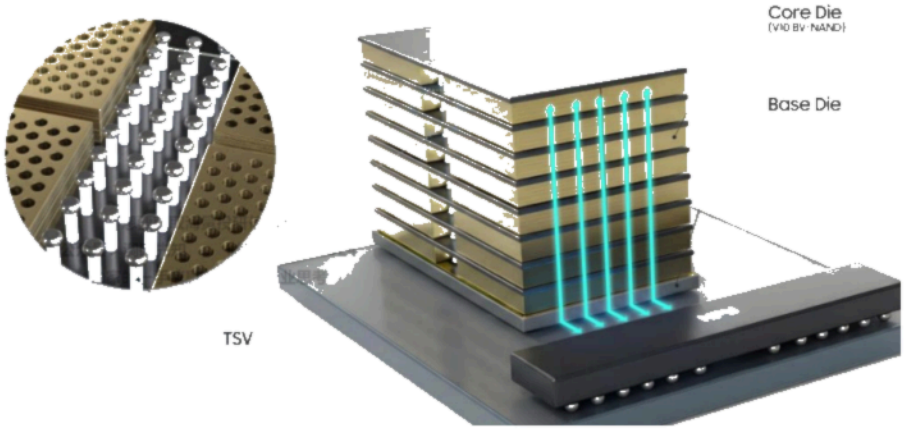
EXHIBIT 33: **KIOXIA incorporates different technologies into SSDs for specific parts of AI inference need.**



Source: KIOXIA, Bernstein analysis

on 31-Aug-2026

EXHIBIT 34: **Targeting on-device AI, Samsung's zNAND-O stacks NAND dies together and then packages the stack close to the processor.**



Source: Samsung