Morgan Stanley | RESEARCH

IDEA

October 5, 2026 06:27 AM GMT

Semiconductors | North America

How can the AI ecosystem work around memory bottlenecks?

We expect the memory shortage to be a constraint on AI builds through the duration of this AI cycle – but AI won't wait for DRAM fabs to be built. We look at ways to work around shortages.

Key Takeaways

- Intensity of memory shortages may fluctuate, but AI will basically use all supply for the foreseeable future
- CEO Jensen Huang has said that the industry needs to think differently to work around these shortages. We contemplate what he might be considering
- Among the potential beneficiaries would be MRVL and ALAB for CXL & larger scale-up, CBRS for disaggregation
- Remain OW MU and SNDK as we don't see the shortage ending

What was the catalyst for this note? In recent conversations with compute companies we have heard a great deal about the need to work around memory bottlenecks - NVIDIA CEO Jensen Huang among them. This was not a negative memory conversation - it's very clear that the industry anticipates multiple years of shortages.

But if memory supply cannot keep up with token growth, the industry will need to become more memory efficient. NVIDIA has talked about using its control over compute, networking, and storage to mitigate those bottlenecks.

In this note, we look at some of the ways that the industry might try to work around memory shortages, notably:

De-speccing, reducing the rack scale central memory (LPDDR5), and selectively reducing high bandwidth memory. This is in many ways the least favorable, but sometimes needed, approach. Our view is that there will continue to be products that maximize memory content, but that there will be selective de-speccing.

Potential beneficiaries: None directly, but allows compute companies to make their numbers - and shifts some of the bottleneck from DRAM to NAND. At the margin, de-speccing pushes for a larger scale up domain, which would help scale up plays such as ALAB and MRVL.

Disaggregation of AI workloads. This involves breaking up AI tasks into memory intensive and less memory intensive portions, where areas such as prefill which do not require as much memory.

We also note that as DRAM prices rise, some of the lower latency solutions from

Joseph Moore
Equity Analyst
Joseph.Moore@morganstanley.com +1 212 761-7516

Ella Tulchinsky
Research Associate
Ella.Tulchinsky@morganstanley.com +1 212 761-2222

Mason Wayne
Research Associate
Mason.Wayne@morganstanley.com +1 212 761-6012

Cate Folan
Research Associate
Cate.M.Folan@morganstanley.com +1 212 296-3520

Nicole Kozhukhov
Research Associate
Nicole.Kozhukhov@morganstanley.com +1 212 761-1636

Shane Brett
Equity Analyst
Shane.Brett@morganstanley.com +1 212 761-1022

SEMICONDUCTORS
North America
Industry View **Attractive**

investors should be aware that the firm may have a conflict of

Research as only a single factor in making their investment decision.

For analyst certification and other important disclosures, refer to the Disclosure Section, located at the end of this report.

companies such as Groq (now within NVIDIA) or Cerebras or others use on chip SRAM, which has historically been much more expensive than DRAM, but currently is not.

Potential beneficiaries: compute solutions that can be a factor in disaggregated workloads, such as Cerebras, and NVIDIA through Groq

CXL memory centralization - CXL is a switched access to DRAM which allows for substantial efficiency advantages to centralized memory "pooling", allowing several processors to access the same memory. This has generally been considered a general purpose compute technology, but recent comments from chip suppliers are pointing to large opportunities within AI as well.

Potential beneficiaries: ALAB, MRVL.

Isn't this all bad for DRAM? In terms of near term earnings maximization, yes, a little. An environment in which the AI ecosystem comes to a halt and stops building because of the DRAM shortages probably leads to the best near term pricing outcomes.

But that was never realistic, and we have consistently argued for duration, not amplitude, as the primary driver of the cycle from here. And we would make the argument that reducing memory specs because of supply constraints - importantly, not because of pricing - leads to a situation where more supply will be very easily absorbed as we respec to higher levels.

While we don't take this for granted - we do extensive monthly checks into memory conditions at this point - our working hypothesis is that because of this pent-up demand, the risk to the memory cycle is the same as the risk to the compute space, that is a deceleration in AI itself. (We should introduce one risk factor here - that if AI slows, not because of demand, but because of land/power/shell bottlenecks, such a pause could cause memory disruption that we would not see with the compute names.

For that reason, we continue to be positive on MU and SNDK in our coverage.

The Brute Force reduction in memory specifications

Summary: Memory is a key aspect to compute performance, and generally there is reluctance to lower specifications. But the world is not going to wait for DRAM fabs to get built, so there are optimizations around lower memory levels. We would note that so far every reduction in specifications causes memory bottlenecks to surface in other parts of the compute spectrum.

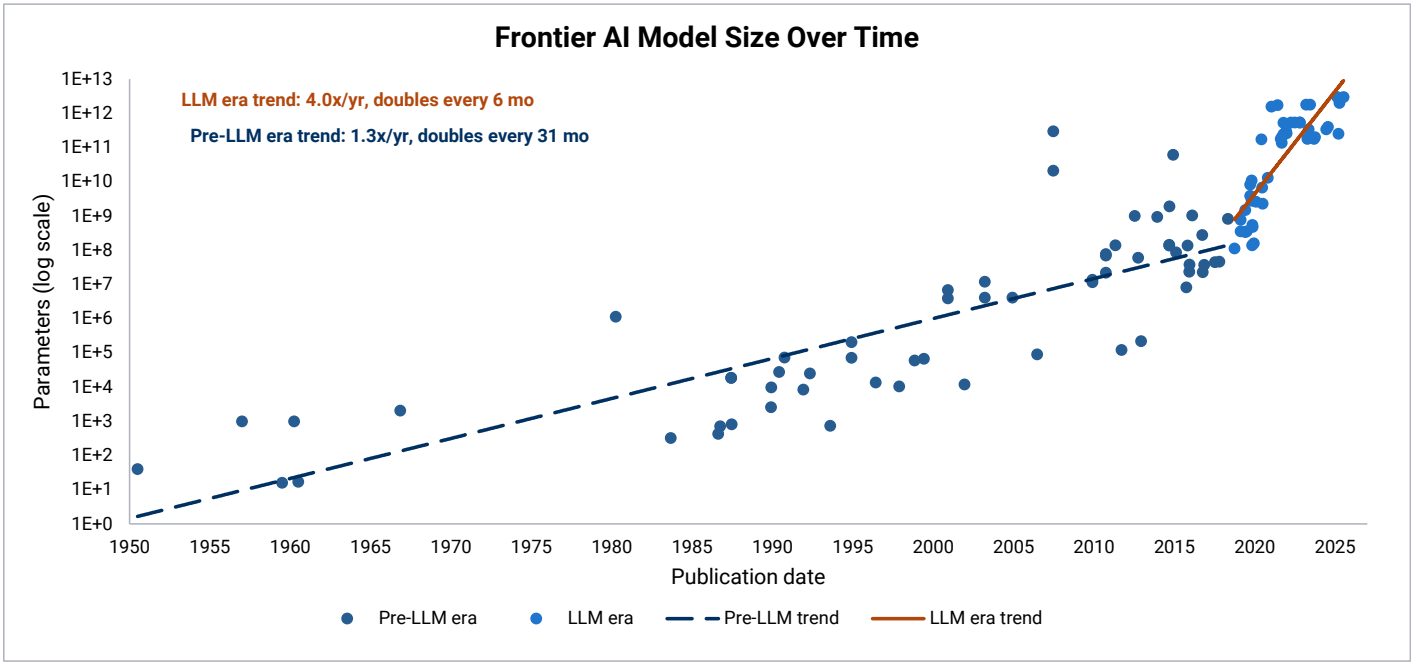
We are seeing Nvidia offer reduced memory content skews for rubin, including both HBM and main memory, in an attempt to offset rising prices and enable continued unit growth but that's less of a negative for memory companies than it appears. Increasing memory demand for AI a secular trend, and reducing memory content at one level of the memory hierarchy puts incremental pressure on the remaining tiers, shifting the memory bottleneck more than eliminating it.

What's happening with de-speccing today? The current shortages and rising prices of both DRAM and NAND has put significant pressure on memory buyers, and in response companies throughout the ecosystem are looking for ways to help ease those constraints. The most straight forward response, especially for DRAM, is to reduce the amount of memory per device or per server. We're seeing this happen out of necessity in most cases, where reducing the amount of DRAM is a measure taken to enable higher unit shipments. We believe Nvidia is offering additional rubin skews to customers, with LPDDR5 per rack content of 28TB vs the original 54TBs by asking DRAM vendors to supply 96GB SOCAMM2 modules vs 192GBs. For HBM Rubin was originally expected to be 288GB of HBM per GPU (8 stacks of 12hi HBM4), but we now expect skews with 192GB (8 stacks of 8hi HBM4) to be offered. For Rubin ultra content was originally debuted at 1TB HBM4e per 4 die accelerator, the move to keep rubin ultra at a 2 die design makes 512GB the relevant comparison vs the originally announced specs. Versus that 512 we could see Nvidia stay at HBM4 rather than move to 4e, as well as offer 8hi vs 12hi alternatives lowering overall capacity to the 192GB-384GB range.

What drives the need for more memory capacity in AI workloads? Before understanding what's lost when memory capacity is reduced, we think its important to look at why there has been a persistent push to add more memory over time. We highlight three factors below:

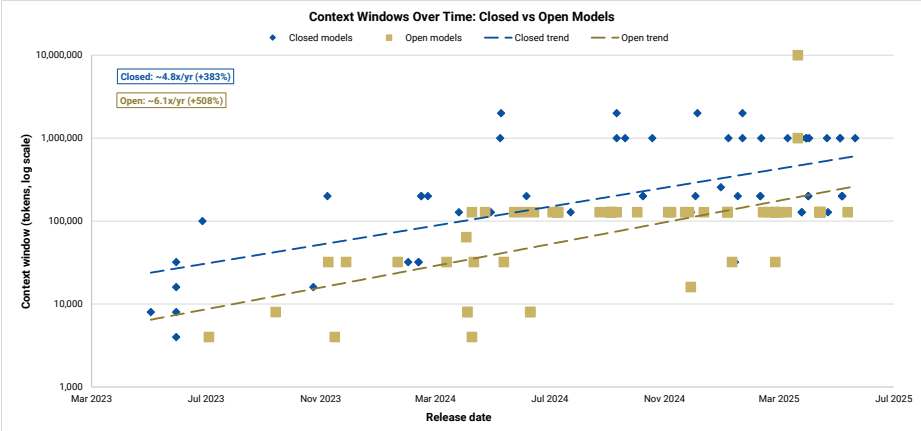
Model size: Over decades of machine learning history there has been a consistent desire for larger and larger models, as research has consistently shown the larger the model the better its performance. Model weights are almost always stored in HBM, increasing the need for larger HBM capacity over time.

Exhibit 1: Frontier AI model sizes are doubling every 6 months



Context length: LLMs can only take into account a certain number of tokens at one time, including the prompt, uploaded documents, conversation history, and the model output. That limit is called the context length or context window, with most leading models today at about 1M tokens. As AI use cases increasingly revolve around long horizon agentic tasks the context window is a critical bottleneck for the amount of work an agent can accomplish.

Exhibit 2: LLM context windows are going ~5x annually

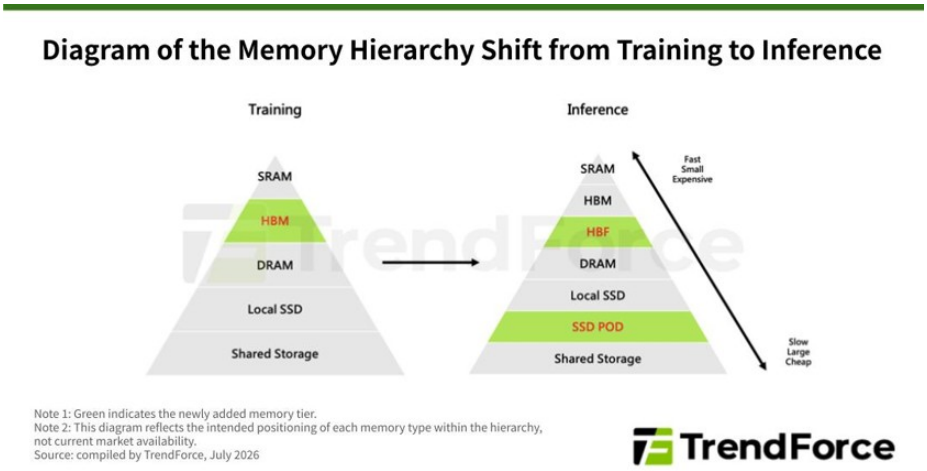


Inference concurrency: The number of requests an AI system serves at the same time. LLM inference has two steps, prefill and decode. Prefill is usually limited by compute, while decode is usually limited by memory. More simultaneous requests require more memory capacity, because each session stores its own context (the KV cache), and more memory bandwidth, because every decode step needs to process data from the KV cache.

We expect over time LLMs will become smarter (model size), need more in-context information (context length), and serve more users/agents over time (inference concurrency). That means de-speccing memory is likely to only prove temporary as each of those vectors continue to require greater amounts of memory across the hierarchy.

When capacity is reduced at one level of the memory hierarchy it puts pressure on the remaining tiers, de-speccing HBM/LPDDR is creating opportunities for networked DRAM (CXL) and NAND. Given the factors outlined above creating incremental demand for memory capacity for AI, when one tier is reduced (i.e HBM) the data needs to move elsewhere creating new pain points in memory and IO. We are seeing this in real time, as Nvidia moves to de-spec LPDDR5 racks we are seeing incremental demand for NAND as KV cache storage demands increase for that level of the memory tier. We are also seeing lower DRAM content putting increasing weight on networking, as bits that were stored in HBM need to be spread across more GPUs (putting pressure on scale-up bandwidth), or read from main memory (putting pressure on CPU IO).

Exhibit 3: De-speccing HBM/LPDDR should open up opportunities lower in the memory hierarchy



There will also be opportunities that emerge for new tiers in the memory hierarchy; we discuss CXL and SRAM based architectures below but going forward, optically connected memory is an interesting avenue for the industry as well as new technologies such as HBF.

Is cutting HBM capacity a "free lunch" for bandwidth bound workloads? If HBM content is being reduced by lowering HBM stack heights, like the rubin de-spec (moving from 12hi to 8hi), bandwidth isn't affected, as HBM bandwidth is function of the interface and pin speeds. That can work well for small context and smaller models, where there is lower capacity requirements but it creates risks. If trends outlined above (long context, agentic workloads) continue the number of workloads that fall into
to shrink, requiring more GPUs and/or increasing latency as more da
lower memory tiers. That said bandwidth is a key bottleneck for LL
lowering stack heights is a straight forward way to reduce costs. Go
watch if HBM vendors can charge a premium for higher bandwidth a
progress - that has been true historically. However, as custom base
the driver of HBM bandwidth (the DRAM PHY) can be designed by t



which may shift the value away from DRAM vendors.

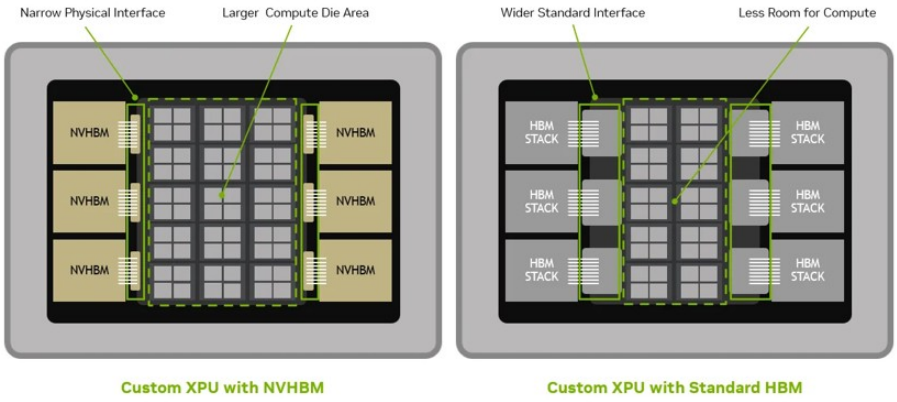
The scale-up domain is also a factor to consider with Rubin Ultra and beyond. An effective scale-up mechanism effectively creates a rack that acts as a single GPU, by using a multiplexed switching mechanism to connect all to all within the rack. We don't know what Rubin Ultra looks like at this point - the company has said they have altered the Kyber form factor and will have a higher scale up domain - ie more GPUs per rack - and this would allow the total memory content at the rack level to stay constant with less memory per GPU.

Exhibit 4: Nvidia custom HBM4e offers 30% more bandwidth vs JDEC standard HBM4e

Feature	NVHBM **benefit**
Bandwidth	Up to 30% more memory bandwidth compared with standard HBM4e
Area	More efficient interface connections allow up to 25% more compute die area for additional XPU capabilities
Power	Up to 15% lower HBM power usage compared with standard HBM4e adds up savings across thousands of XPUs

Table 1. NVHBM brings three main platform-level advantages to AI accelerator programs: higher memory bandwidth, more package and silicon area, and lower HBM power usage

Exhibit 5: Nvidia custom HBM reduces PHY and support area 67% vs JDEC standard HBM4e



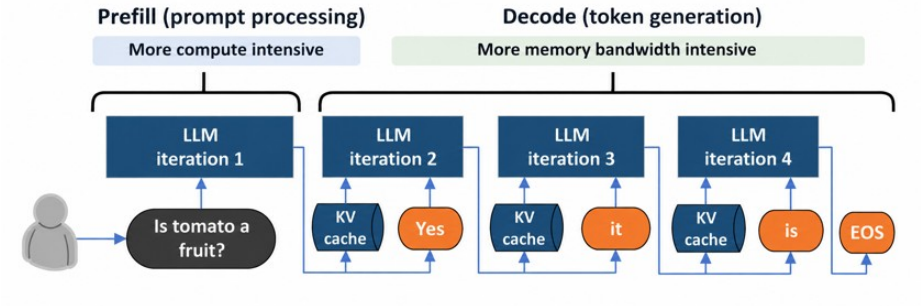
Not every step requires so much memory: Disaggregation

The opportunity in disaggregation comes from optimizing hardware around the different requirements of each stage of inference. Not all parts of inference place the same demands on compute and memory. Prefill processes the input prompt in parallel and is relatively compute intensive. Decode generates output one token at a time and repeatedly accesses model weights and the growing KV cache, making memory bandwidth and capacity more important. The same accelerator is therefore being asked to handle two workloads with very different resource requirements.

Disaggregation separates these stages so each can run on hardware better suited to the task. Compute dense systems can handle prefill, while architectures optimized for memory bandwidth and low latency can handle decode. This allows operators to scale the two resources independently and improve utilization, rather than provisioning the same mix of compute and memory for every stage of inference.

From a memory perspective, disaggregation allows the industry to use different types of memory where they are most valuable. Memory bandwidth intensive portions of inference can move to architectures built around high bandwidth SRAM or other approaches that reduce data movement, while HBM and larger capacity memory can remain focused on workloads that need their capacity and flexibility. Disaggregation does not eliminate memory demand, but heterogeneous architectures can improve how efficiently scarce and expensive memory is used as inference scales.

Exhibit 6: AI inference consists of two distinct stages: prefill and decode



Why now? The memory demands of inference are increasing as models process longer context and generate more output tokens. Reasoning models, coding agents and other agentic applications can require substantially more inference time than traditional question and answer workloads, increasing the importance of both memory capacity and bandwidth. At the same time, faster interconnects and improvements in inference software are making it more practical to move the KV cache between separate prefill and decode engines, helping address one of the main implementation challenges of disaggregation.

Interest in disaggregated inference accelerated sharply around the end of 2025. NVIDIA acquired Groq around this time, AWS subsequently announced a disaggregated architecture pairing Trainium for prefill with Cerebras for decode over EFA, and Cerebras has since announced a similar approach with AMD. Meanwhile, d Matrix has moved Corsair

into volume production. Taken together, these developments point to a broader move toward heterogeneous inference, where different processors are optimized for different parts of the workload.

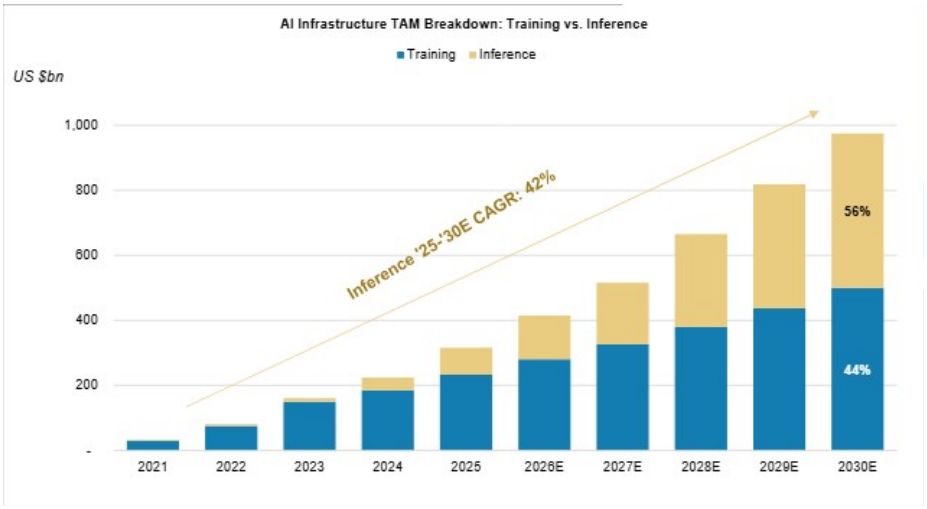
The emerging solutions take different approaches to the memory problem, ranging from replacing external memory with high bandwidth SRAM to combining SRAM, HBM and DDR within the same inference system:

- **CBRS:** Cerebras uses its wafer scale processor to combine a large amount of SRAM and compute on the same piece of silicon, reducing the need to move data between processors and external memory. The architecture provides very high memory bandwidth, which is particularly useful during decode when model weights must be accessed repeatedly for each generated token. Cerebras can run both prefill and decode, while its disaggregated architectures with AMD and AWS use GPUs or Trainium for the more compute intensive prefill stage and Cerebras for the memory bandwidth intensive decode stage.
- **NVIDIA Groq:** Groq's LPU uses high bandwidth on chip SRAM and a compiler scheduled architecture designed for low latency inference. NVIDIA is integrating Groq technology alongside its Vera Rubin platform, combining HBM based GPUs with SRAM based LPUs. Rubin GPUs handle prefill and the attention portion of decode, including workloads that require large KV caches, while Groq handles the feed forward and mixture of experts portions of decode. NVIDIA Dynamo coordinates the workload and transfers activations between the two systems.
- **D Matrix:** d Matrix uses a digital in memory compute architecture that places compute directly alongside high bandwidth on chip SRAM, allowing model weights to be processed where they are stored rather than repeatedly moved between separate memory and compute resources. This is designed to reduce the data movement bottleneck that becomes particularly important during decode, where model weights are accessed repeatedly for each generated token. d Matrix combines this high bandwidth local memory with larger capacity memory for data that does not need to remain on chip, and is positioning its accelerators alongside GPUs in heterogeneous systems where GPUs handle prefill and d Matrix handles decode.
- **SambaNova:** SambaNova uses a reconfigurable dataflow architecture that maps the model's computation directly onto the processor, allowing data to flow from one operation to the next without repeatedly returning to memory between each step. The architecture is paired with a three tier memory hierarchy spanning on chip SRAM, HBM and DDR, with the fastest memory used for the hottest local data and larger memory tiers providing capacity for model weights, KV cache and other less latency sensitive data. SambaNova is using this architecture in heterogeneous inference systems where GPUs handle prefill, its RDUs handle decode and CPUs manage orchestration and agentic tool execution.

We expect inference spending to grow faster than training as AI usage expands and each interaction consumes more compute. Disaggregated inference should be particularly relevant for workloads where prefill and decode place very different demands on the underlying hardware. Coding assistants and agentic applications are good examples: they can ingest large amounts of context, such as codebases, documents or prior interactions, and then generate substantial output across repeated model calls. This

creates a mix of compute intensive prefill and memory bandwidth intensive decode, making it more attractive to separate the two stages and scale each independently. Long context workloads can also benefit because a large prefill request can otherwise interfere with token generation for other users. As these workloads become a larger share of inference, disaggregation provides another way for the ecosystem to scale AI output without requiring every stage to be provisioned with the same mix of compute and high performance memory.

Exhibit 7: Total AI infrastructure spend forecast



Potential Stock Beneficiaries: CBRS, NVDA

We see Cerebras as one of the clearest pure play beneficiaries of greater adoption of disaggregated inference. The company is already working with both AMD and AWS on systems that pair third party processors for prefill with Cerebras for the memory bandwidth intensive decode stage. With AMD, Helios handles prompt processing and large context windows while the Cerebras Wafer Scale Engine handles decode, with the solution expected to enter production in 4Q26. AWS is taking a similar approach with Trainium for prefill and Cerebras for decode, with the combined solution expected to reach Amazon Bedrock in 1Q27. These partnerships give Cerebras a path to participate in heterogeneous systems even where another accelerator remains the primary compute engine

Disaggregation can also materially improve the economics of Cerebras Cloud. Cerebras and AMD have disclosed that their combined system can increase throughput by up to 5x while maintaining Cerebras inference speed, allowing each Cerebras system to produce substantially more tokens from the same underlying Cerebras capacity. Because Cerebras operates its own cloud infrastructure, lower cost per token can accrue directly to the company through higher gross margin, lower customer pricing, or some combination of the two. This makes disaggregation more than an incremental hardware opportunity for Cerebras, it can increase the amount of revenue generated from each deployed system while lowering the cost of producing each token.

Other potential beneficiaries include compute solutions that can participate in disaggregated inference workloads. NVIDIA is incorporating Groq technology into Vera

Rubin, pairing HBM based GPUs with SRAM based LPUs for different portions of inference. As disaggregated inference becomes more common, we see an opportunity for a broader set of processors to participate in AI infrastructure by specializing around the portions of the workload where their architecture is most efficient.

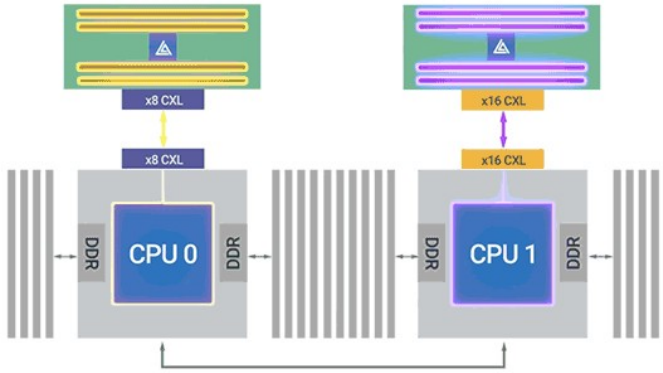
CXL: Using networking/interconnect technologies to optimize memory usage

The CXL Opportunity

CXL, or Compute Express Link, is a high speed interconnect designed to give processors more flexible access to memory. It uses the same underlying physical infrastructure as PCIe, but adds specialized memory and cache protocols that allow external memory to behave much more like memory directly attached to the processor. The key benefit is that memory no longer needs to be rigidly tied to a specific CPU or server, allowing data center operators to use available DRAM more efficiently and add capacity without necessarily adding more compute.

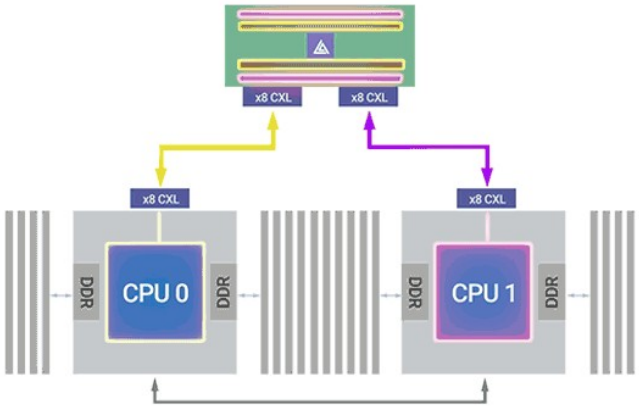
We see three main use cases: **Expansion** allows a server to add memory beyond the capacity supported by its local memory channels. **Sharing** allows multiple processors to access the same memory resources, reducing duplication and improving overall memory utilization. **Pooling** creates a common pool of memory that can be dynamically allocated across multiple processors or servers, reducing stranded capacity and overprovisioning.

Exhibit 8: Memory Expansion: Adds additional memory capacity to a single processor beyond its locally attached DRAM.



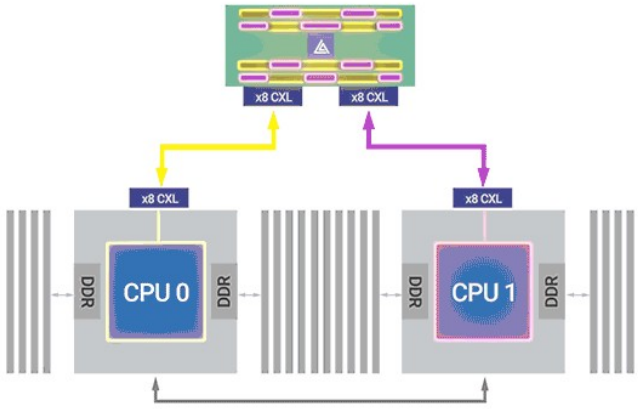
Source: Astera Labs

Exhibit 9: Memory Sharing: Allows multiple processors to access the same memory resource and data.



Source: Astera Labs

Exhibit 10: Memory Pooling: Creates a common memory pool that can be dynamically allocated across multiple processors.



Source: Astera Labs

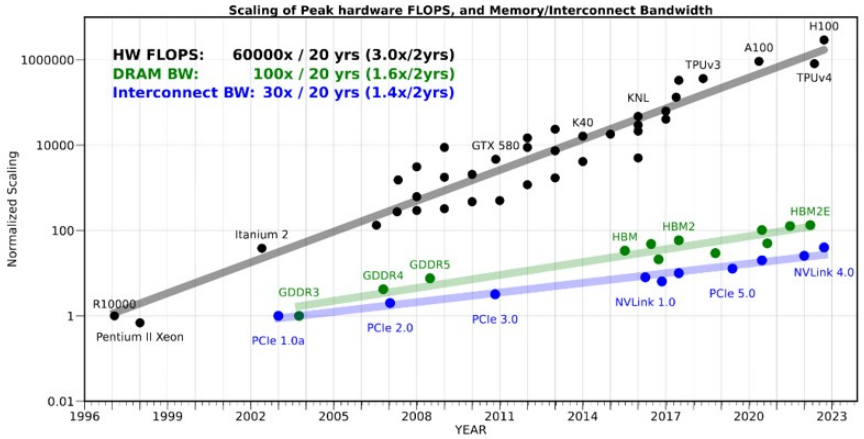
Why CXL for AI?

In the past, CXL has been used primarily for traditional CPU based compute, where workloads often benefit more from additional memory capacity than from optimizing for the absolute lowest latency and highest bandwidth. CXL introduces an additional hop between the processor and memory, making CXL attached DRAM slower than locally attached DRAM. **This has historically limited CXL's role in the most latency and bandwidth sensitive AI workloads, where GPUs rely on HBM for extremely high bandwidth**.

The broader backdrop is that compute performance has scaled much faster than memory and interconnect bandwidth, as shown in Exhibit 11. This widening gap makes memory placement and utilization increasingly important to overall system performance. It also increases the value of a tiered memory architecture, where HBM is reserved for the most bandwidth sensitive data while larger pools of DRAM can support workloads that are more capacity constrained.

More recently, CXL has attracted growing interest for AI inference as longer context windows, agentic workloads, and growing KV caches create much larger memory capacity requirements, while much of that data does not necessarily need to reside in HBM at all times. **CXL attached DRAM can therefore serve as a lower cost, higher capacity memory tier, with the most performance sensitive data remaining close to the accelerator and less sensitive data moved into external memory.**

Exhibit 11:



Source: "AI and Memory Wall," Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, Kurt Keutzer.

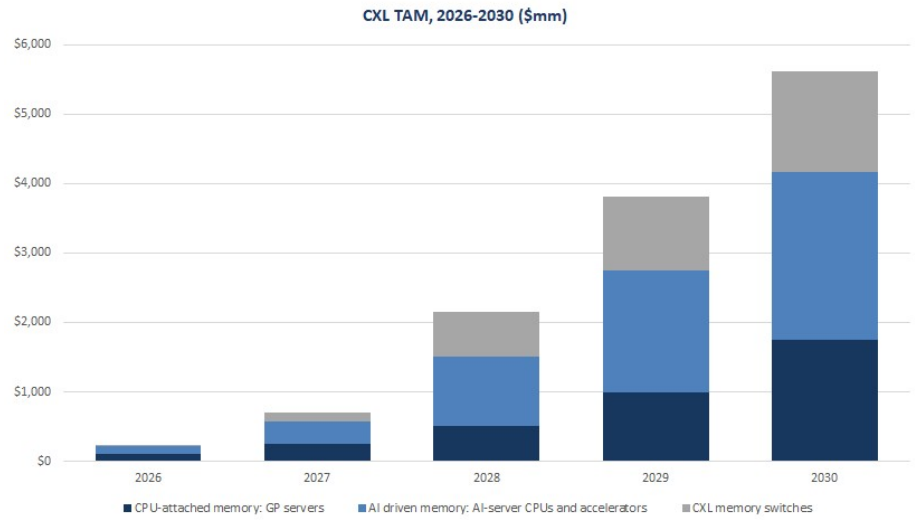
CXL TAM broadens as memory constraints intensify

CXL has typically been sized as a CPU-attached memory expansion market. Astera, for example, has previously framed the CXL memory controller opportunity at over $4bn based on the volume of CXL-capable server CPUs, a market addressed by products such as Astera's Leo and Marvell's Structera X. While adoption has been slower than initially thought, we see conditions improving. Higher DRAM pricing and tighter supply are pushing hyperscalers to add capacity more efficiently and reuse memory they have already deployed. Astera has pointed to renewed customer interest on the back of these memory supply dynamics.

We now see AI expanding the opportunity to roughly $6bn by 2030, largely by creating new places to deploy CXL in AI systems. Within AI servers, host-attached CXL memory adds capacity for inference and agentic workloads. Beyond the host, suppliers are also building memory tiers that accelerators can reach directly over a fabric. Astera's Leo X targets this accelerator-side memory demand, and Astera has a custom KV-cache offload design win that ramps in 2027. As these tiers scale, sharing and pooling memory across larger groups of CPUs and accelerators also adds CXL switching content, a use case Marvell's Structera S targets at rack scale.

Supplier commentary supports this direction. Marvell has framed CXL as a >$1bn revenue contributor around CY2028 within its XPU-attach business, a significant portion of our 2028 estimate. We therefore see the broader CXL silicon opportunity expanding from a > $4bn memory controller market toward a roughly $6bn opportunity for total CXL and related memory-attach semi content by 2030, as AI drives higher memory requirements, greater CXL content per system and increasing adoption of pooled memory architectures.

Exhibit 12:



Sizing this market with precision is difficult. Many of the AI use cases are still in early deployment, the boundaries between CPU-attached, accelerator-attached and pooled memory are still being defined, and hyperscalers are pursuing different architectures, including some built on non-CXL fabrics, larger HBM stacks or flash-based context tiers. Both the mix across use cases and the timing of adoption could look meaningfully different from our assumptions. But our central view is that AI has expanded the opportunity materially beyond the >$4bn controller market that has framed CXL to date, and we expect to refine it as deployments scale.

Potential Stock Beneficiaries: ALAB and MRVL

Astera Labs. We see Astera as one of the most direct public beneficiaries through its Leo memory controller family. Management expects both standard and custom Leo products to ramp in volume in 2027 across two US hyperscalers, spanning traditional compute and AI inference, including a custom KV cache offload design. We do not fully reflect a broad CXL adoption cycle in our Astera estimates today, making a stronger Leo ramp largely incremental to our model.

Marvell. Marvell has a broad CXL portfolio spanning memory expansion through Structera X and rack level pooling through Structera S. Management has previously framed CXL as a greater than $1bn revenue contributor around CY2028, while more recent commentary points to additional design wins and stronger demand as inference and memory constraints increase. Some of this opportunity is already contemplated in Marvell's longer term framework, but continued expansion in CXL could provide additional upside. We expect to learn more at Marvell's upcoming Investor Day, where we expect more detail on its memory expansion opportunity.

The content addressing private companies is being provided for informational purposes only and does not constitute a solicitation or imply future research coverage if the company goes public. Content is based on unaudited information. No investment recommendation is provided as there is limited public information available for private companies. Investors should conduct their own due diligence and be aware that additional or different

information may be provided by the company.