MemVerge

Memory Machine™ X

Big Memory Software for the AI Era

Server Memory Expansion

Tiering enables up to 32TB capacity at half the cost.

Increases GPU utilization by 77%.

Fabric-Attached Memory

Multiple servers can share CXL memory.

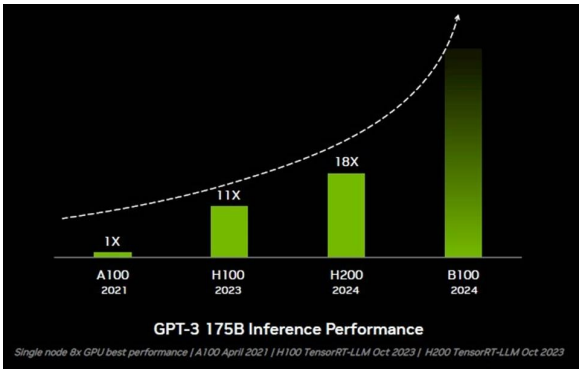
Maintains cache coherency.

Big Memory Computing in the AI Era

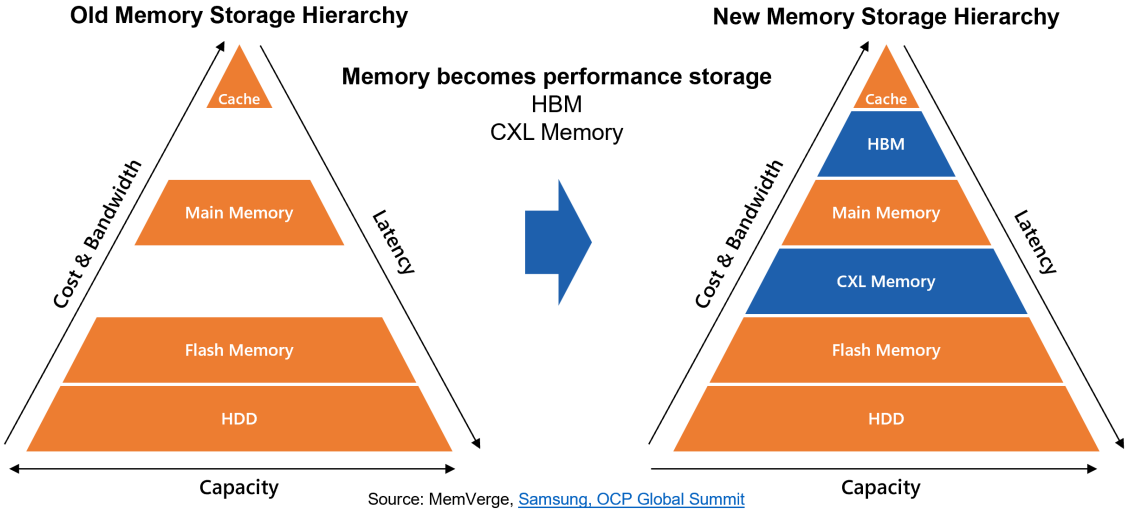
In the fast-evolving landscape of artificial intelligence (AI), where models are growing larger and more complex by the day, the demand for efficient processing of vast amounts of data has ushered in a new era of computing infrastructure. With the advent of transformer models, Large Language Models (LLM), and generative AI, the reliance on matrix computation across extensive tensor datasets has become paramount. This shift has propelled GPUs and other AI processors into the spotlight, as they boast increasing power to handle large-scale matrix multiplications efficiently.

However, as AI models balloon in size, a critical bottleneck emerges: the limitation of high-bandwidth memory (HBM) accessible to these processors and the constrained bandwidth of the interconnecting fabric between them. Today's mammoth models, often comprising hundreds of billions or even trillions of weights, demand an exorbitant amount of memory – ranging from hundreds of gigabytes to terabytes – for both training and inference tasks. Despite the exponential growth in GPU processing power, the rate of expansion in high-bandwidth memory on GPUs has been comparatively modest, leaving many models unable to fit entirely within the memory of a single GPU. This problem is commonly referred to as the "memory wall".

"...it looks to our eye like we can expect a lot more inference performance, and we strongly suspect this will be a breakthrough in memory, not compute, as the chart below suggests that was meant to illustrate the performance jump..." – The Next Platform



To address this challenge, the concept of Big Memory Computing has come to the fore. Big Memory Computing encompasses a suite of technologies aimed at scaling memory both vertically and horizontally, thus expanding the capacity of memory systems to accommodate the burgeoning needs of AI workloads. Horizontal scaling involves distributing tasks across multiple GPUs, necessitating various parallelism techniques to shard data between them. However, this approach often incurs significant data transfer overhead between GPUs, leading to suboptimal GPU utilization and slower overall performance.

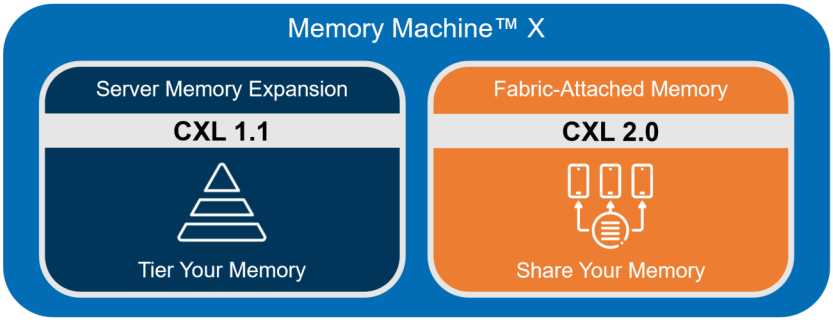


Source: MemVerge, Samsung, OCP Global Summit

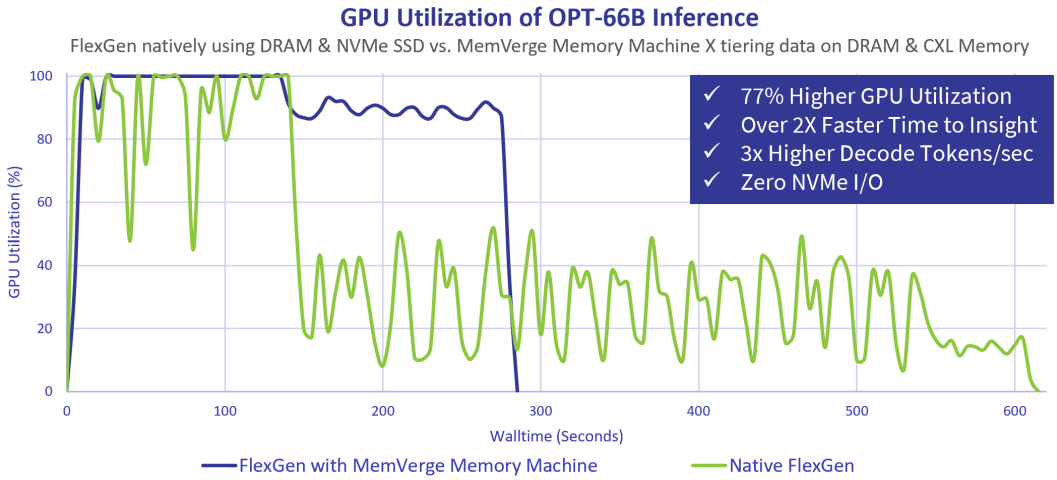
Conversely, vertical scaling focuses on extending memory capacity by introducing additional tiers of memory, such as main system memory or CXL memory, to complement the High Bandwidth Memory on GPUs. While less prevalent, vertical scaling holds promise, especially with ongoing software and hardware innovations aimed at enhancing its performance and feasibility.

Central to the realization of Big Memory Computing is the intelligent optimization of the memory-storage hierarchy within AI infrastructure. Technologies such as shared memory enable multiple processors on different server nodes to access the same memory region, facilitating memory sharing and reducing communication overhead for horizontal scaling of GPU memory. Furthermore, intelligent memory tiering and offloading, along with multi-node memory sharing technologies, are poised to play a pivotal role in maximizing GPU utilization and enhancing overall system performance.

A company at the forefront of Big Memory Computing is MemVerge, which has dedicated six years to developing cutting-edge technologies in this domain. Our flagship product, Memory Machine X, incorporates best-of-breed memory tiering and memory sharing technologies, offering a robust solution for AI use cases.



One example is a groundbreaking joint solution from MemVerge and Micron that leverages intelligent tiering of CXL memory, boosting the performance of large language models (LLMs) by offloading from GPU memory to CXL memory. The chart below shows the use of intelligently tiered memory increases the utilization of precious GPU resources by 77% while more than doubling the speed of OPT-66B batch inference.



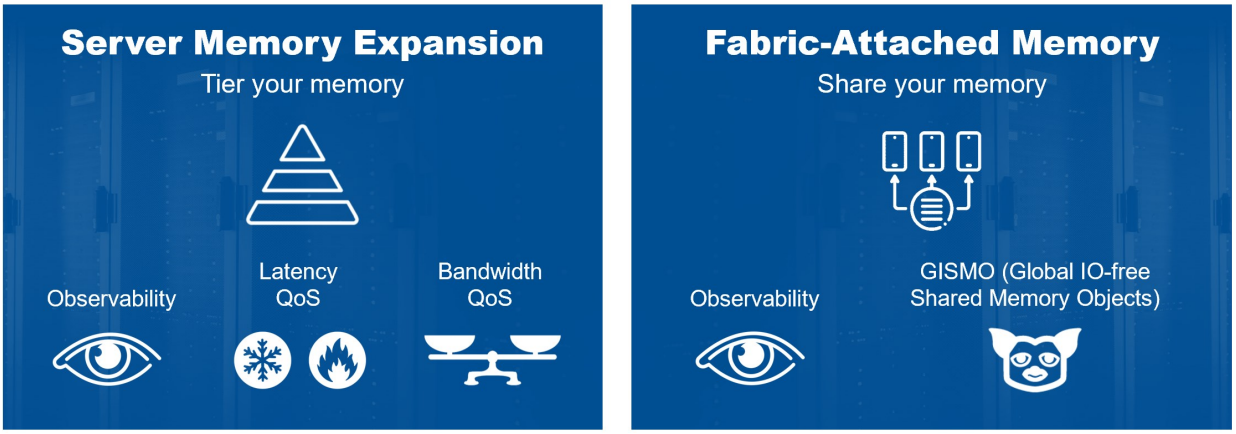
In summary, the growth in size and complexity of AI models is driving the need for innovative solutions to overcome the memory wall; Big Memory Computing answers the call with solutions that unlock the full potential of AI with vertical and horizontal memory scaling, coupled with intelligent optimization techniques; and MemVerge stands ready to leverage its expertise in Big Memory Computing and CXL to help prospective clients tackle the intricate challenges presented by the era of AI.

Introducing Memory Machine X

Big Memory Computing software that optimizes the cost and performance of AI and other data-intensive workloads by intelligently managing the memory-storage hierarchy and fabric-attached memory to overcome the memory wall.

Memory Machine X consists of 2 modules: Server Memory Expansion that allows you to tier your DRAM and CXL memory in a server, and Fabric-Attached Memory that allows you share the CXL memory in a fabric-attached memory system.

Memory Machine X

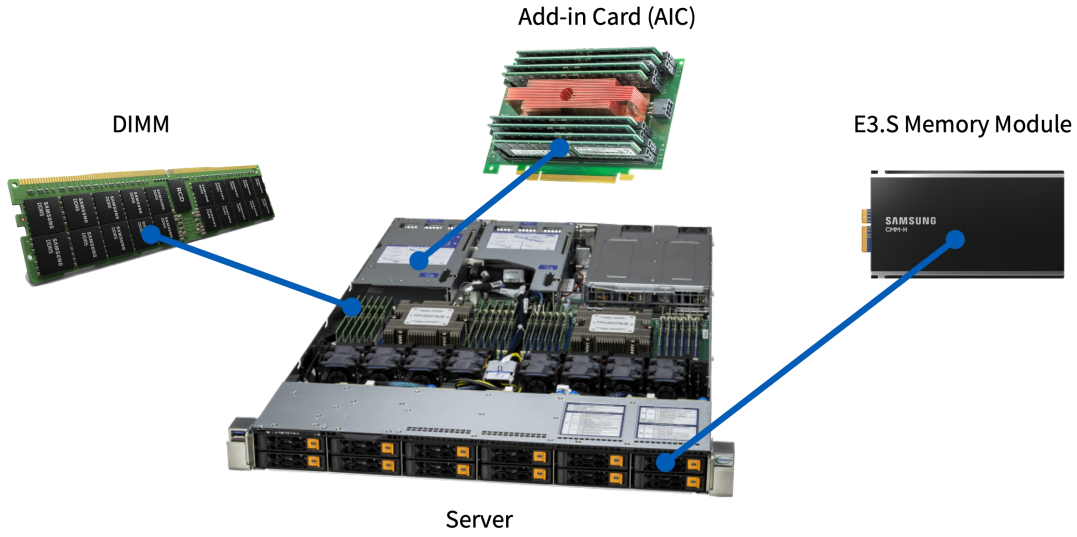


Server Memory Expansion

Server hardware: You can now expand memory 3 different ways

With support for Compute Express Link (CXL) 1.1, servers offer a new architectural model supporting 2 new products for dramatically lower cost and radically greater capacity: CXL Memory Add-in Cards (AICs) and E3.S memory modules. The availability of new classes of memory creates the need for software that provides transparent access to mixed memory and to automatically place hot and cold data in the right tier.

New Server Expansion Model



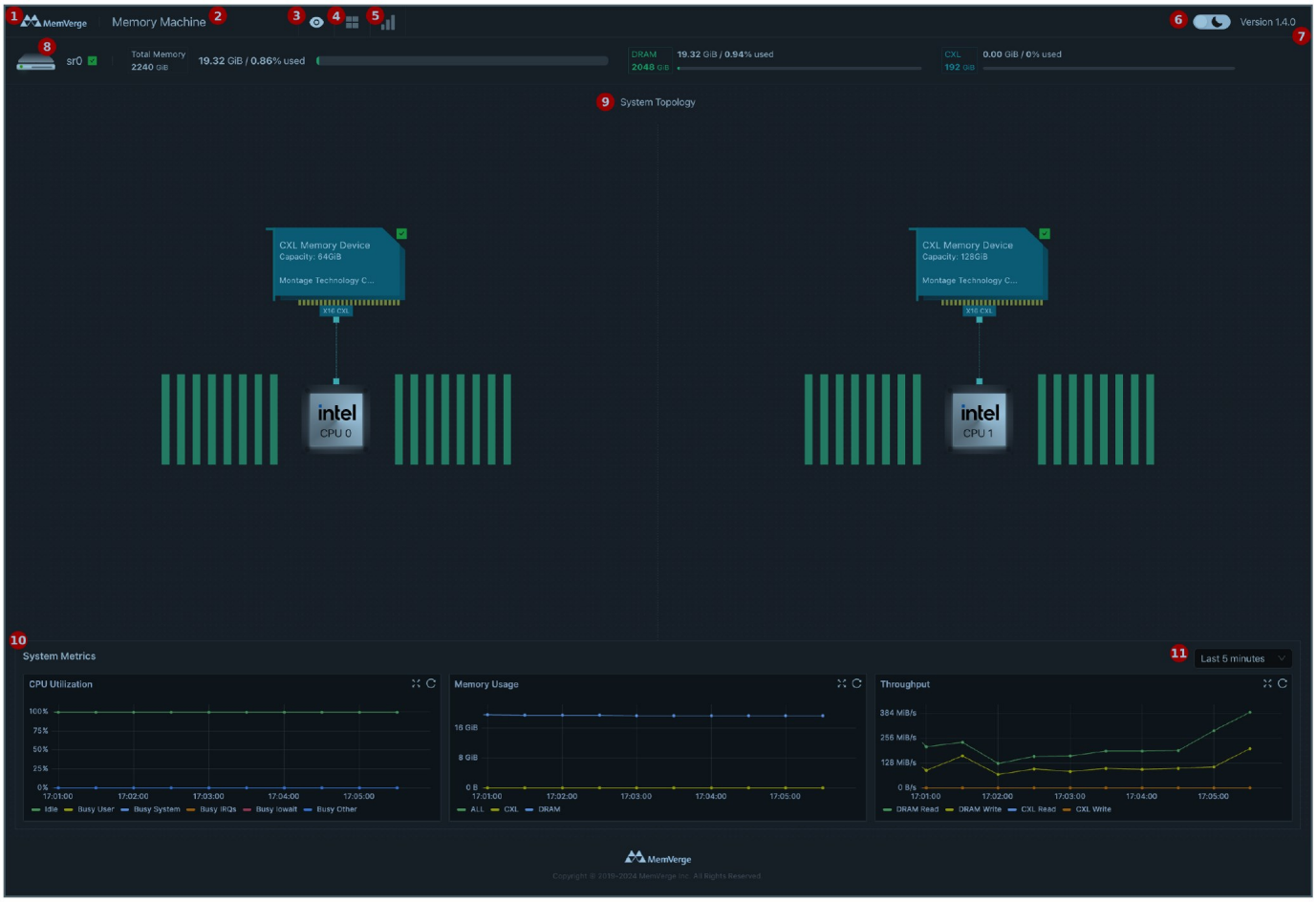
Memory Machine X | Server Expansion

Provides IT organizations with the tools needed to determine if adding CXL memory is suitable to their environment, and ensure bandwidth or latency quality-of-service (QoS) with intelligent management of the memory-storage hierarchy. Intelligent tiering can also boost utilization of precious GPU resources.

Key Features & Benefits

Observability—Memory Machine starts by providing valuable insights into server resource usage to help IT organizations understand if and how servers can benefit from CXL memory. Once in production, Memory Machine continues to report the performance of applications using mixed memory configurations.

Memory Machine System Topology Dashboard



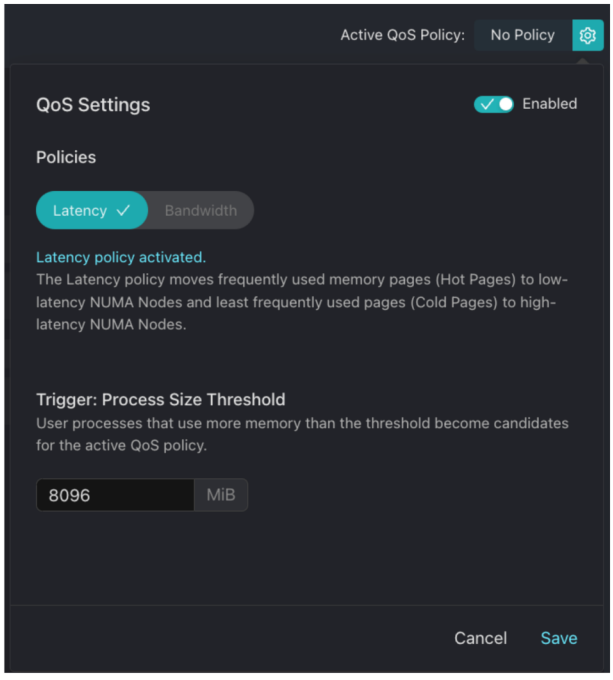
Legend: 1. Your company logo, 2. MemVerge product name, 3. Dashboard & system topology menu, 4. Quality of Service (QoS) menu item, 5. Insights menu item, 6. Light or dark UI mode selector, 7. MemVerge product version, 8. System information, 9. System topology, 10. System metrics, 11. System metrics date/time selector

Key Features & Benefits (cont.)

Intelligent Tiering of Mixed Memory—Memory Machine X makes it possible for IT organizations to scale memory in a server to 32TB at half the cost without sacrificing application performance. The Server Memory Expansion software makes this possible by adapting to varying workloads with intelligent placement policies and memory page movement to optimize latency or bandwidth.

- **Latency Policy**—Latency tiering intelligently manages data placement across heterogeneous memory devices to optimize performance based on the "temperature" of memory pages, or how frequently they are accessed. The MemVerge QoS engine moves hot pages to DRAM, where they can be accessed quickly. Cold are placed in CXL memory. By ensuring that frequently accessed data is stored in DRAM, the system reduces the average latency of memory accesses, leading to faster application performance. See how intelligent tiering with a latency policy was used to improved performance of MySQL based on TPC-C benchmark tests and how GPU utilization was boosted by 77%.
- **Bandwidth Policy**—The goal of bandwidth-optimized memory placement and movement is to maximize the overall system bandwidth by strategically placing and moving data between DRAM and CXL memory based on the application's bandwidth requirements. The bandwidth policy engine will utilize the available bandwidth from all DRAM and CXL memory devices with a user-selectable ratio of DRAM to CXL to maintain a balance between bandwidth and latency.

Setting Latency QoS

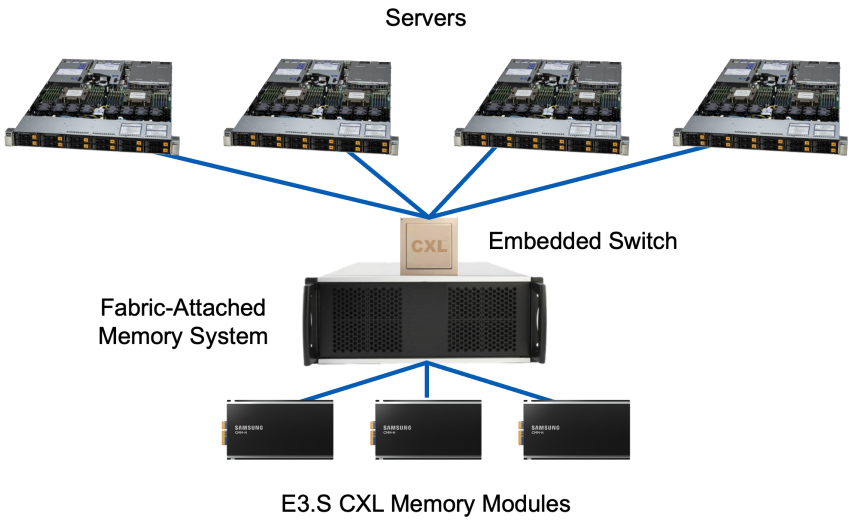


Setting Bandwidth QoS

Fabric-Attached CXL Memory

CXL 2.0 allows multiple servers to connect and share physical memory addresses from a memory pool.

New Memory Sharing Model

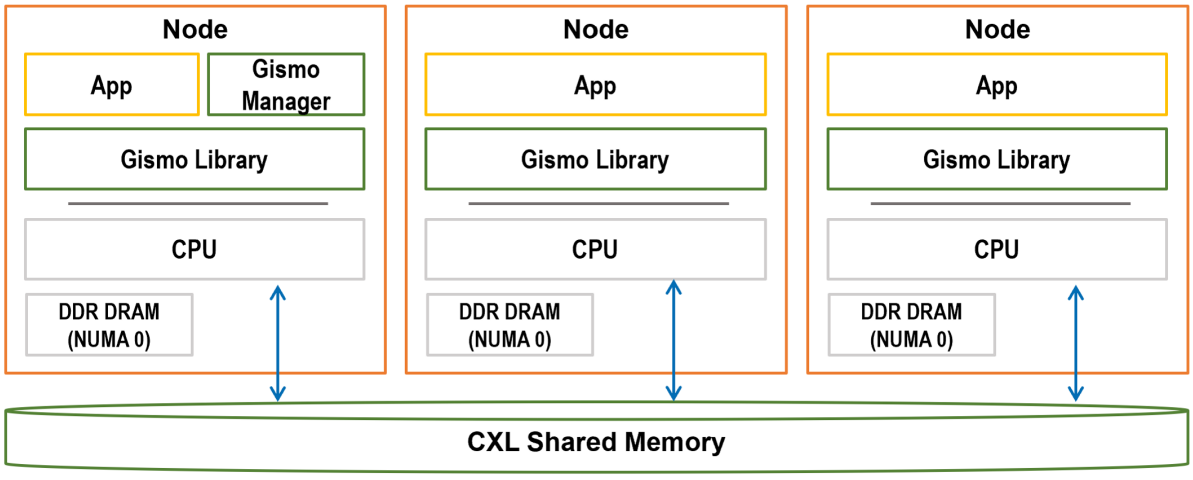


Memory Machine X | Fabric-Attached Memory

With our hardware partners, MemVerge delivers a memory appliance that enables IO-free data sharing between the servers with shared memory. Memory Machine X ensures cache coherency through software.

Memory Machine X | Fabric Attached Memory software includes a memory object store API called GISMO that allows applications to create and access memory objects across multiple nodes using memory semantics. GISMO reduces or eliminates transferring data over the network, the most costly step of network-based message passing, by allowing applications to directly access data in the shared memory pool and maintain cache coherence between processors in different servers. See how GISMO powered the Ray AI framework to 675% faster remote gets and 280% faster shuffles across 4 nodes.

GISMO (Global IO-free Shared Memory Objects)

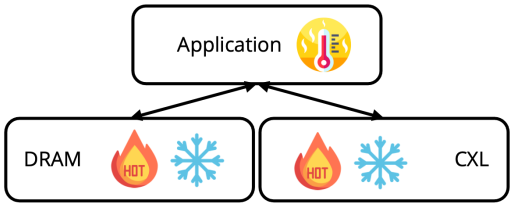


Weaviate Vector Database

Server Memory Expansion Use Case

Vector databases complement generative AI models by providing an external knowledge base for generative AI chatbots and by helping to ensure they provide trustworthy information. Weaviate is an AI-native vector database used by developers to create intuitive and reliable AI-powered applications. Weaviate benchmark tests are available to measure queries per second and latency.

For Weaviate, the Memory Machine X bandwidth-optimized memory placement and movement maximized the overall system bandwidth by strategically placing and moving data between DRAM and CXL memory based on the application's bandwidth requirements.

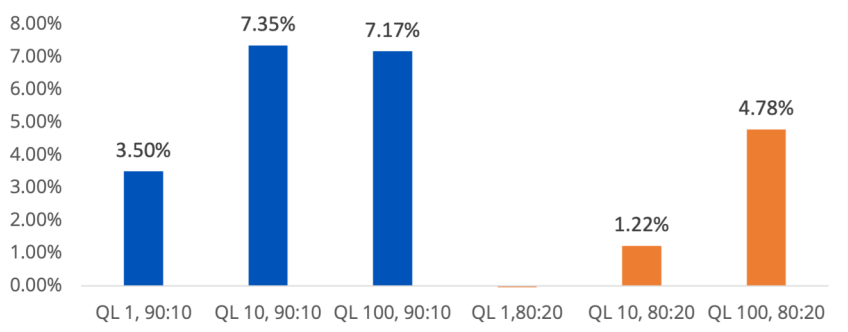


The bandwidth policy engine utilized the available bandwidth from all DRAM and CXL memory devices with a user-selectable ratio of DRAM:CXL to maintain a balance between bandwidth and latency.

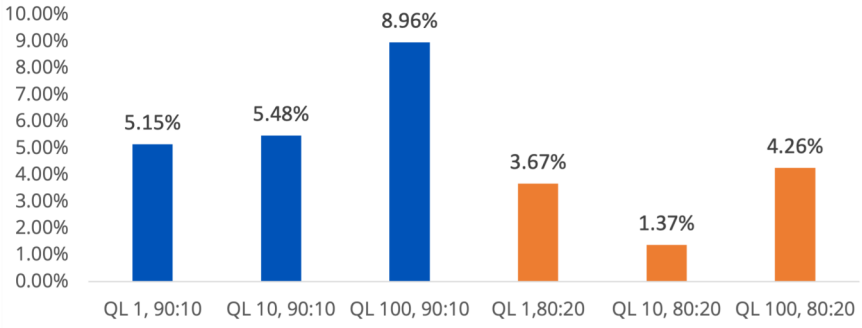
Shown below is QPS testing performed by MemVerge. Using 10% CXL and 20% CXL memory across different query limits (QL), Memory Machine X powered Weaviate to deliver up to 7.35% more queries per second.

Weaviate queries per second with Memory Machine X

(gist-960-Euclidian-128-32 – Queries per Second – EF512)



Shown below is latency testing performed by MemVerge. Using 10% CXL and 20% CXL memory across different query limits (QL), Memory Machine X enabled Weaviate to deliver up to 8.96% lower latency.



Weaviate latency with Memory Machine X

(gist-960-Euclidian-128-32 – P95 Latency (ms) – EF512)

FlexGen

Server Memory Expansion Use Case

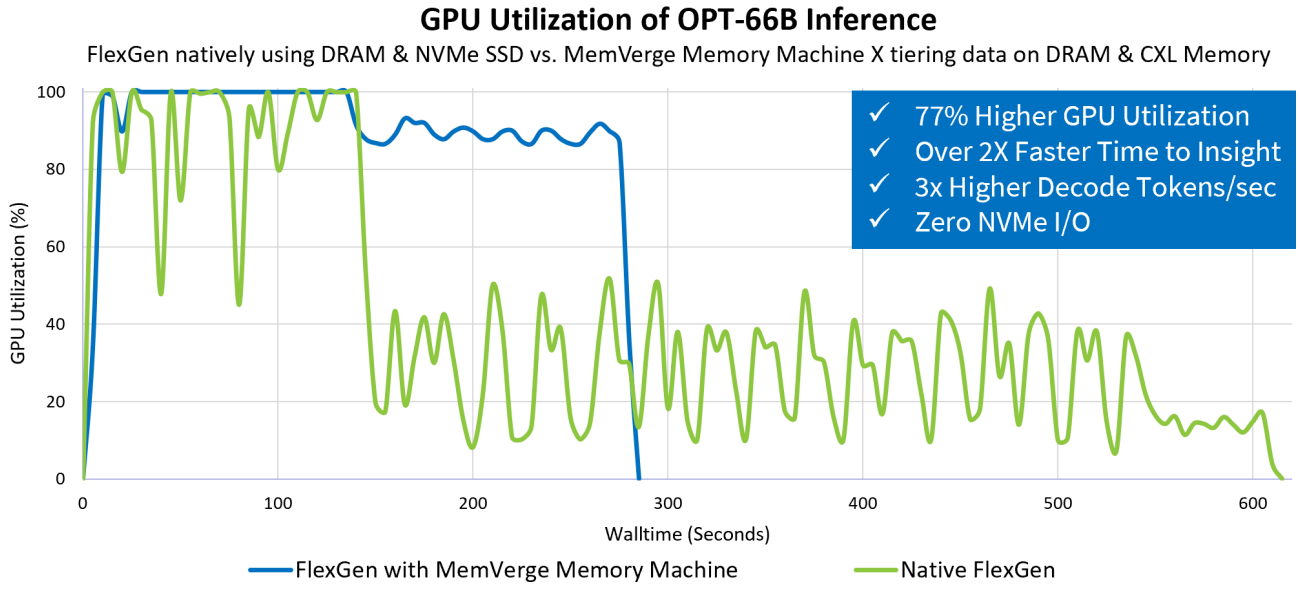
FlexGen is a high-throughput generation engine for running large language models with limited GPU memory. FlexGen allows high-throughput generation by IO-efficient offloading, compression, and large effective batch sizes.

MemVerge joined forces with Micron to develop a groundbreaking solution that leverages intelligent tiering of CXL memory, boosting the performance of large language models (LLMs) by offloading from GPU memory to CXL memory.

The solution, developed by engineers from MemVerge and Micron featured a FlexGen high-throughput generation engine and OPT-66B large language model running on a Supermicro Petascale Server equipped with an AMD Genoa CPU, Nvidia A10 GPU, Micron DDR5-4800 DIMMs, CZ120 CXL memory modules, and MemVerge Memory Machine™ X intelligent tiering software.

The results of the solution were impressive. The FlexGen benchmark, utilizing tiered memory, completed tasks in less than half the time compared to conventional NVMe storage methods. Simultaneously, GPU utilization soared from 51.8% to 91.8%, thanks to the transparent management of data tiering across GPU, CPU and CXL memory facilitated by MemVerge Memory Machine X software.

FlexGen Benchmark Results



MySQL

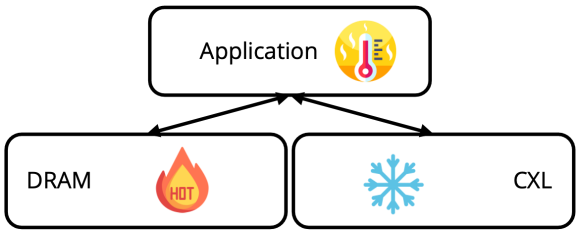
Server Memory Expansion Use Case

MySQL is the world's most popular open source database. Many of the world's largest and fastest-growing organizations including Facebook, Twitter, Booking.com, and Verizon rely on MySQL to save time and money powering their high-volume Web sites, business-critical systems and packaged software.

MemVerge used TPC-C benchmark tests to compare the performance of MySQL using Transparent Page Placement in the kernel vs. MySQL using the Memory Machine Server Expansion latency policy.

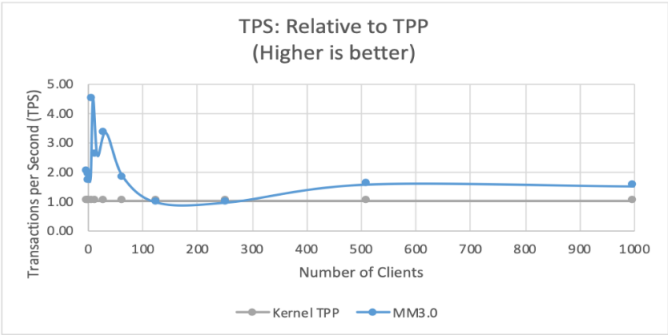
The Memory Machine X QoS policies adapt to various application workloads with memory page movement to optimize latency or bandwidth.

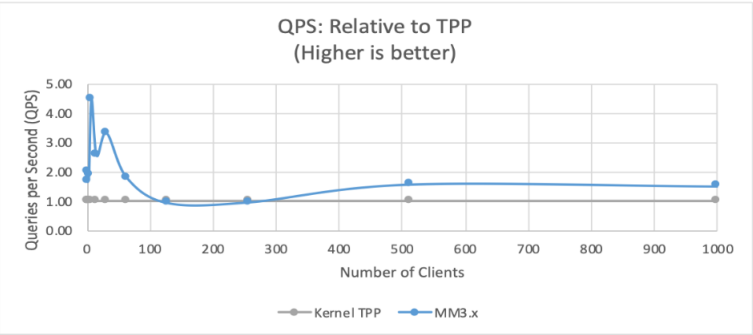
Latency tiering intelligently manages data placement across heterogeneous memory devices to optimize performance based on the "temperature" of memory pages, or how frequently they are accessed. The MemVerge QoS engine moves hot pages to DRAM, while cold pages are placed in CXL memory to reduce the average latency of memory accesses, leading to faster application performance.

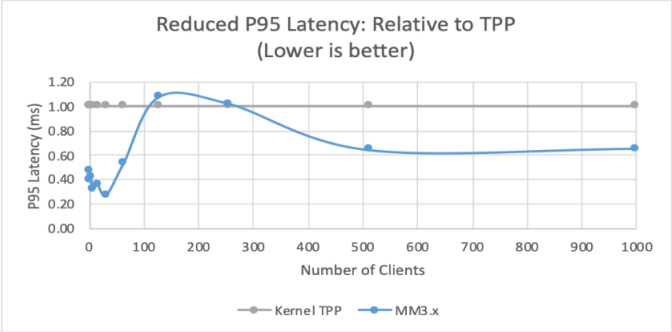


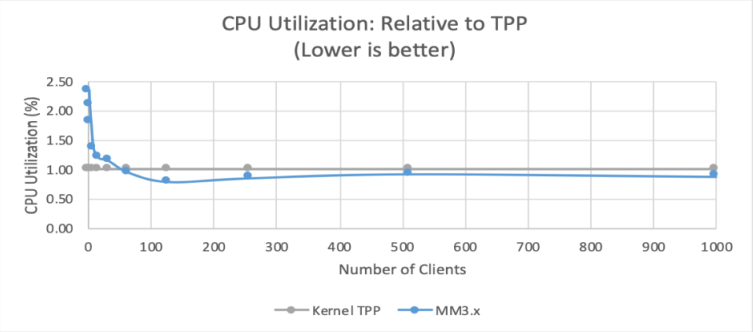
Memory Machine enabled 20-40% lower latency (lower left chart) which allowed MySQL to deliver higher transactions per second (TPS), queries per second (QPS), and CPU utilization.

TPC-C Benchmark Results









½ Cost, 3x Capacity with CXL

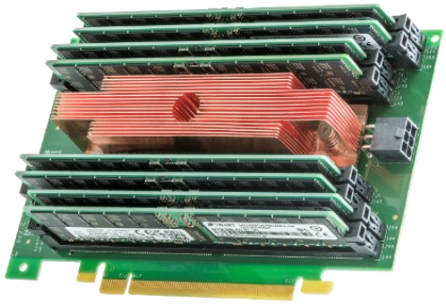
Server Memory Expansion Use Case

50% less cost—If you want to lower the cost of your server memory (who doesn't?), CXL technology offers a lower cost alternative to128GB DIMMs for scaling to multi-terabyte configurations.

A new class of PCIe Add-In Cards (AICs) support up to 8 lower cost 64GB and/or DDR4 DIMMs. The result you see below in the table is you can scale capacity up to 8TB for half the cost with a mixed memory configuration.

Memory Machine X Server Expansion software tells you if CXL memory is appropriate for your application environment and automatically places your data in the right memory tier for optimum performance.

CXL Add-in Card with DDR4 DIMMS



Mixed DIMM and CXL Memory Configurations

Memory Configurations	DIMM size, type	Total Capacity (GB)	$/GB	Total Cost	Savings $	Savings %
4TB						
A. DIMM-only	128GB	4,096	$11.25	$46,080	-	-
B. DIMM & CXL	64GB, DDR5	4,096	$5.60	$22,928	$26,008	50%
C. DIMM & CXL	64, DDR4	4,096	$4.90	$20,072		-
8TB						
D. DIMM-only (quad-socket server required)	128GB	8,193	$11.25	$92,160	-	
E. DIMM & CXL	64, DDR4	8,192	$5.77	$47,288	$44,872	49%

300% more capacity—With a mix of DIMMs and AICs, it's possible to configure a 2-socket server with up to 11.26TB of memory, or 38% more than the 8TB capacity possible using only DIMMs. With a mix of DIMMs and AICs, each with 8 x 256GB DIMMs, it's possible to configure a 4-socket server with 32TB of memory, or 300% more than the 8TB capacity possible using only DIMMs.

Mixed DIMM and CXL Memory Configurations

Memory Configurations	DIMM size, type	Total Capacity (GB)	$/GB	Total Cost	Capacity >8TB	Added Capacity %
11TB						
F. DIMM & CXL	96GB , DDR4	11,264	$5.56	$62,656	3,072	38%
32TB						
G. DIMM and CXL (quad-socket server required)	256GB, DDR5	32,768	$12.99	$425,600	24,576	300%

Ray AI Framework

Fabric-Attached Memory Use Case

OpenAI is using Ray to train its largest models, including ChatGPT. Ray powers their solutions to the thorniest of their problems and allows OpenAI to iterate at scale much faster than before.

In a baseline Ray environment, sharing data between processes using message passing involves a 3-step process:

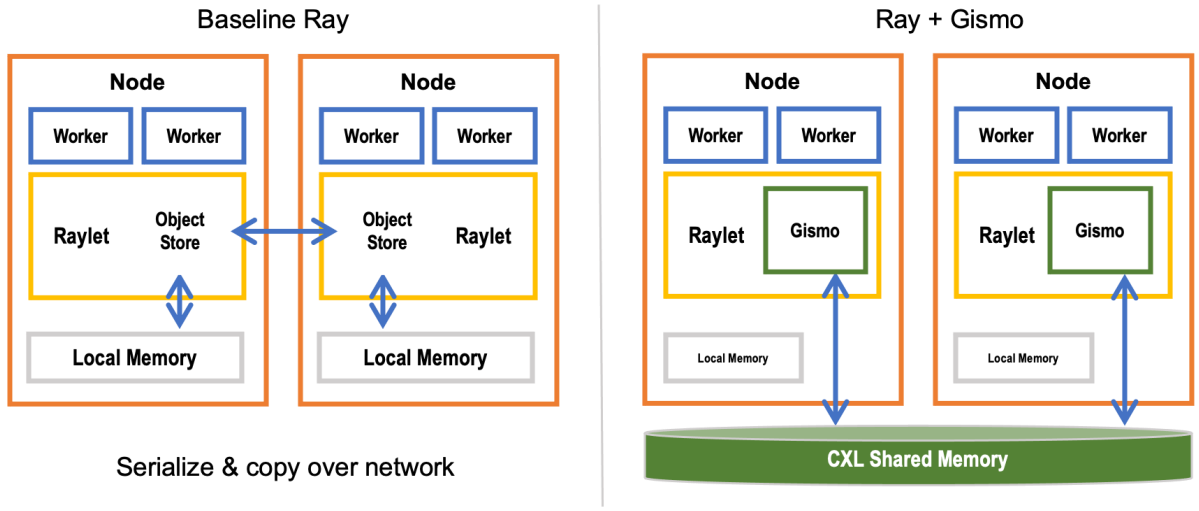
Writing data to local memory in node A

Passing the message across the network

Writing the data to local memory in node B

Memory Machine X Fabric-Attached Memory makes Ray clusters IO-free by eliminating object serialization and transfers over the network for remote object access. It includes a memory object store API called GISMO that allows applications to create and access memory objects across multiple nodes using memory semantics.

Using GISMO, node A writes to shared memory and node B reads from shared memory. GISMO maintains cache coherence between the nodes and delivers high throughput and low latency in single-writer, multiple-reader application environments such as Ray-based AI.



In testing performed by MemVerge using software emulation of a pooled CXL memory sharing environment, Memory Machine X Fabric Attached Memory software delivered the same access time for a local get object, 675% faster access time for a remote get object, and 280% better performance for a shuffle across 4 nodes.

Shuffle Benchmark Results

	Baseline Ray	With Gismo	Difference
Local Get 1GB object	0.4 sec	0.4 sec	**CXL shared memory as fast as local memory**
Remote Get 1GB object	2.7 sec	0.4 sec	**675% faster**
Shuffle 50GB, 4 nodes, each 4 cores, 128 GB object store	515 sec	185 sec	**280% faster**

Learn More

MemVerge.com

CXL specifications at the CXL Consortium website

Dozens of CXL videos on the Memory Fabric Forum YouTube channel

Daily CXL news in the Memory Fabric Forum LinkedIn Group

Get Started

Request a Server Memory Expansion
or Fabric-Attached Memory PoC now

MemVerge