

# CXL Server Memory Expansion and Fabric Attached Memory

Presenter: Steve Scargall, Sr. Product Manager & Software Architect, MemVerge



## Server Memory Expansion

Tier your memory



Observability



Latency  
QoS

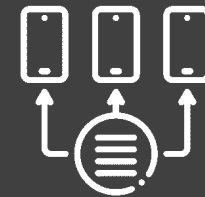


Bandwidth  
QoS



## Fabric-Attached Memory

Share your memory



Observability

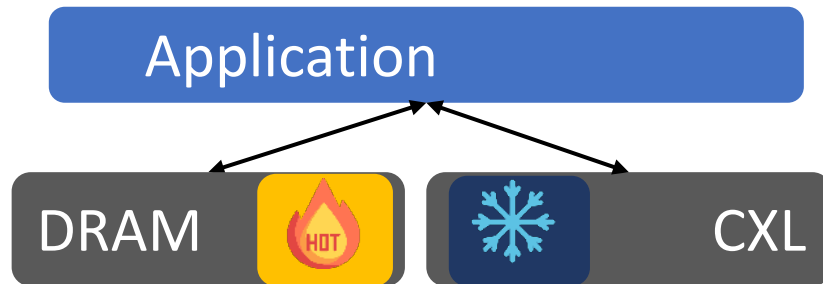


GISMO (Global IO-free Shared  
Memory Objects)



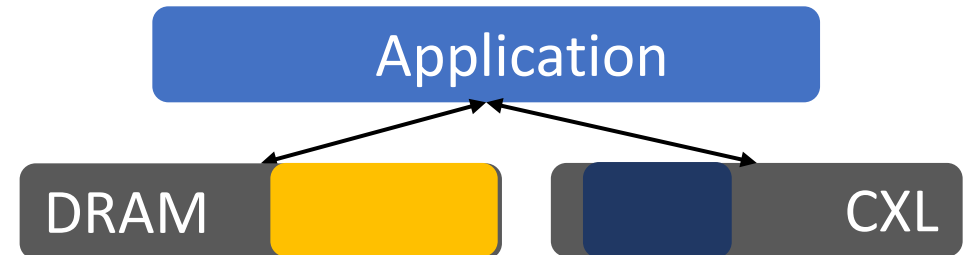
# Memory Machine X: Intelligent Memory Placement Engine

## Latency Tiering



Latency tiering policies intelligently manage data placement and movement to match the "temperature" of memory pages – Hot or Cold – with the right tier of memory devices.

## Bandwidth Tiering



Bandwidth Tiering Policies strategically place data between different tiers of memory proportionate to the bandwidth ratio of the tiers



# MemVerge CXL Benchmark Suite

- <https://github.com/cxlbench/cxlbench>
  - Intel MLC
  - CloudSuite
    - Graph Analytics
    - In-Memory Analytics
  - MemCacheD
  - Redis
  - STREAMS (Modified to support CXL)
  - TPC-C (MySQL + Sysbench)



# MemVerge's Open Source Contributions

- QEMU
  - Expansion
  - Sharing
  - DCD
  - Multi-Head DCD
- Linux Kernel Weighted Interleave (6.9 onwards)
  - Introduces bandwidth optimization for CXL
  - Blog: <https://memverge.com/introducing-weighted-interleaving-in-linux-for-enhanced-memory-bandwidth-management/>



# Weighted NUMA Interleaving

- Contributed by MemVerge & SK hynix
- Available in Kernel 6.9 or newer:  
<https://github.com/torvalds/linux/blob/master/mm/mempolicy.com>
- Weighted interleave is a **new policy** intended to use heterogeneous memory environments appearing with CXL.
- Weighted interleave distributes memory across nodes according to a provided weight. (Weight = # of page allocations per round).
- As bandwidth is pressured, latency increases. [Figure 1](#)
- Allows greater use of the total available bandwidth in a heterogeneous hardware environment. [Figure 2](#)

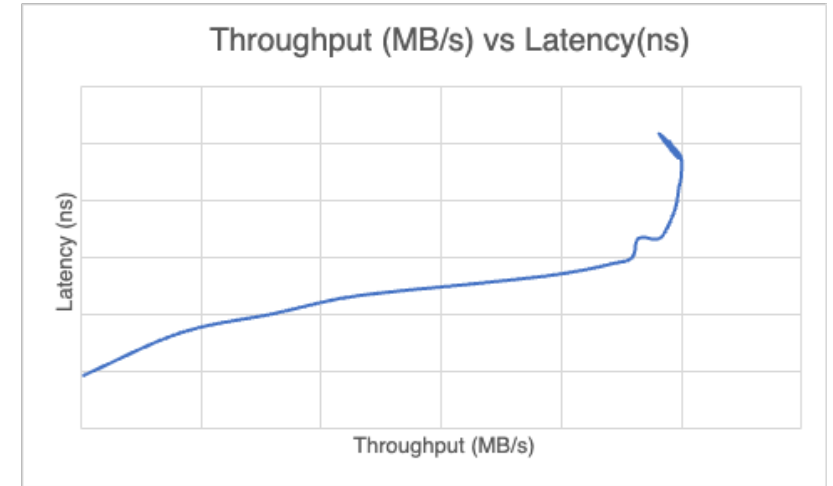


Figure 1: Throughput vs Latency (Hockey Stick)

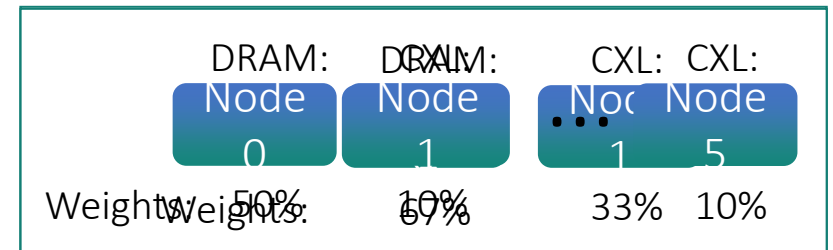
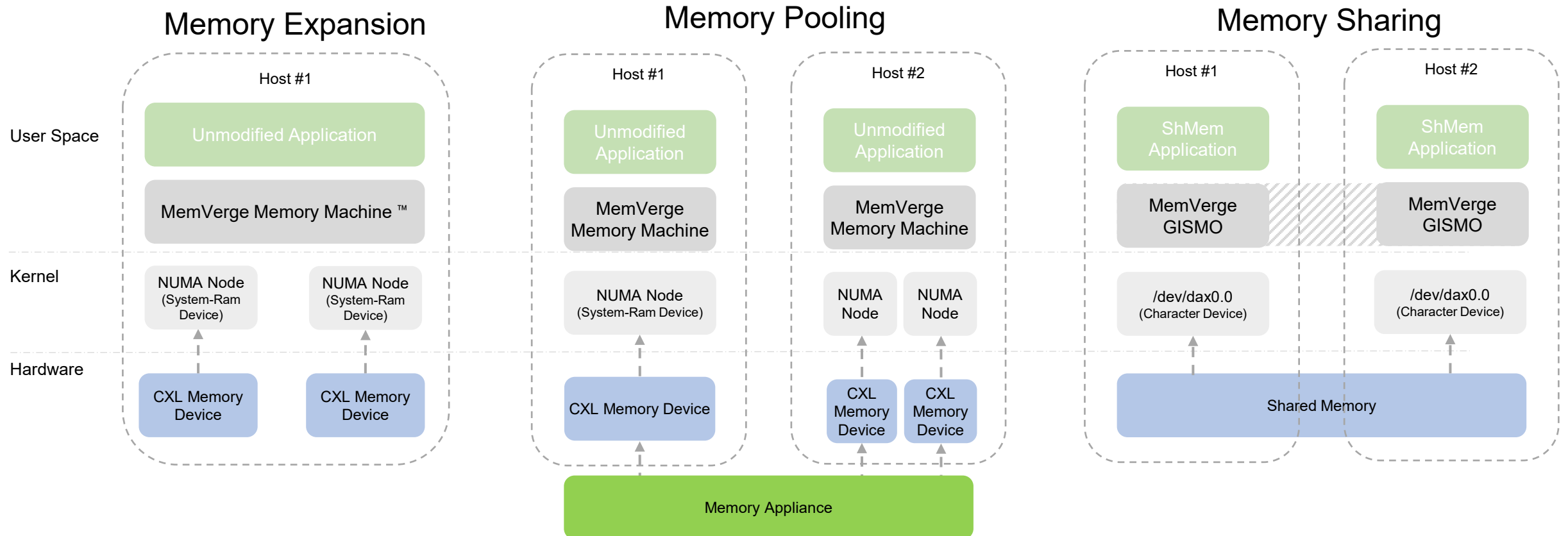


Figure 2: Weighted Interleave 5x CXL Example



# How Applications will use CXL for Large-Scale Datasets



# CXL for AI/ML Workloads

- FlexGen is a high-throughput generation engine for running large language models with limited GPU memory.
- FlexGen allows high-throughput generation by efficient offloading to main memory (or CXL)
- Paper: <https://arxiv.org/abs/2303.06865>
- GitHub: <https://github.com/FMInference/FlexGen>



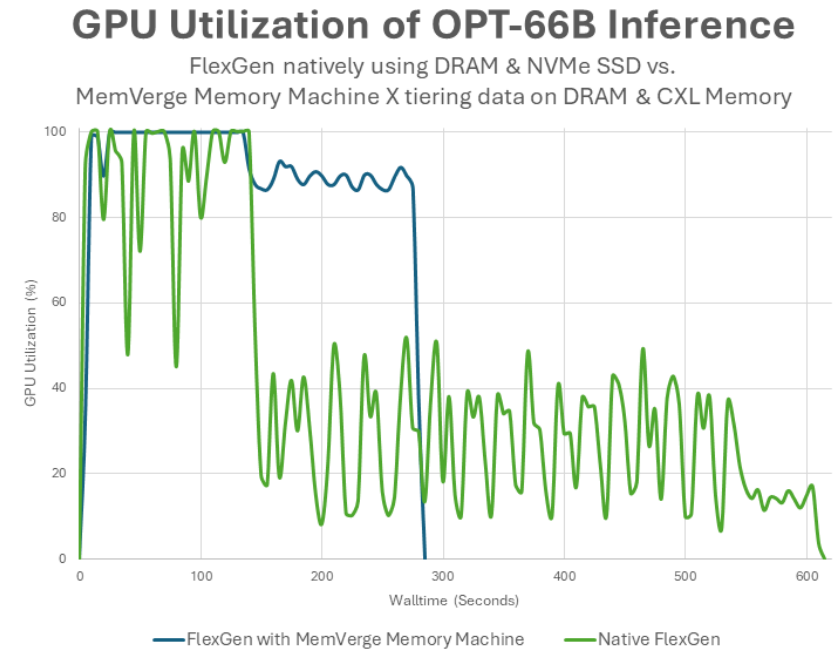
**MemVerge and Micron Rescue Stranded NVIDIA GPUs with CXL® Memory**

*AI data intelligently tiered on DIMM and CXL memory to maximize utilization of one of the world's most precious resources, GPU.*

- ✓ 77% Higher GPU Utilization
- ✓ Over 2X Faster Time to Insight
- ✓ 3x Higher Decode Tokens/sec
- ✓ Zero NVMe I/O

**NVIDIA GTC** See the demo in Micron booth #1030

**MemVerge** **Micron**



**FlexGen Native:** Supermicro Petascale Server, AMD Genoa 9634 DP/UP 84C/168T, 8 \* 32GB Micron DDR5-4800, 2 x Micron 7450 960GB M.2, Nvidia A10 GPU, Ubuntu 22.04.04 using Kernel 5.15.0, [FlexGen AI](#)

**FlexGen using MemVerge Memory Machine:** Supermicro Petascale Server, AMD Genoa 9634 DP/UP 84C/168T, 8 \* 32GB Micron DDR5-4800, 2 x Micron 7450 960GB M.2, 1 x Micron CZ120 CXL, Nvidia A10 GPU, Ubuntu 22.04.04 using Kernel 5.15.0, MemVerge Memory Machine 2.5.1, [FlexGen AI](#)



# Ollama inference server with CXL acceleration



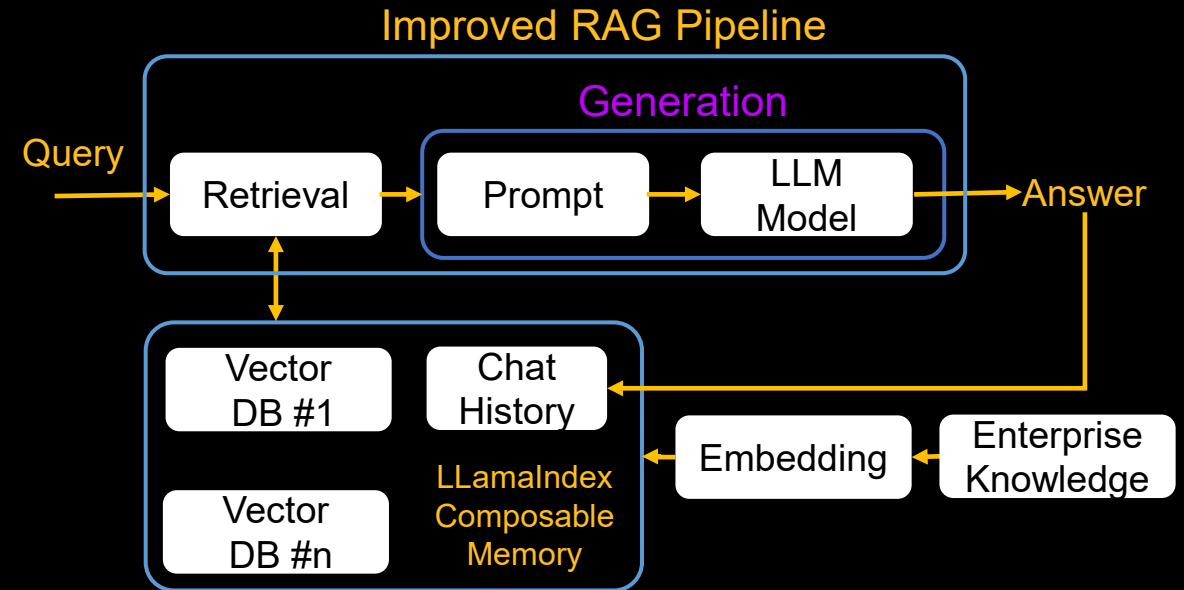
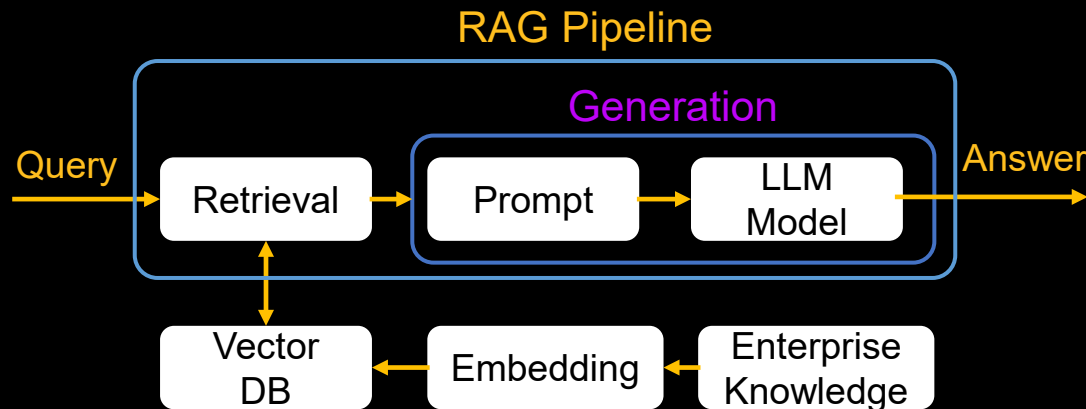
## System configuration

Platform	Supermicro platform
CPU family	Intel Xeon Gen 5 processor
Native DRAM	Micron DDR5 – 64GB – 5600 MTs Capacity: 1TB (up to 8 RDIMM slots)
CXL DRAM	Micron CZ120 Capacity: 1 TB (4 * Micron CZ-120 256 GB) Total B/W: 4 x 26 = 104 GB/s (100% Read)
GPU	Nvidia A100

## Software specification

Applications	LlamaIndex, Qdrant Vector DB, LLAMA 2
OS	Red Hat Enterprise Linux
Kernel	6.8 rc-5 – with weighted s/w interleaving enabled

# Improved RAG Pipeline- LlamaIndex example



## Common Bottlenecks in a RAG Pipeline:

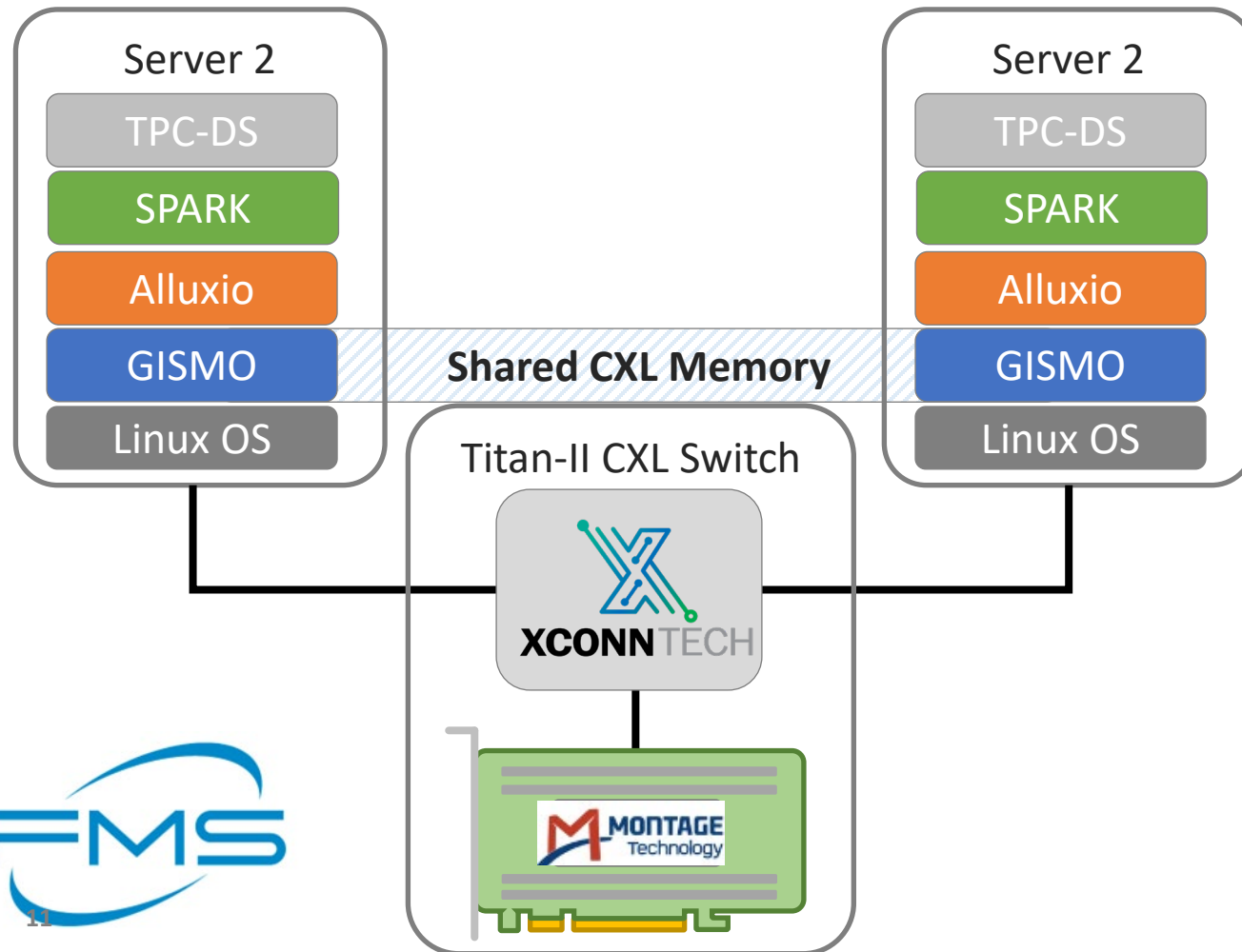
- CPU: Embedding generation
- GPU: LLM inference
- Memory: Storing large datasets and embeddings spills to disk
- Storage I/O: Reading/writing vector data

## Benefits of Adding CXL Memory in a RAG Pipeline:

- Efficient retrieval of large-scale and multiple vector databases
- Improved User answer quality due to rich retrieved context
- Large memory capacity can be shared and pooled across nodes
- Improved RAG performance improves GPU utilization



# A TPC-DS + Spark Workload Accelerated by MemVerge GISMO, an X-Conn Switch, and Montage CXL Memory in Sharing Mode




## Benefits of this Solution:

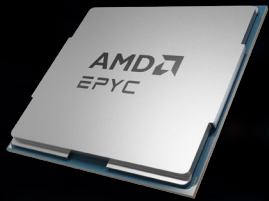
- Alluxio uses a **shared CXL cache (GISMO)** vs local node caching, reducing DRAM requirements
- Significantly reduces Disk and Network I/O (HDFS).
  - Data is accessed over the CXL memory bus instead.
- TPC-DS Lower Query Latency vs 1G Ethernet
- TPC-DS Higher Queries/Requests per Second (QPS) vs 1G Ethernet
- TPC-DS Reduced time to result vs 1G Ethernet



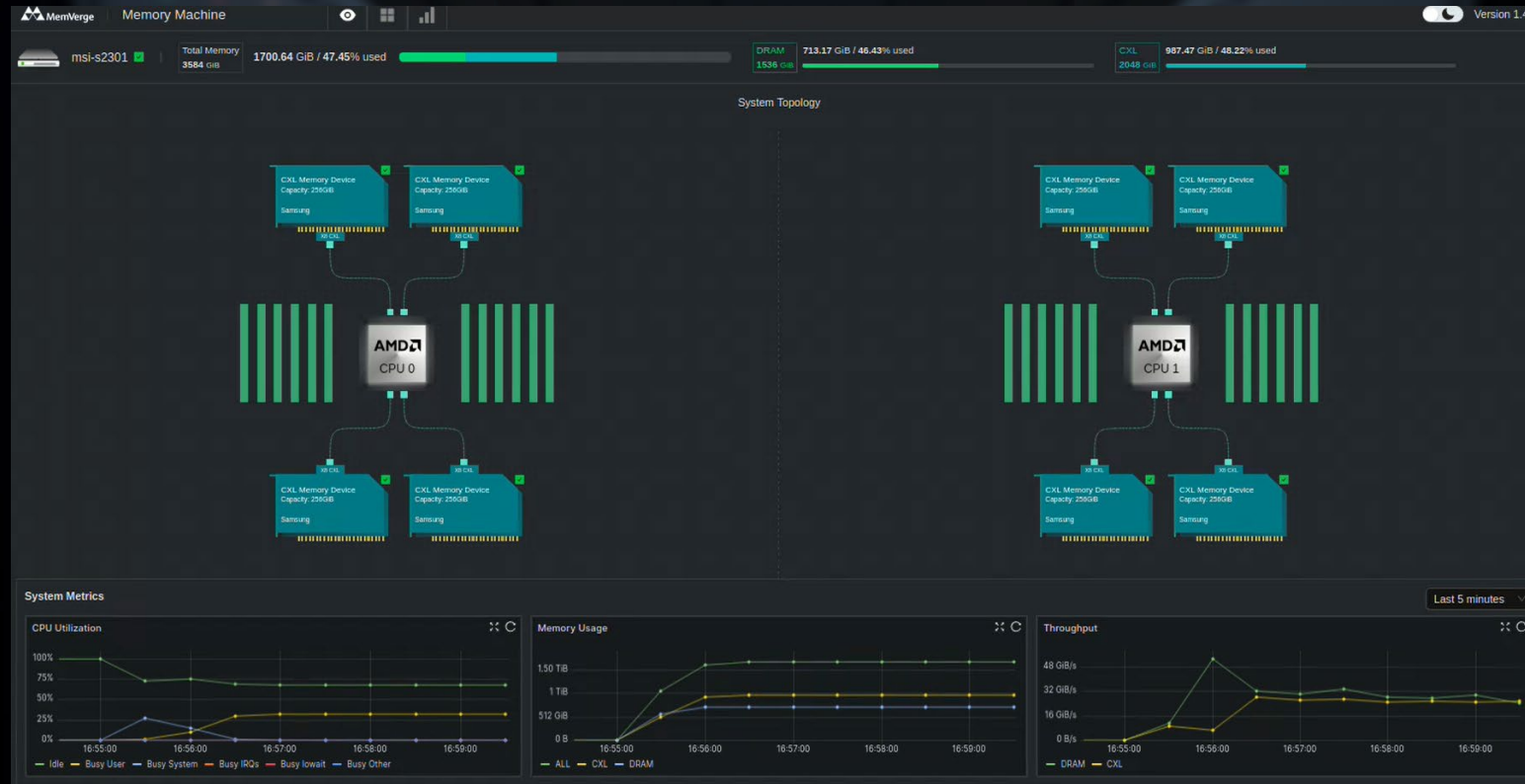
# CXL® Memory Expansion Server

Now Available for  
Enterprise PoCs

 **MemVerge** Memory Machine™ X Intelligent Memory Tiersing Software



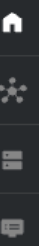
**AMD**  
Epyc Processors




**SAMSUNG**  
CXL Memory Modules





**msi** | SERVER




### Dashboard

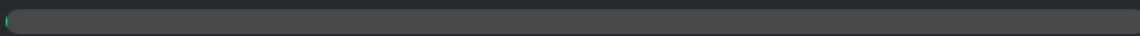
 Compute Nodes 5

 Switches 2

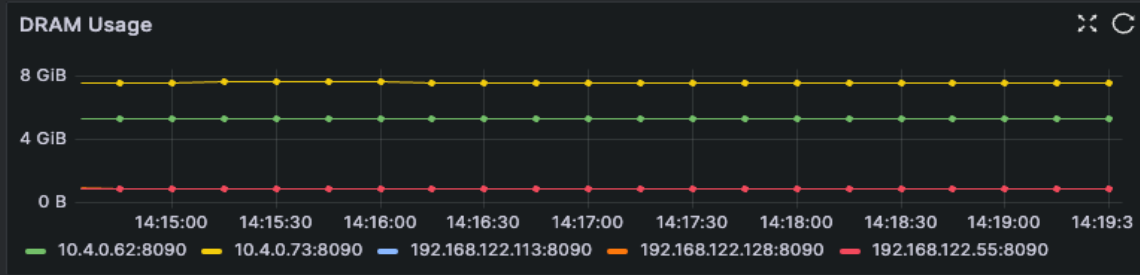
 CXL Devices 5

 GPUs 1

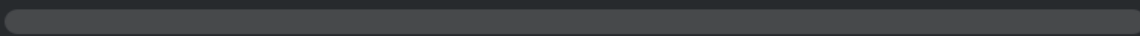
**DRAM Usage** Total: 2656 GiB  
Used: 0.87 GiB Free: 2656 GiB



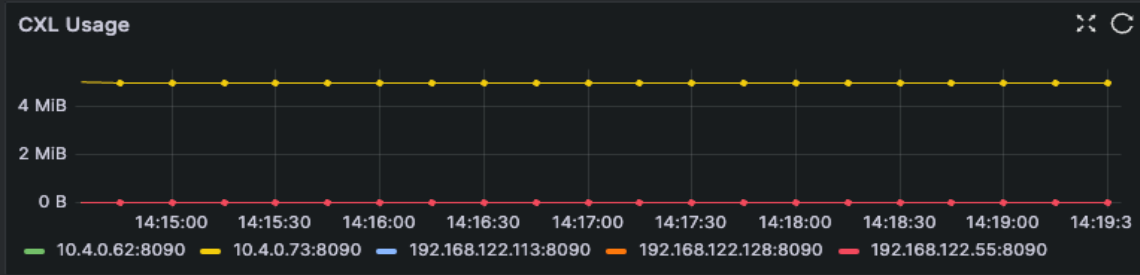
Usage Trend Last 5 minutes ▾



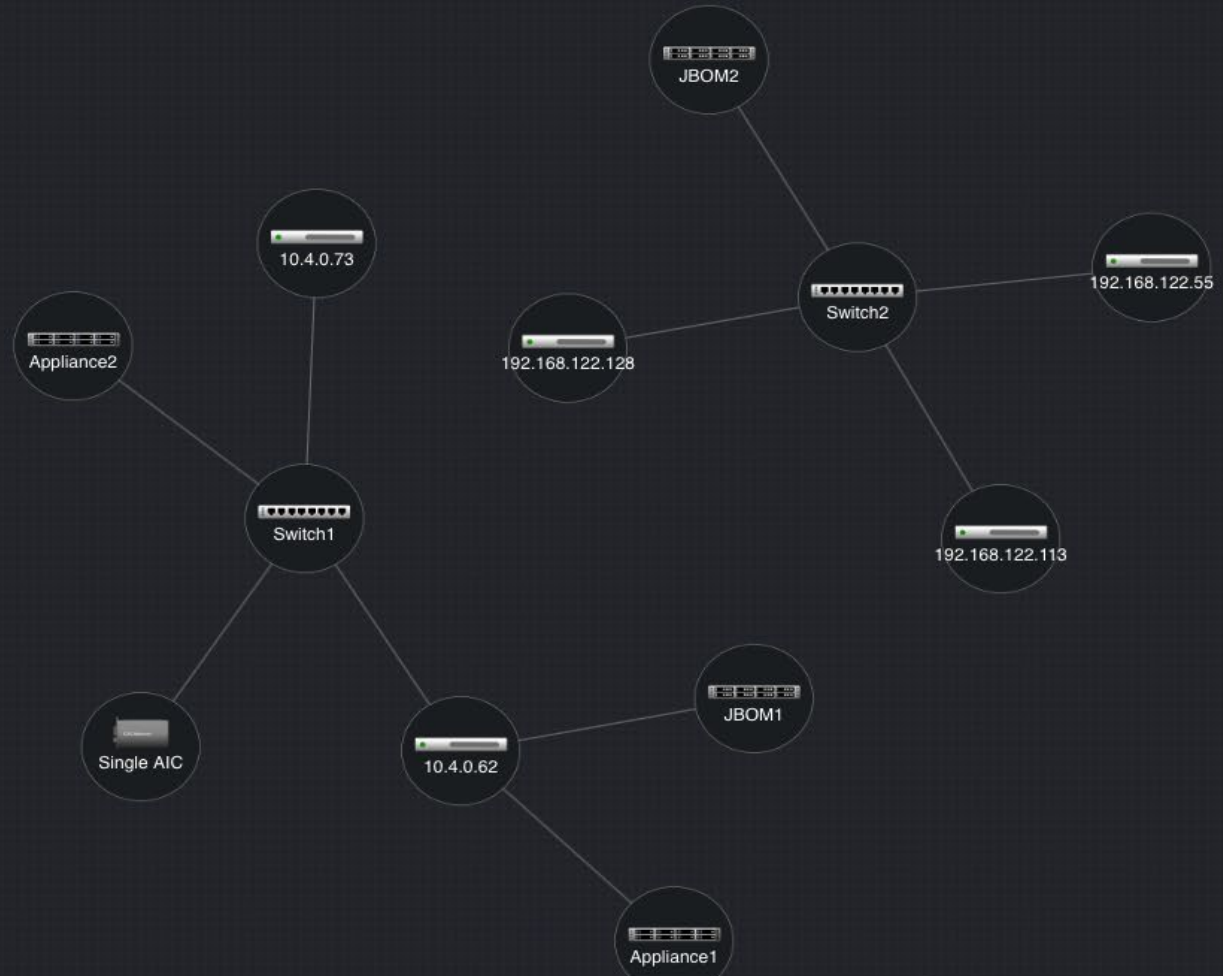
**CXL Memory Usage** Total: 256 GiB  
Used: 0 GiB Free: 256 GiB



Usage Trend Last 5 minutes ▾



Fabric Topology



Navigation and legend controls:

- Zoom in (+) and zoom out (-) icons.
- Zoom level: 100%
- Fit button.
- Legend items: Compute Nodes, Switches, Appliances, JBOMs, Single E3.S, Single AIC.

# Visit the MemVerge Booth #1251

- Demos
  - MemVerge Memory Machine X
  - CXL Expansion & Sharing
  - GPU Utilization
  - Acceleration of RAG pipelines
  - GISMO with Ray.io and Alluxio; TPC-DS Benchmark
- Hardware Showcase
  - CXL Devices & Servers
- Give Aways
- And more .....





**Thank you!**

