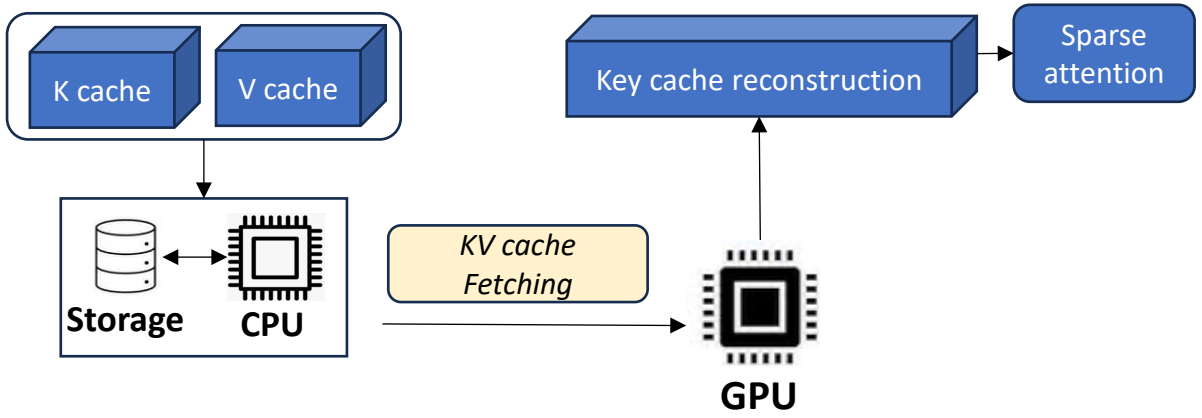
Accelerating AI with Real-World CXL Platforms

Sandeep Dattaprasad

Andy Mills

AI Inferencing Bottlenecks

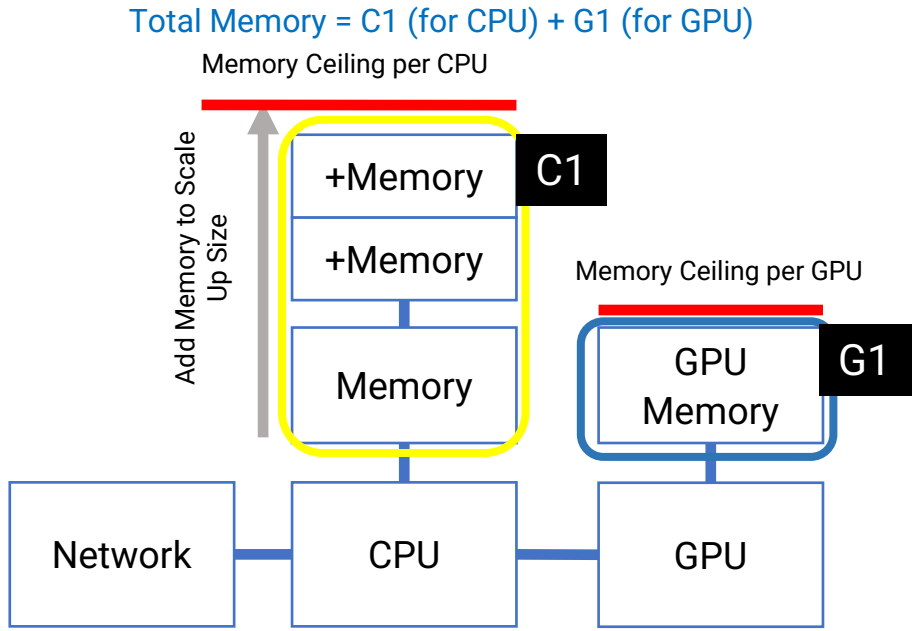
KV Caching Requirements



KV cache consumes a significant amount of memory

- Attention models may consume on the order of **~1MB/token**
- KV Cache size depends on precision, i.e., FP 8/16/32, BF16, INT8, etc.

Traditional Memory Scale up limitations

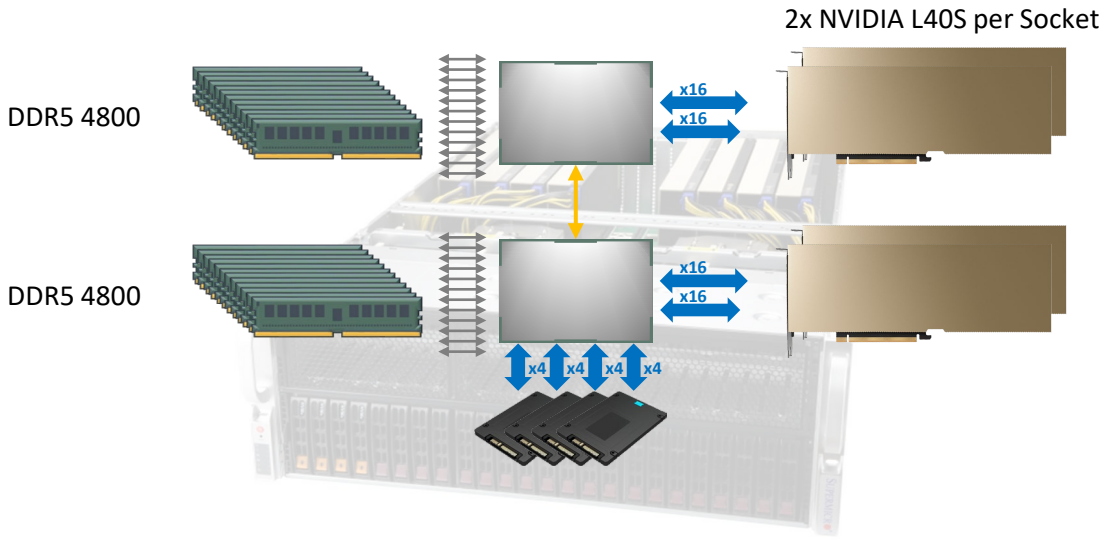


Limitations

- Compute (GPU & CPU) utilization
- CPU storage results in low performance due to high latency
- Overhead due to limited LLM instances per server

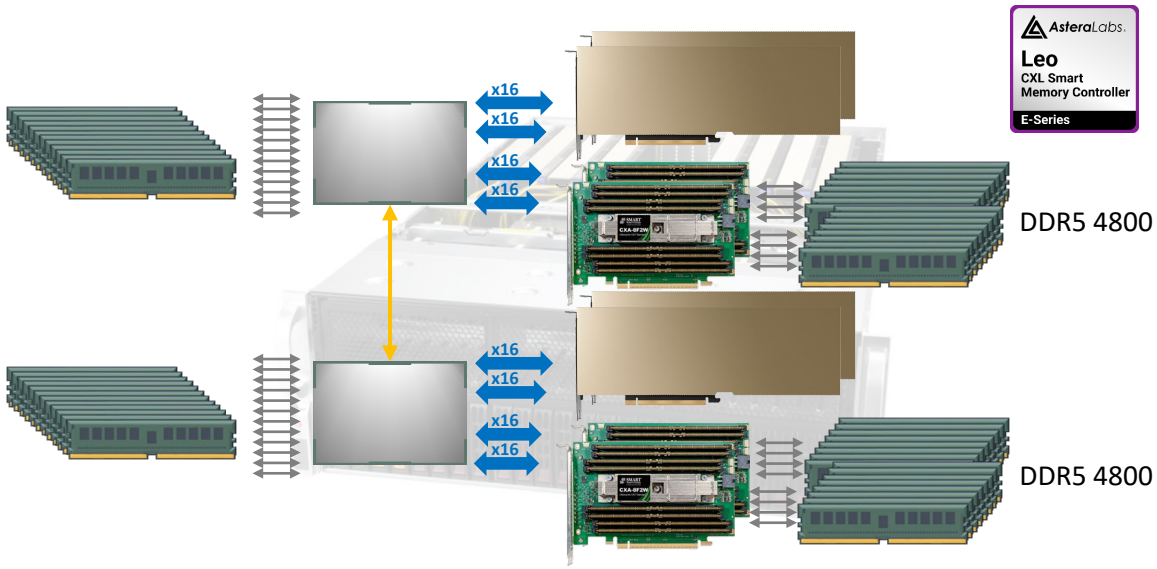
Unlock AI Performance with SMART Modular & Astera Labs' Leo

Four GPUs without CXL (24 DIMMs)



- **High latency of 4743 s w/ NVMe cache**
- **35% GPU utilization on average w/ high CPU overhead**
- **192 concurrent LLM instances on memory**

Four GPUs with Four CXL Smart Modular AICs (56 DIMMs)

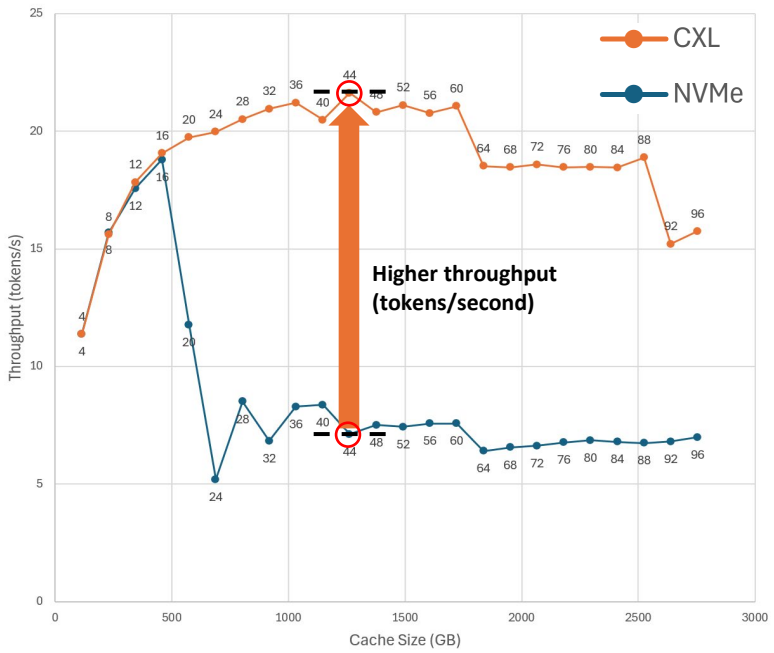


- **67% Lower latency of 1562.29 s w/ CXL**
- **75 % GPU utilization on average w/ low CPU overhead**
- **528 concurrent LLM instances on memory**

PROMPT_LENGTH = 512, GEN_LENGTH = 32, GPU_BATCH_SIZE = 12, NUM_BATCHES = 44, CACHE_ON_CPU%_NVME = 24, CACHE_ON_CPU%_CXL = 100

Performance Improvement w/ CXL Memory

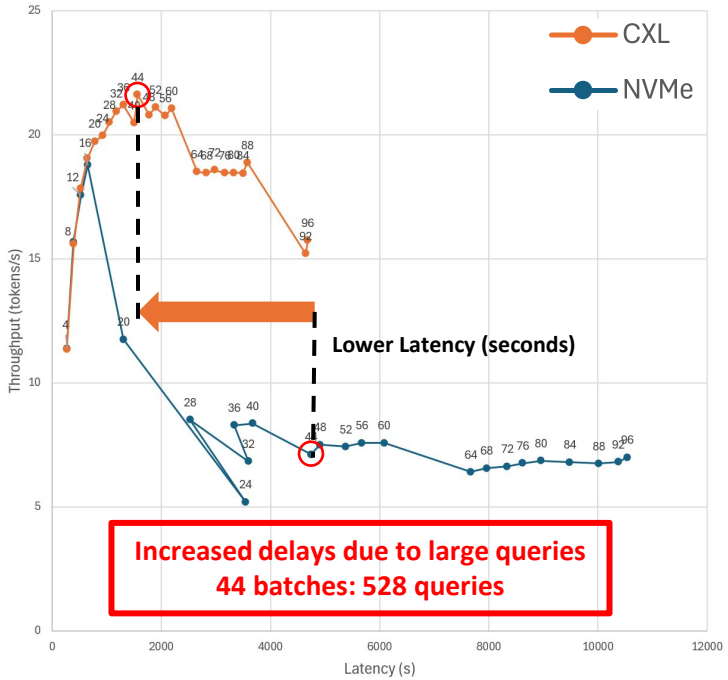
Cache Size Vs. Throughput



Steep performance drop with NVMe

More performance with CXL

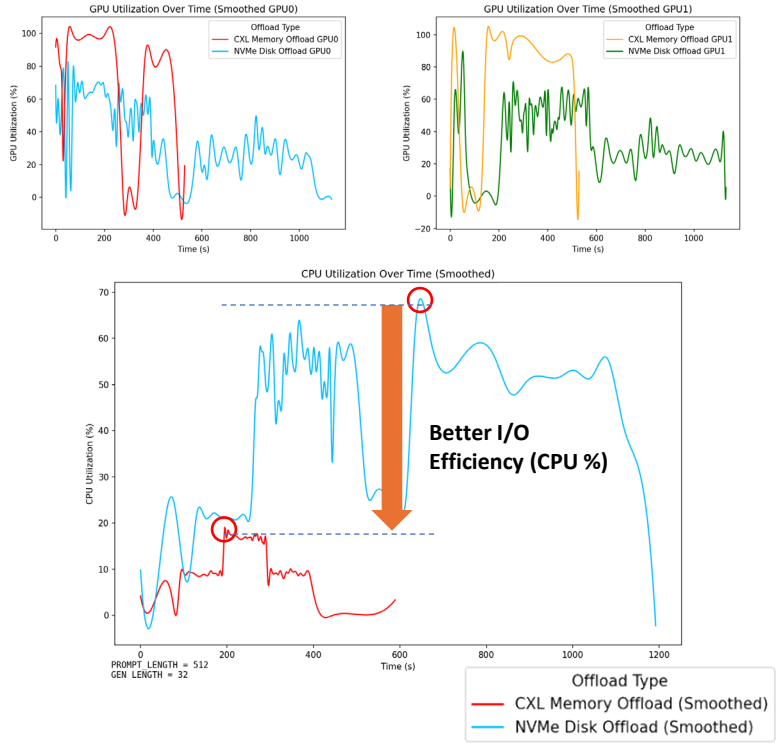
Latency Vs. Throughput



Slower response time with NVMe

Faster response time with CXL

GPU & CPU Utilization



Overburdened CPU due to slower media

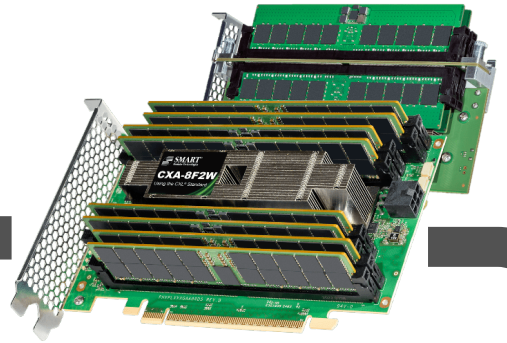
Efficient I/O from CXL Memory to GPU

PROMPT_LENGTH = 512, GEN_LENGTH = 32, GPU_BATCH_SIZE = 12, NUM_BATCHES = 44, CACHE_ON_CPU%_NVME = 24, CACHE_ON_CPU%_CXL = 100

Add-in-cards As a Critical CXL Building Block

CXL 2.0 Expansion Add-in-Cards

Fundamental component for first generation CXL memory appliances

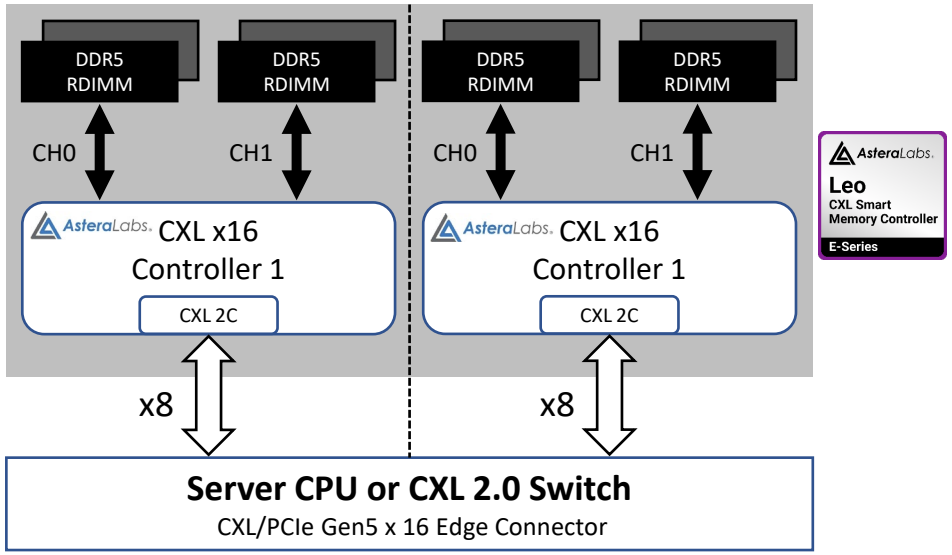


Fundamental component for first generation CXL standalone memory appliances

Data Center and Enterprise Servers

CXL 2.0 Retimer Add-in-Cards

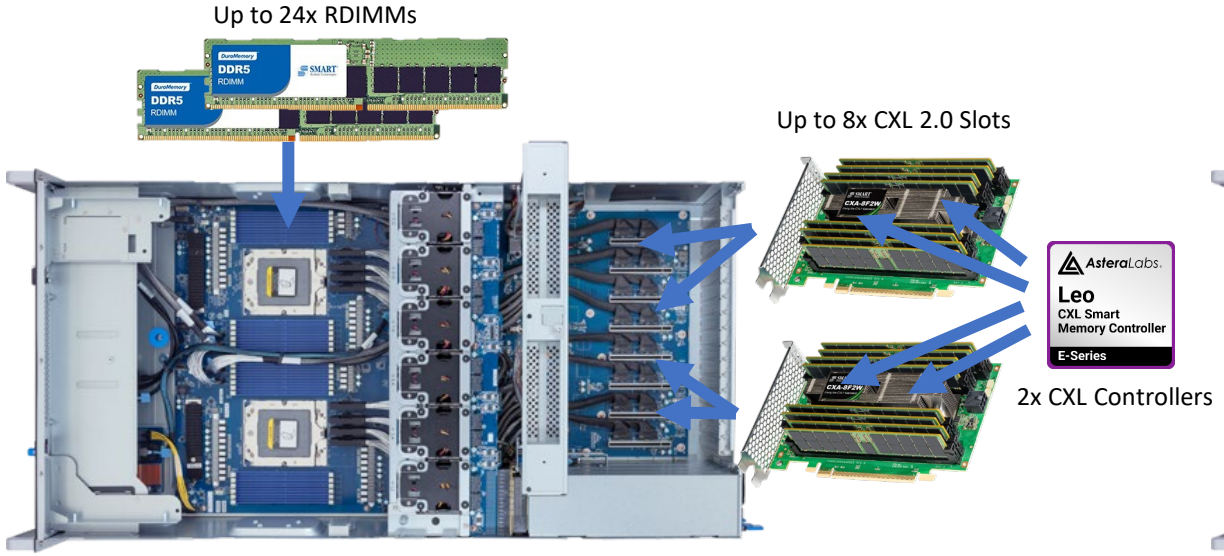
CXL 2.0 Memory Pooling Appliances



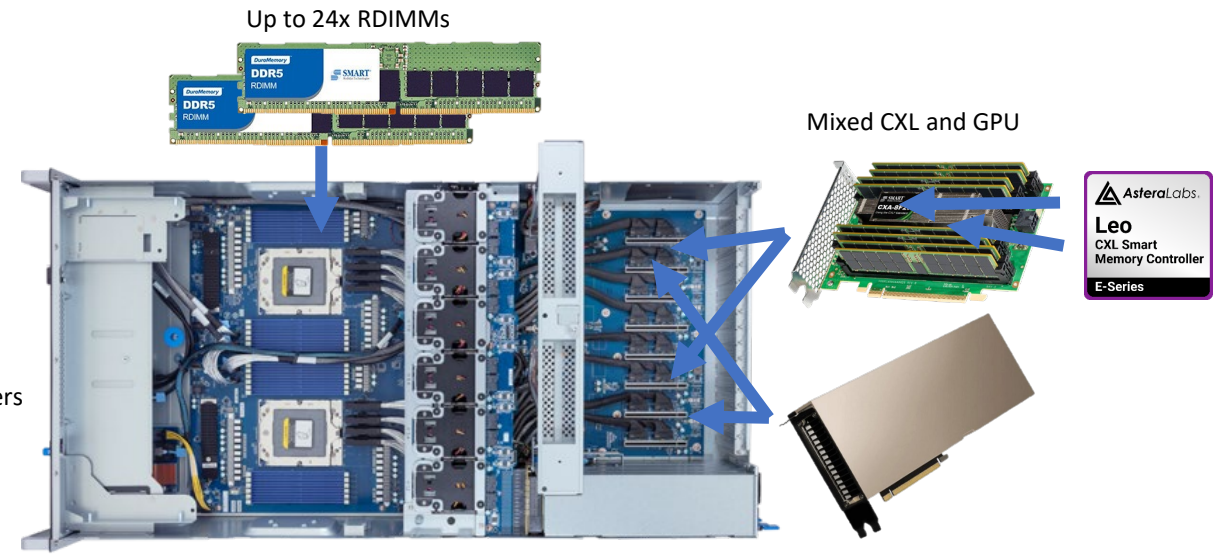
- CXL Expansion Add-in-cards provide the highest performance and memory density per PCIe slot
- Can support up to 1-2TB per card
- Early 8-DIMM use two controllers but plan to migrate to single controller in next generation
- CXL Retimer Add-in-cards enables robust, extended, and low-latency memory pooling

Scaling Server Memory Using CXL AICs

Maximum Memory Configuration



Mixed GPU-Memory Configuration



Penguin Solutions GPU Class 4U Server

	MOTHERBOARD MEMORY ONLY	MAX MEMORY WITH ALL CXL AICS	MAX MEMORY WITH 4X GPU
	24 RDIMMS + 64 CXL RDIMMS	24 RDIMMS + 64 CXL RDIMMS	24 RDIMMS + 32 CXL RDIMMS
64GB RDIMMS	1.536 TB	5.632 TB	3.584 TB
96GB RDIMMS	2.304 TB	8.448 TB	5.376 TB
128GB RDIMMS	3.072 TB	11.264 TB	7.168 TB
256GB RDIMMS	6.144 TB	22.538 TB	14.336 TB

Based on AMD Turin configuration.

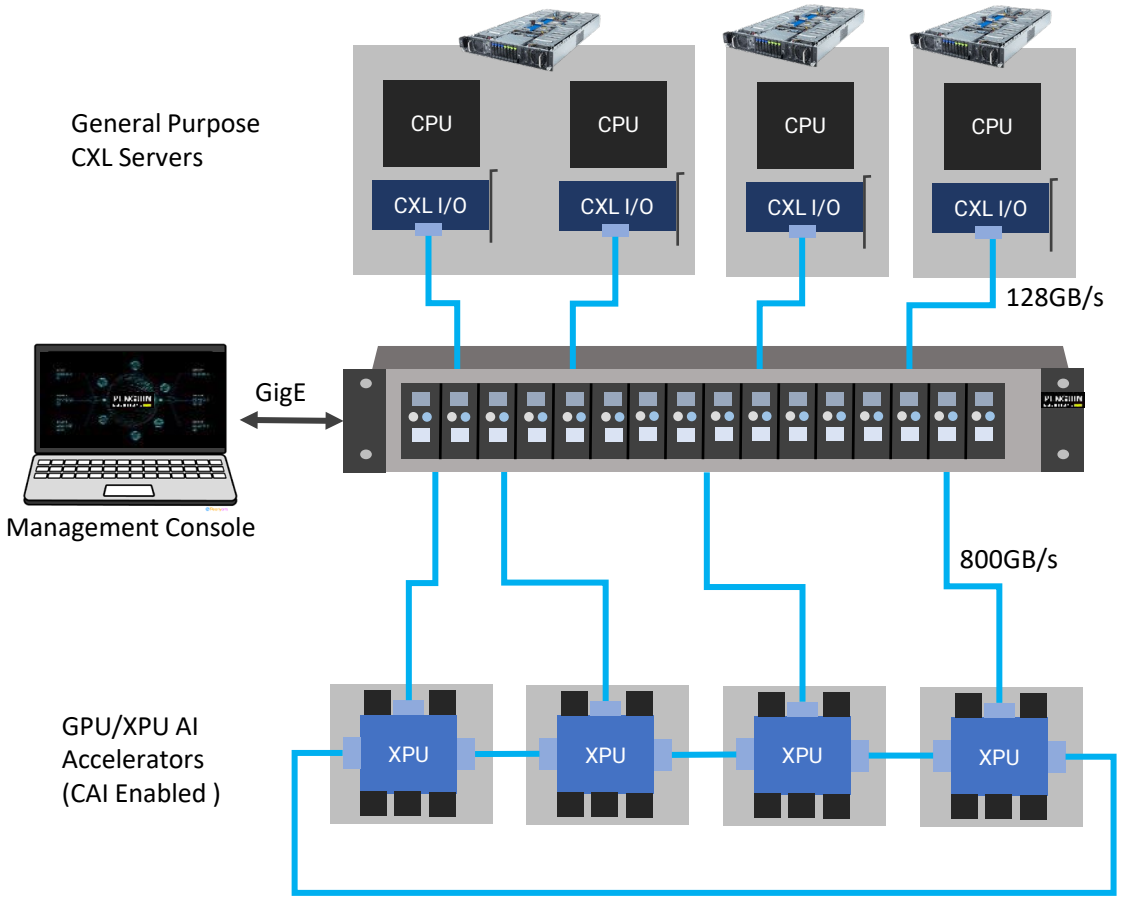
GPU servers from server suppliers provides the easiest adoption route for CXL in AI application

CXL is now enabled on many server SKUs

Enables large memory options for AI RAG, KV-cache, In-memory database and other memory heavy use cases

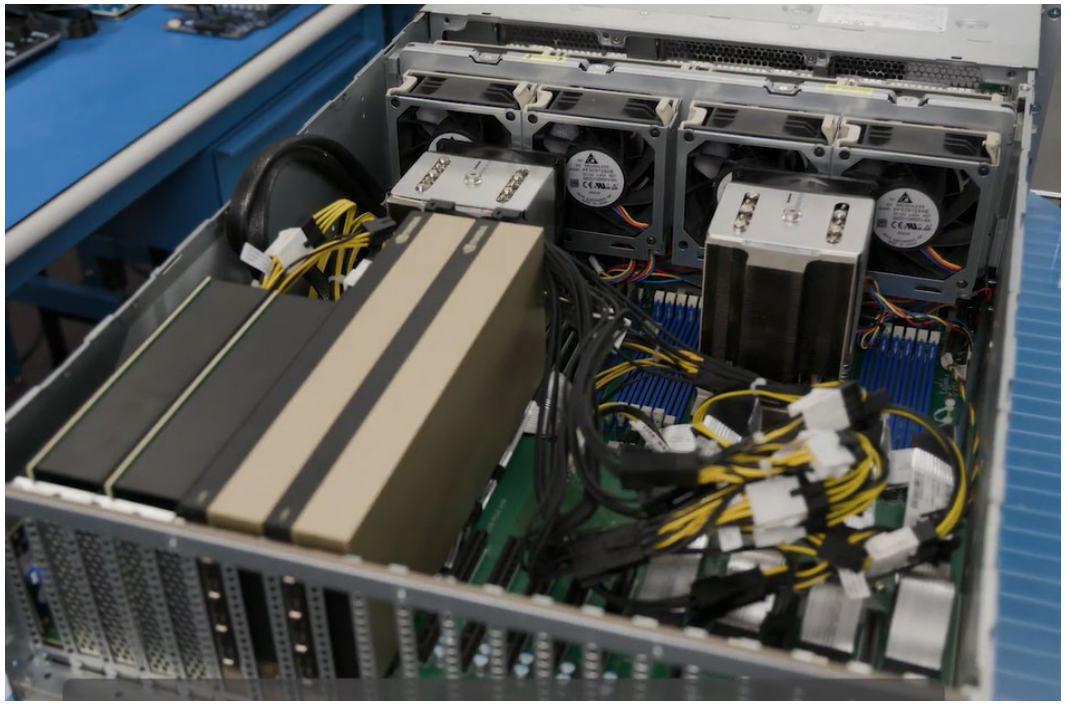
A Peek at the Future

Optical HBM Based Memory Pooling Appliance



- Share HBM Class Memory between multiple Servers and Custom XPU AI Accelerators
  - Significant reductions in data movement hence power
  - Support custom, CXL, UALink type links
- Example Memory Supported
  - Up to 32TB DDR5 RDIMMs
  - Up to 1TB HBM3E
- Example Memory Access Modes
  - HBM Direct Mode
  - CXL 3.2 Mode
  - Cached HBM-DDR mode
- Fabric Management Software
  - Statically or dynamically allocate memory to each XPU and/or CPU based on job needs

See the Demo – CXL Consortium Booth 725



AsteraLabs®

SMART® Modular Technologies

FMS the Future of Memory and Storage