SNIA | Compute, Memory, and Storage

Accelerating AI with Real-World CXL Platforms

A SNIA CMS Community Webinar
July 22, 2025, 9:00 am PT

Webinar access:
https://www.brighttalk.com/webcast/663/646407

Arthur Sainio
SNIA Persistent Memory Special Interest Group

Andy Mills
SMART Modular Technologies

Anil Godbole
Intel Corporation

Steve Scargall
MemVerge

Our Moderator and Presenters

Arthur Sainio

Co-Chair

SNIA Persistent Memory Special Interest Group

Steve Scargall

Director of Product Management for CXL and AI/ML

MemVerge

Anil Godbole

CXL Marketing Working Group Co-Chair/Senior Marketing Manager

Intel Corporation

Andy Mills

Vice President, Advanced Product Development

SMART Modular Technologies

SNIA | CMS A SNIA Community

The SNIA Community

200
Corporations, universities, startups, and individuals

2,500
Active contributing members

50,000
Worldwide IT end users and professionals

SNIA | CMS A SNIA Community

SNIA | Compute, Memory, and Storage

What We Do

- Engage technology users
- Educate on compute, memory, and storage technologies
- Accelerate SNIA standards
- Propel technology adoption

How We Do It

- Demonstrate at industry events
- Host Programming Memory Workshops and Hackathons
- Present educational webinars, podcasts, and blogs

Learn more at www.snia.org/groups/cms

SNIA Persistent Memory (PM) Special Interest Group (SIG)

- The place to go for information on:
  - Persistent Memory advances
    - www.snia.org/forums/cmsi/NVDIMM
  - PM and CXL Programming Workshops and Hackathons
    - www.snia.org/pmhackathon
- Why you should join
  - Engage and educate the industry
  - Expand reach of memory technologies to end users
- Learn more
  - askcms@snia.org

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be, construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

- The CXL Advantage for Real-World AI Workloads
- Boost Your AI Workload Performance Using CXL Memory
- Deploying CXL in Next Generation AI/ML Systems

SNIA | CMS
A SNIA Community

The CXL Advantage for Real-World AI Workloads

Steve Scargall

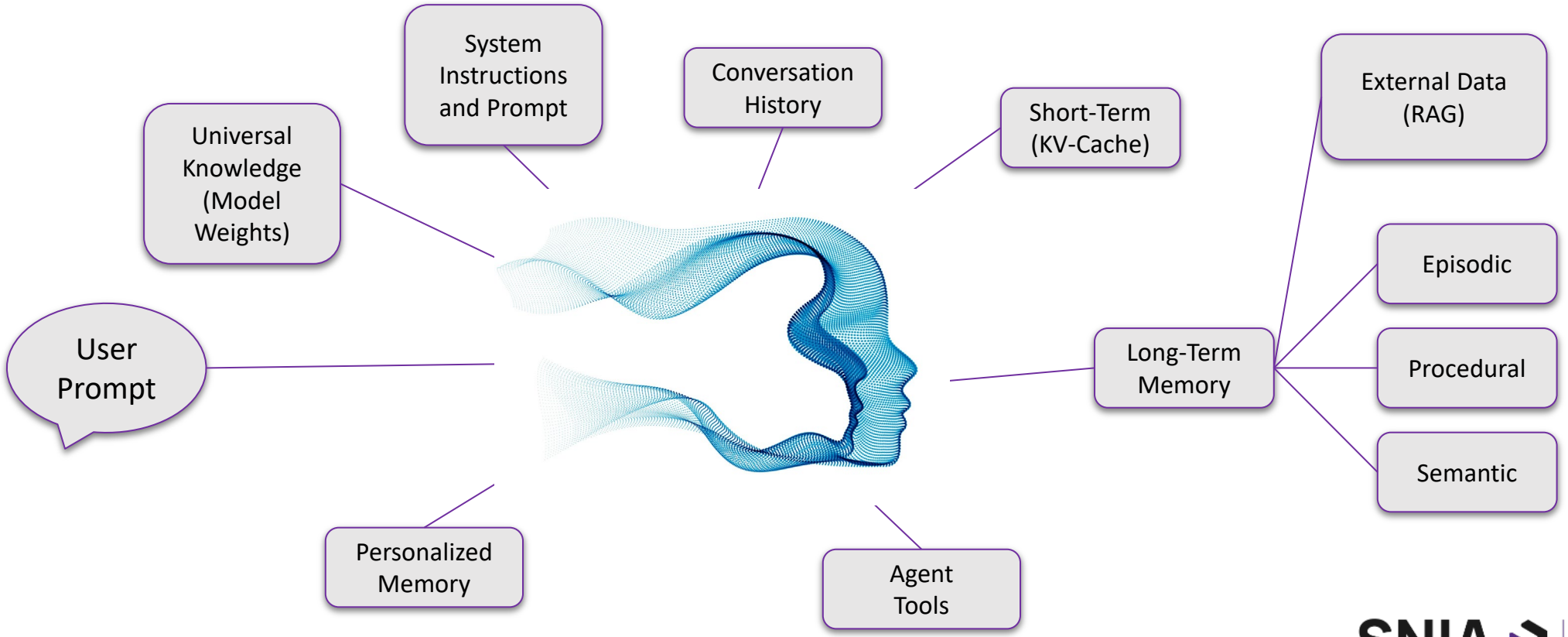
MemVerge

Accelerating AI: CXL and High-Capacity Memory Solutions

- In the rapidly evolving landscape of Artificial Intelligence, the demand for greater memory capacity and bandwidth is insatiable.
- Compute Express Link (CXL) is emerging as a pivotal technology in 2025, directly addressing the memory bottlenecks and capacity that can throttle AI workloads.

The Context Engineering Era: The CXL Opportunity

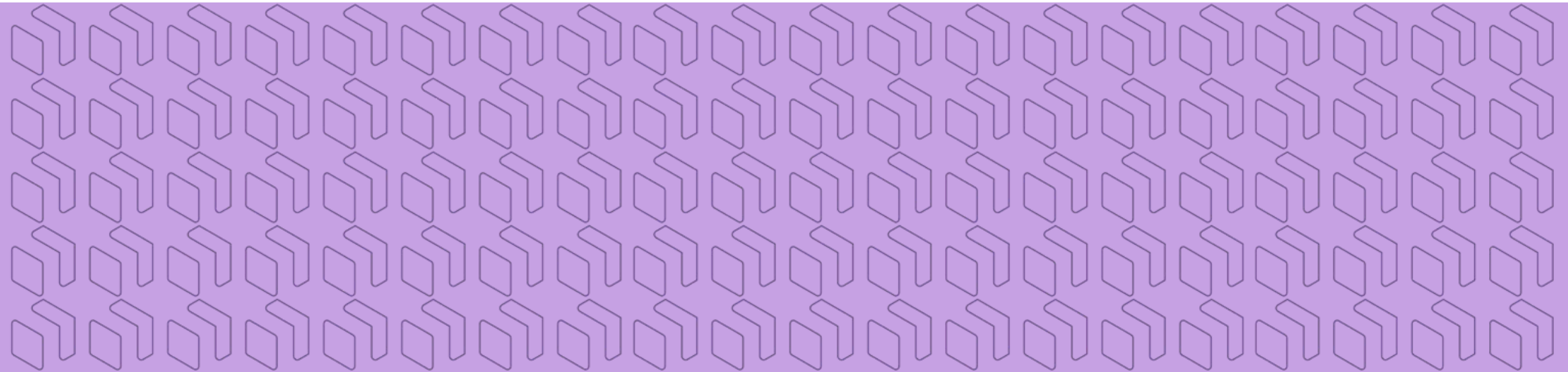
Context engineering is the practice of designing systems that provide large language models (LLMs) with the necessary information and tools to perform tasks effectively. It goes beyond simple prompt engineering, which focuses on crafting specific instructions. Instead, context engineering involves assembling a comprehensive context that includes:



SNIA | CMS A SNIA Community

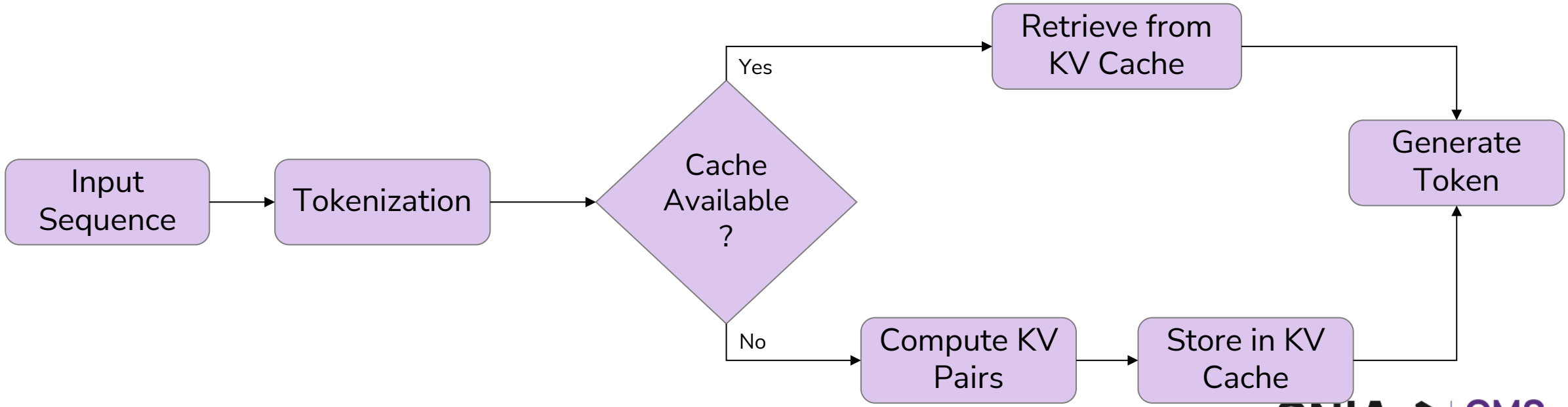
SNIA | CMS
A SNIA Community

KV-Cache: The Heart of LLM Inferencing



The Hidden Cost of Long Context: KV Cache

- **The KV-Cache Problem:**
  - Large Language Models (LLMs) rely on a Key-Value (KV) Cache to avoid re-computing attention for previous tokens.
  - For long prompts or conversations, this KV Cache can grow to hundreds of gigabytes, consuming precious and expensive GPU High-Bandwidth Memory (HBM).
  - This "memory tax" limits the number of users (batch size) a single GPU can serve, creating a throughput bottleneck.



How Much GPU Memory Do You Need for Efficient LLM Serving?

Total Required GPU Memory = (Model Weights) + (KV Cache) + (Activations) + (Overhead)

The main components contributing to it are:

- **Memory for Model Weights:** (Model Size in Parameters) x (Bytes per Parameter)
- **Memory for KV Cache:** (Batch Size) x (Sequence Length) x (2 tensors [K,V]) x (Num Layers) x (Hidden Size) x (Bytes per Parameter)
- **Memory for Activations:** This is temporary memory used during the model's forward pass, dependent on model architecture and sequence length.
- **System & Framework Overhead:** This is a typically fixed amount of memory reserved for CUDA kernels, the serving framework, and other system processes.

Note: Model Weights, Activations, and Overhead are fixed values. The KV-Cache is variable.

GPU Memory Requirements: Scaling Users

Scenario Assumptions

- Model: 7B Parameter LLM
- Precision: FP16 (2 bytes/param)
- Context Length: 8,192 tokens
- Architecture: 32 layers, 4096 hidden size

Static Memory Base

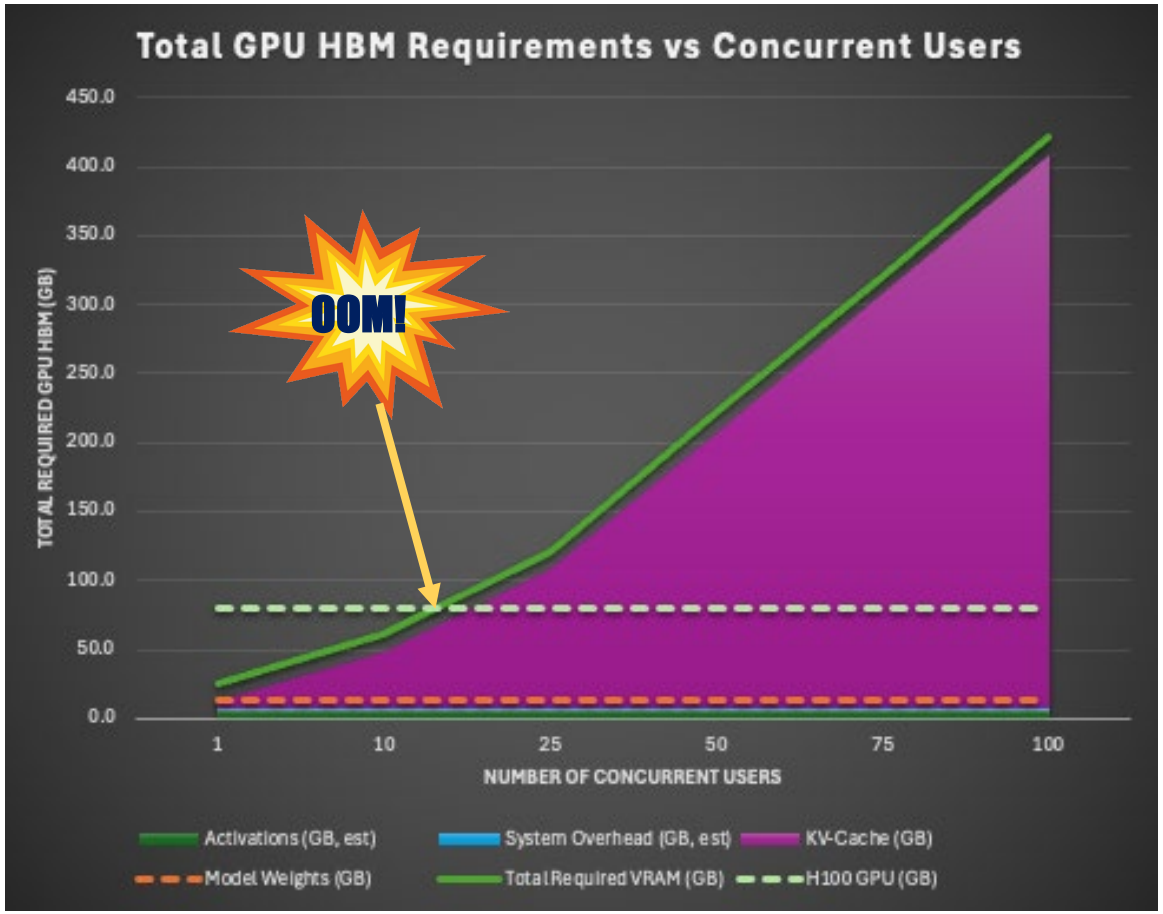
~21GB

The fixed cost of (Model Weights (~16GB)) + (Activations + Overhead (~5GB)).

Max Users on a single NVIDIA H100 with KV-Cache

up to ~16

100 users requires ~6 x NVIDIA H100 (80GB)

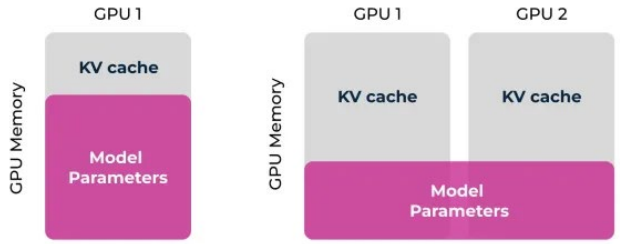


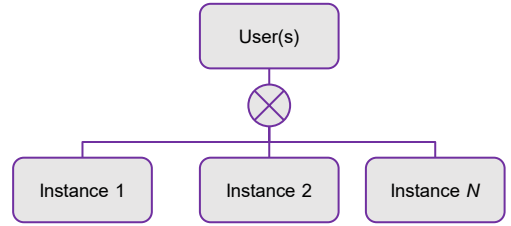
Key Takeaways:

- **KV Cache Dominates:** The memory required for the KV Cache quickly surpasses the size of the model weights, especially with a long context window and multiple users.
- **The Physical GPU Memory Limit:** Memory offloading and expansion strategies are critical for scaling.

Possible Solutions

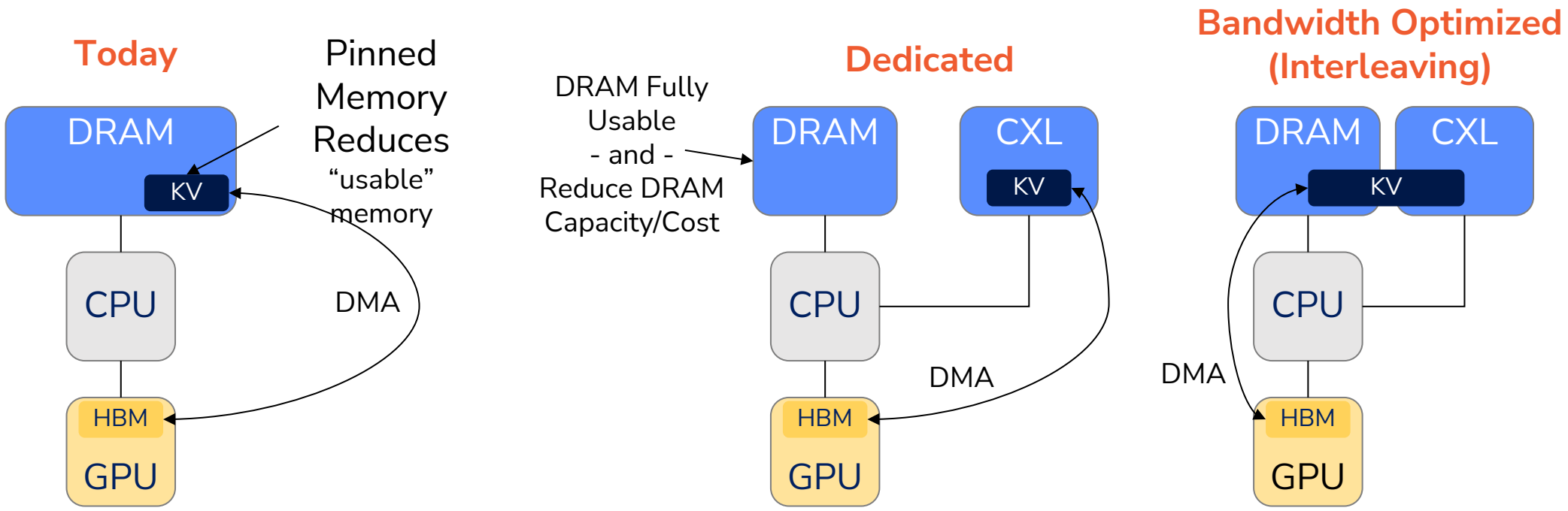
- Buy more GPUs ($$$$)
  - Pro: Increases HBM Capacity
  - Cons:
    - Expensive approach to increase HBM
    - Wasted GPU Tensor Compute Resources (idle)
- Quantize the Model
  - Pro: Reduces Model Parameters Memory Requirement
  - Cons: Reduces Accuracy
- Run multiple instances and route user sessions
  - Pro: Scale-out solution to handle user sessions
  - Con: Requires more Servers/Instances (power/cooling…)
- Offload KV-Caches (Memory Tiering)
  - Pro: Use the capacity and bandwidth of DRAM + CXL
  - Con: Higher Latency for warm memory access (relative to HBM)





KV Cache Offloading

- Use DRAM & CXL memory as a vast, secondary tier for the KV Cache.
- The CXL-to-GPU interconnect offers low latency and high bandwidth, making it a perfect fit for this "warm" tier of data.



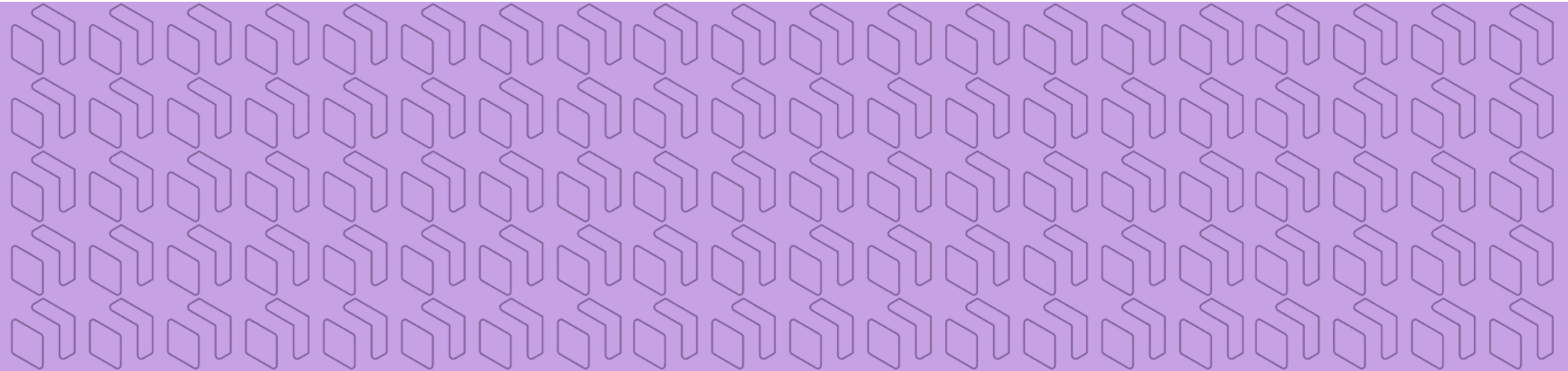
SNIA | CMS A SNIA Community

KV Cache Offload Benefits

- **Massively Increased Batch Size:** Freeing up GPU HBM allows for significantly more concurrent users.
- **Reduced TCO:**
  - Serve larger models with fewer, more efficiently utilized GPUs.
  - Reduce the number of GPUs
- **Enable Future Models:** Provides a path to serve future models with even larger context windows.

SNIA | CMS
A SNIA Community

Improving Distributed LLM Inference @Scale



Distributed Inferencing

The Problem & Challenges @ Scale:

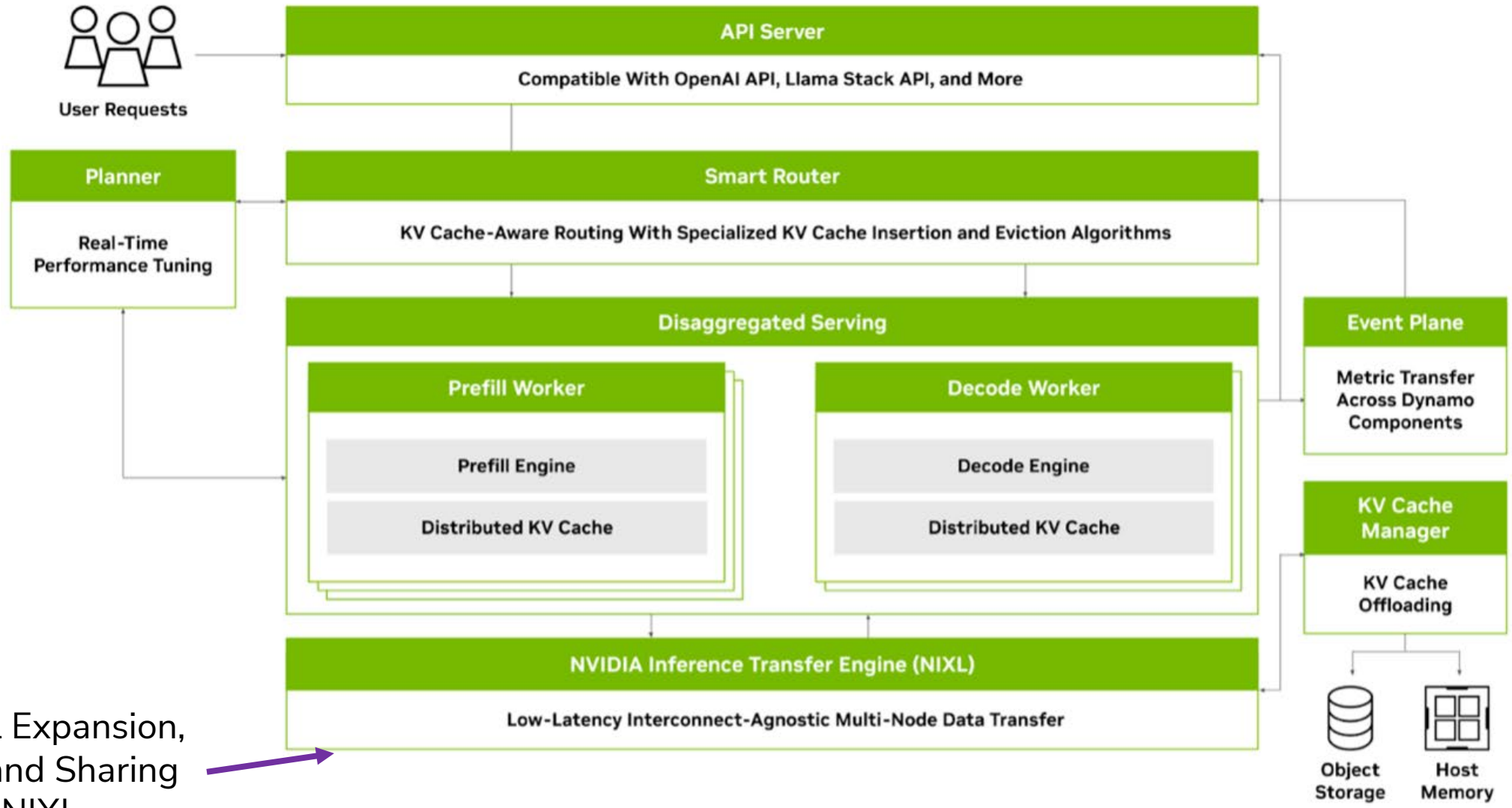
- Scaling inference for generative AI and reasoning models presents complex challenges in three key areas: performance, correctness, and efficiency.
  - ***GPU underutilization***: Traditional monolithic inference pipelines often leave GPUs idle due to the imbalance between prefill and decode stages.
  - ***Expensive KV cache re-computation***: When requests aren't efficiently routed, KV caches often get flushed and recomputed, leading to wasted computation cycles and increased latency.
  - ***Memory bottlenecks***: Large-scale inference workloads demand extensive KV cache storage, which can quickly overwhelm GPU memory capacity.
  - ***Inefficient data transfer***: Distributed inference workloads introduce unique and highly dynamic communication patterns that differ fundamentally from training.

Distributed Inferencing

The Solution:

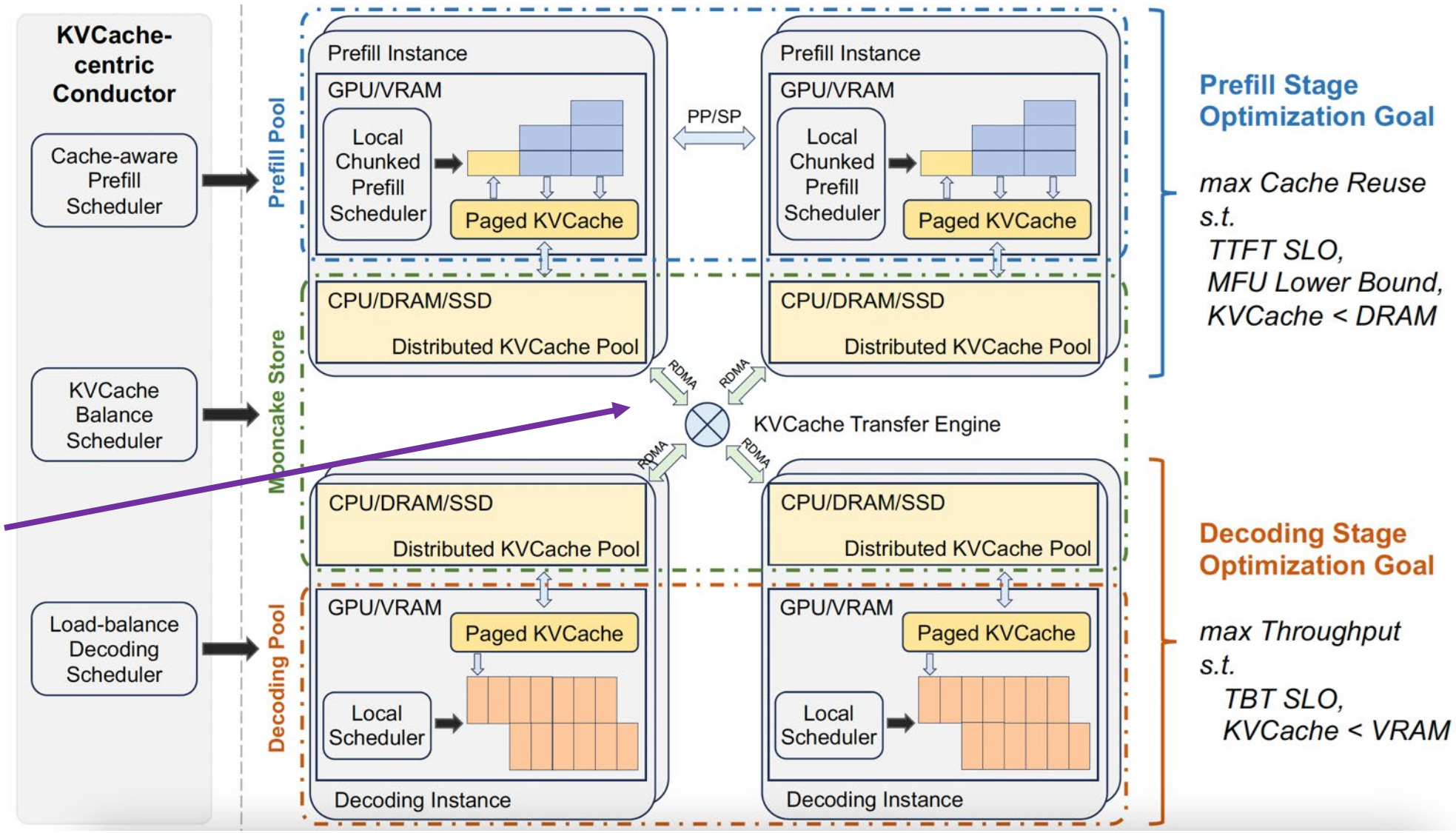
- Leading-edge serving platforms, such as NVIDIA Dynamo, LLM-D, VLLM, and Mooncake, are built on the principle of disaggregation.
- For efficient global memory management at scale, the platforms strategically store and evict KV caches across multiple memory tiers—GPU, CPU, SSD, and object storage - enhancing both time-to-first-token and overall throughput.
- CXL Fabrics can facilitate high-throughput, low-latency, and efficient transfer and interconnect between inference instances, leading to higher optimization and throughput of tokens per second.

NVIDIA Dynamo & NIXL



Add CXL Expansion, PMem, and Sharing in NIXL

Mooncake Architecture



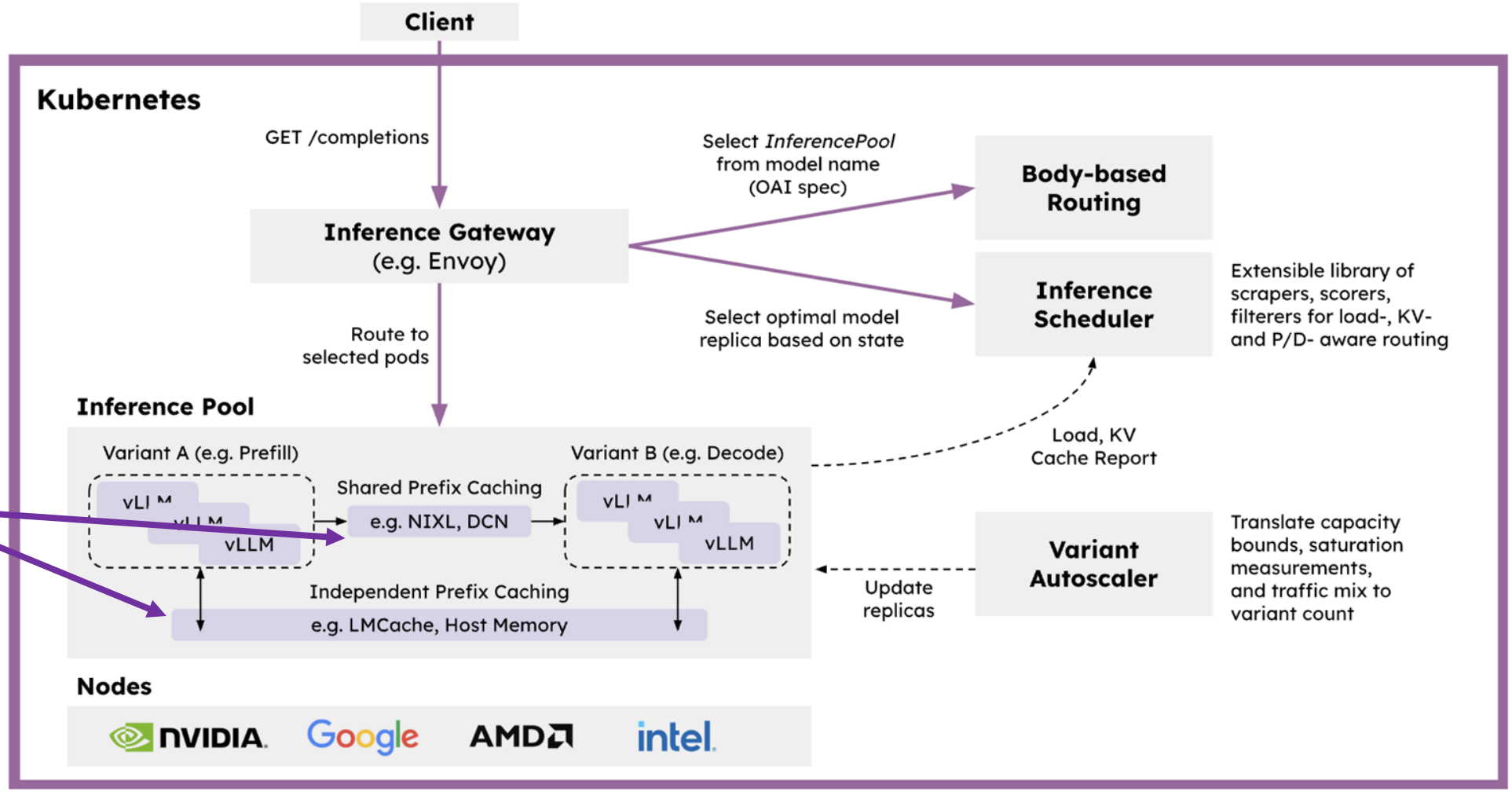
Replace RDMA with CXL Sharing

https://github.com/kvcache-ai/Mooncake

SNIA CMS A SNIA Community

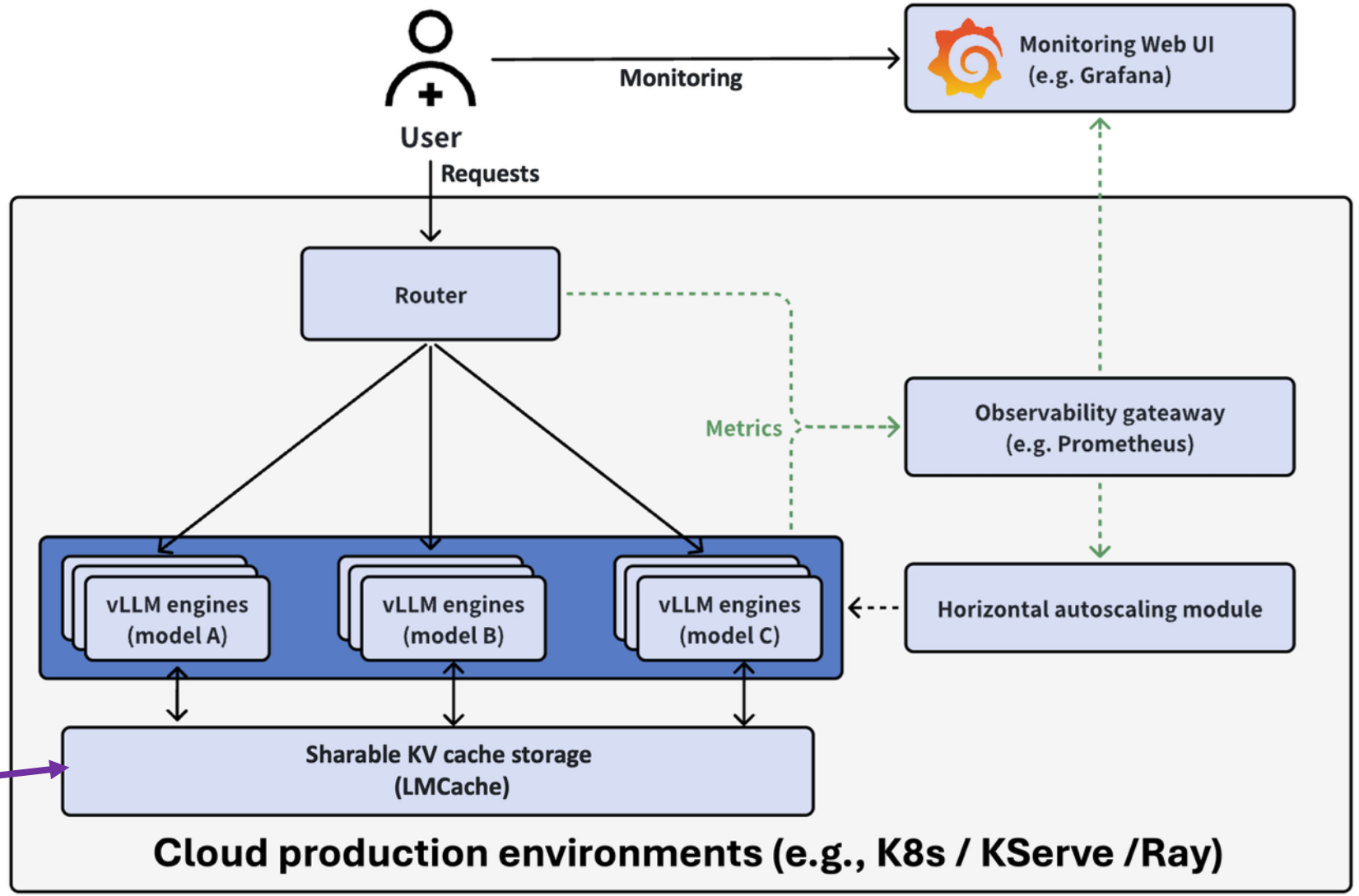
LLM-D

Utilize CXL Expansion & Sharing



SNIA CMS A SNIA Community

VLLM Architecture



Backed by
DRAM & CXL

Source: https://docs.vllm.ai/projects/production-stack/en/latest/getting_started/installation.html

Takeaways

- Context Engineering is happening now, requiring high memory capacity and bandwidth - A great opportunity for CXL.
- KV-Cache dominates the GPU memory requirements @ scale
- Offloading the KV-Cache and/or Model Parameters is an efficient approach to scaling
- With modifications, existing LLM Inference frameworks can adopt CXL Expansion and Sharing

SNIA | CMS
A SNIA Community

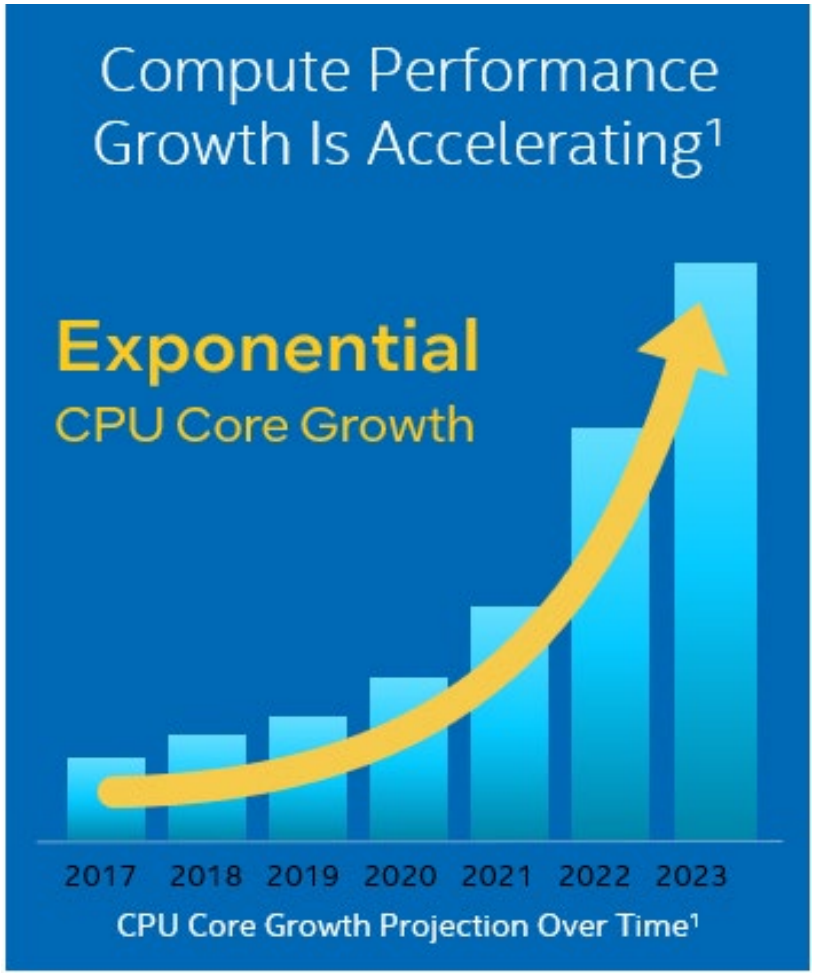
Boost Your AI Workload Performance Using CXL Memory

Anil Godbole

intel

Compute Core Count Keeps Increasing

- Needed to keep up with memory intensive workloads
- W/L Examples
  - Virtualized servers
  - In-memory data bases
  - AI/ML
  - Many others…



1 Source: Intel. Results may vary.

Value Prop of CXL-attached Memory

Increased Memory Capacity

Improve processor perf

- Reduced SSD accesses

Benefitting Workloads

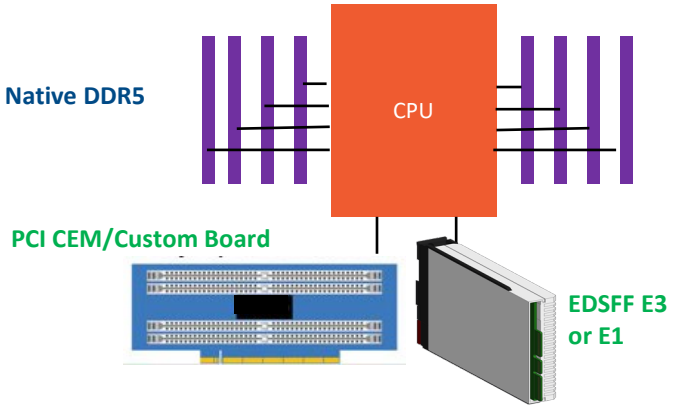
- Virtualized Servers
- In-memory Databases
- AI / ML
- HPC (High Perf Computing)
- Media (CDN, Video 8K)
- Medical (Genomics)

Increased Memory Bandwidth

Improve processor's memory bandwidth using address interleaving

Benefitting Workloads

- AI/ML
- HPC
- Non-relational Databases



Lower Memory TCO

Avoid expensive 3DS DIMMs

- Use standard DIMM capacities for native & CXL
- Use lower-cost memory media on CXL
  - DDR4

Memory Pooling

- For optimal provision of local DRAM on servers

TCO Savings Examples with CXL Memory
Avoid high-cost DIMMs

- 128GB and 256GB DIMMs have high price premiums
- CXL add-in cards with DIMM slots provide more total channels per socket → same system capacity with lower priced DIMMs

Memory per socket	Socket-attached DIMMs only	With CXL	Memory TCO Savings*
1TB	8x 128GB (1DPC)	8x 64GB + 8x 64GB on CXL	24%
2TB	16x 128GB	16x 64GB + 16x 64GB on CXL	27%
4TB	16x 256GB	16x 128GB + 16x 128GB on CXL	16%

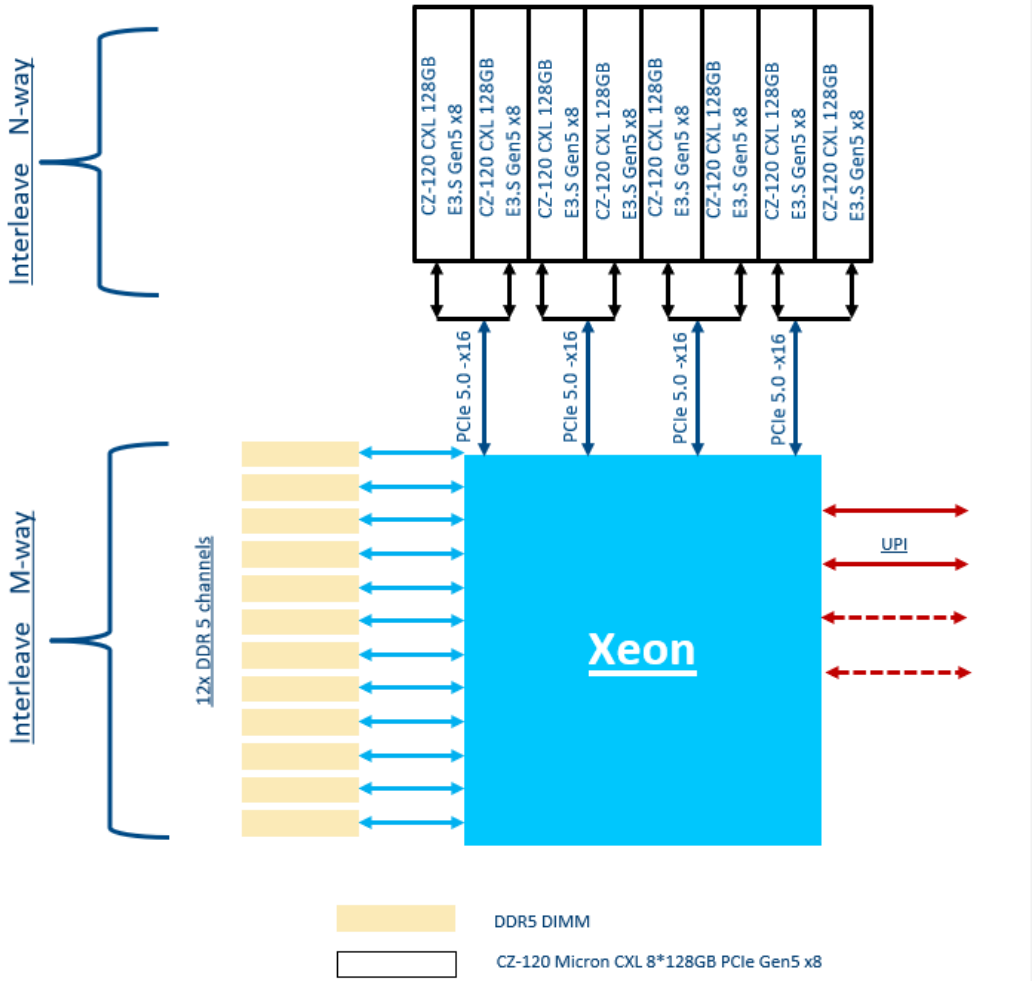
*TCO savings based on Intel modeling using projected DIMM and CXL pricing for 2025

*TCO savings even more if DDR4 RDIMMs are reused using AICs

Use CXL-attached DIMMs to achieve high system memory capacity:
Avoid expensive high-capacity DIMMs

S/W-Assisted B/W-Weighted Memory Interleaving

- SW (Hypervisor/OS/App) responsible for tiering & interleaving*
- Systems boots as two-tier memory (Near & Far)
- S/w ‘stripes’ pages between native & CXL memory
  - Uses page-table entries to assign physical addresses to virtual address pages
- Page-striping ratio (‘M:N’)
  - No. of pages in native DRAM / No. of pages in CXL memory)
  - Typically based on ratio of native DRAM memory wrt CXL memory b/w
  - But completely flexible for S/W to choose
- No page movement involved
  - Pages remain ‘pinned’ in their respective memories



*Upstreamed to Linux with Kernel v6.9 & beyond

Bandwidth Expansion with DDR5 + CXL Memory

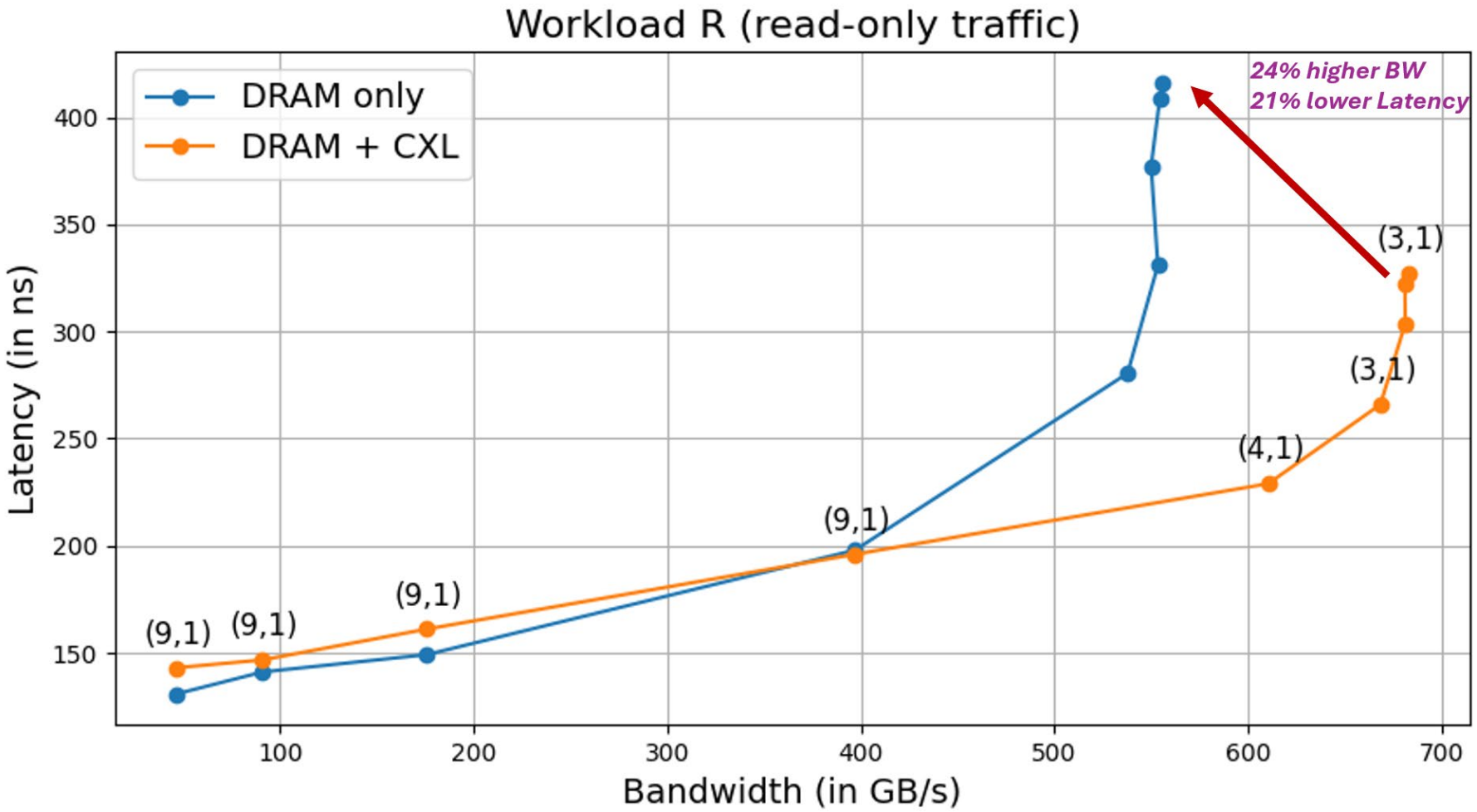
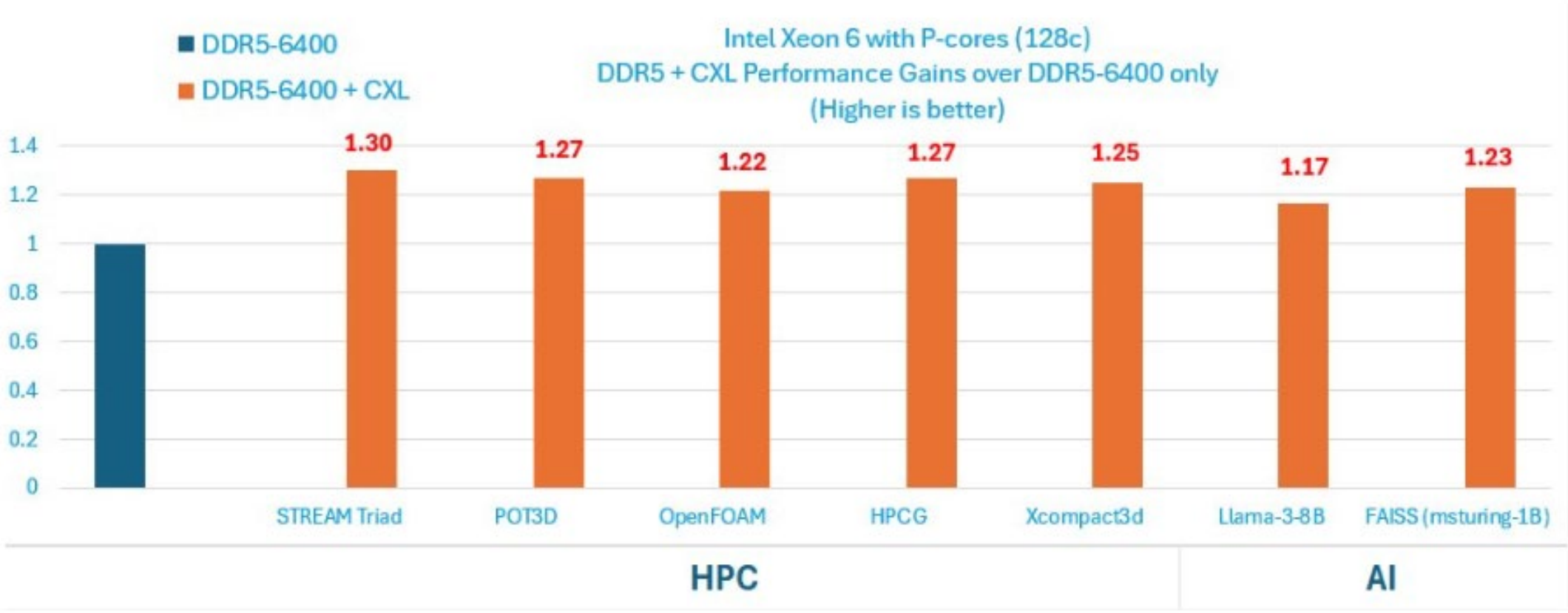


Chart with Intel Xeon 6900 with Micron DDR5 DIMMs & CXL CZ-120 modules

Interleaving weights given by (M,N) pairs

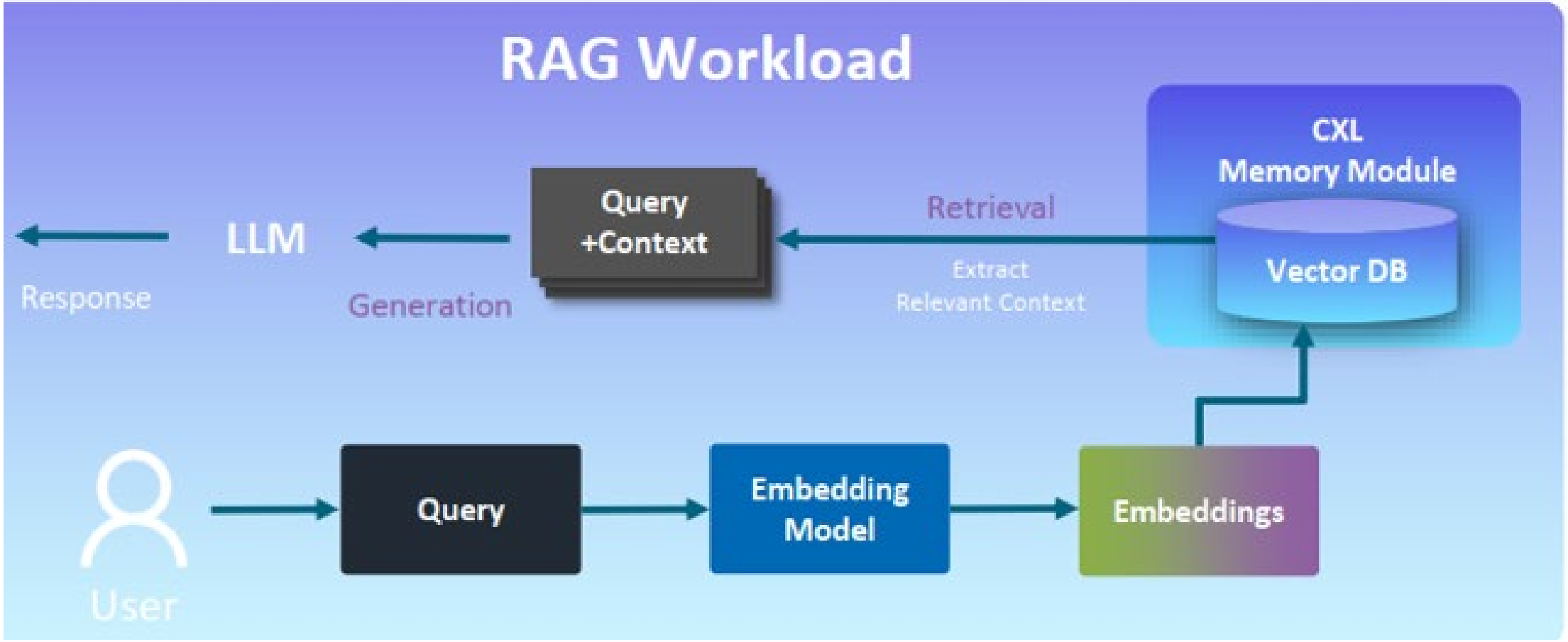
SNIA CMS A SNIA Community

HPC W/Ls Performance Examples

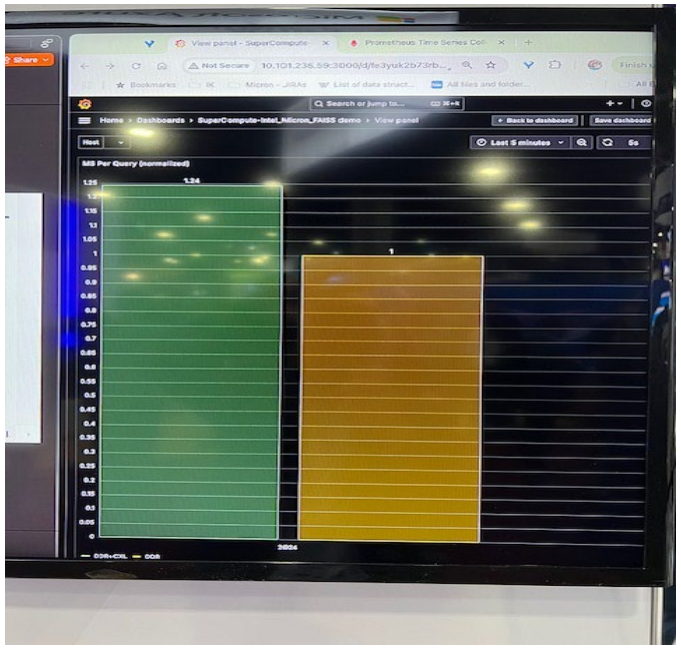


https://community.intel.com/t5/Blogs/Tech-Innovation/Data-Center/Improve-your-HPC-and-AI-workload-performance-by-increasing/post/1647882

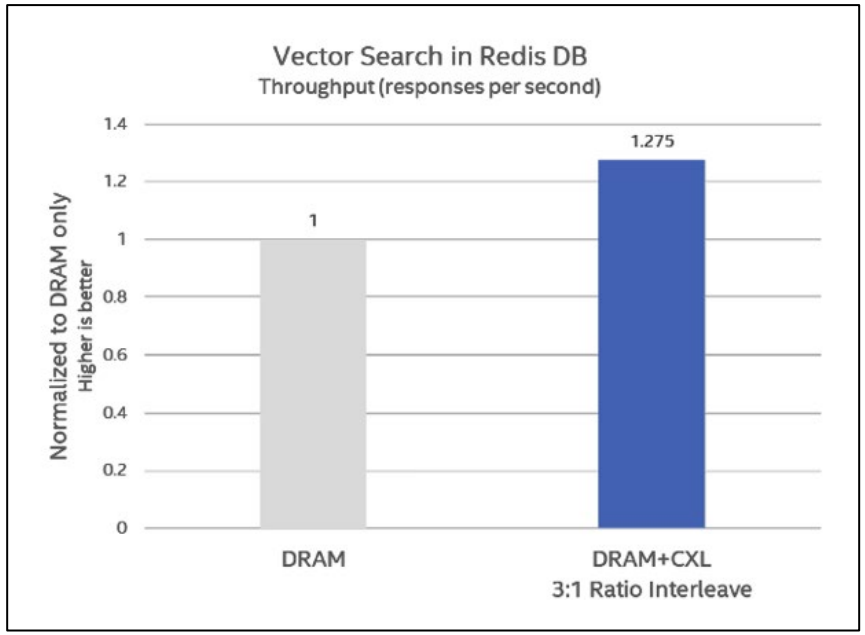
RAG (Retrieval Augmented Generation) based AI



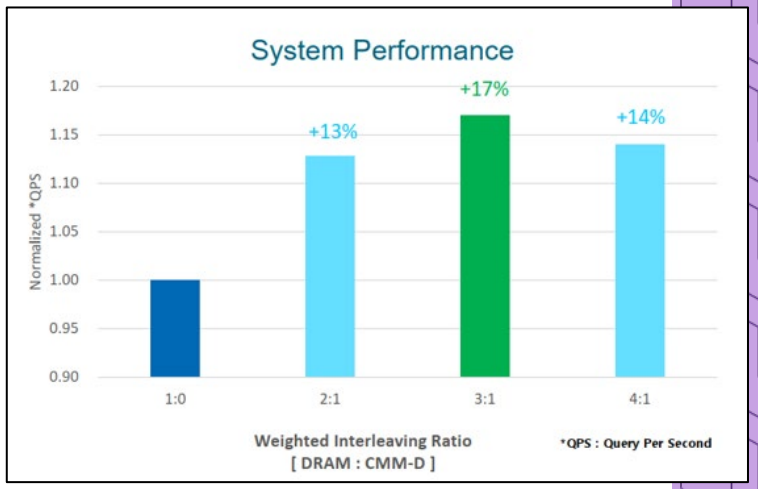
In-Memory RAG Database Performance Gains



23%+ for FAISS Vector Db
CPU: Intel Xeon-6 6900 w/ 128 cores
CXL: Micron CZ122 128GB



27%+ for Redis Vector Db
CPU: Intel Xeon-6 6787P w/ 86 cores
CXL: 2x x16 Astera Leo D5
DRAM: Hynix 64GB RDIMMs



17% Put Perf for Milvus DB
CPU: Intel Xeon-6 6767 w/ 64c
CXL: Samsung CMM-D 256GB

SNIA | CMS A SNIA Community

Intel CXL Enablement Roadmap

4th & 5th Gen Intel® Xeon® Gen4 (SPR) / Gen5(EMR) Eagle Stream Platform)	6th Gen Intel® Xeon® CPU Gen6 (GNR, SRF) Birch Stream Platform*	Future Gen Intel® Xeon® CPU
▪ Supports CXL v1.1 spec ▪ Leadership in CXL ecosystem enablement	▪ Supports CXL v2.0 spec ▪ Enhanced support for CXL Memory	▪ Support for CXL v3.x

Intel Xeon Roadmap Fully Aligned with CXL Roadmap

*Recommend using SKUs at HCC or above

Summary

- CPUs will play a big role in the AI revolution in the coming years

- There are many AI workloads like RAG, small LLM inferencing where CPUs can do the job more economically
  - Without needing a GPU
  - With CXL memory giving a boost
- Intel Xeon 6900/6700/6500 CPUs and coupled with CXL ecosystem of both h/w devices & s/w is available for your AI workloads

SNIA | CMS
A SNIA Community

Deploying CXL in Next Generation AI/ML Systems

Andy Mills

PENGUIN COMPUTING

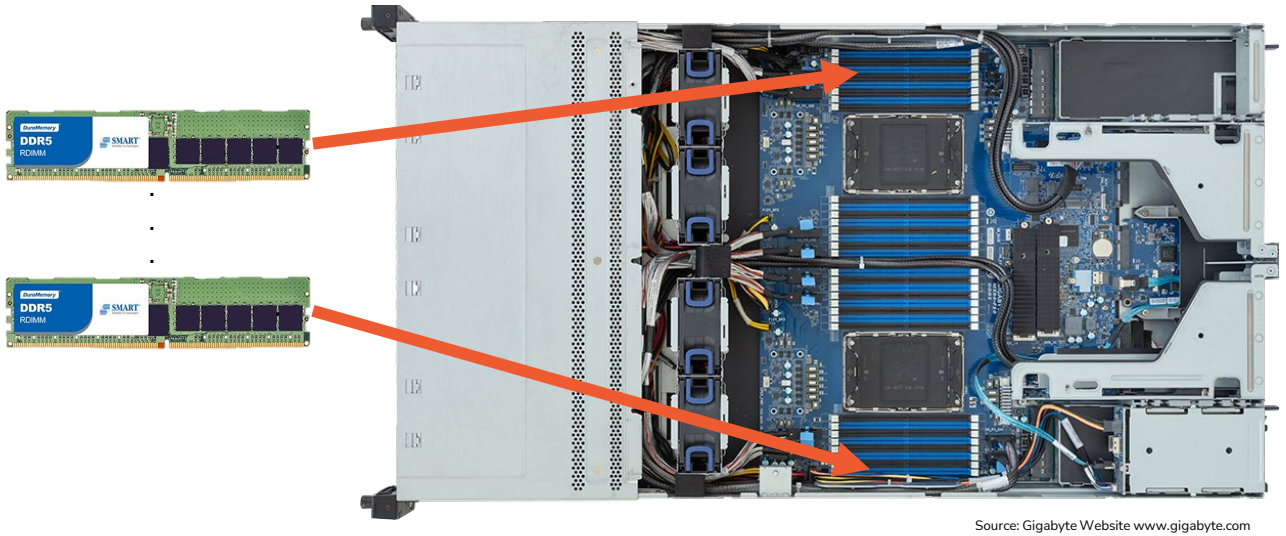
Scaling Memory in AI Servers

- Challenges deploying large memory footprints today
  - Limit to number of RDIMMs per CPU = capacity limit
  - Limit to HBM directly addressable by GPUs = capacity limit
  - Expansion by networking multiple CPUs or GPUs together with a high-performance network = capacity gains, new network performance limits
  - More CPUs + GPUs + network hardware = significantly higher power consumption

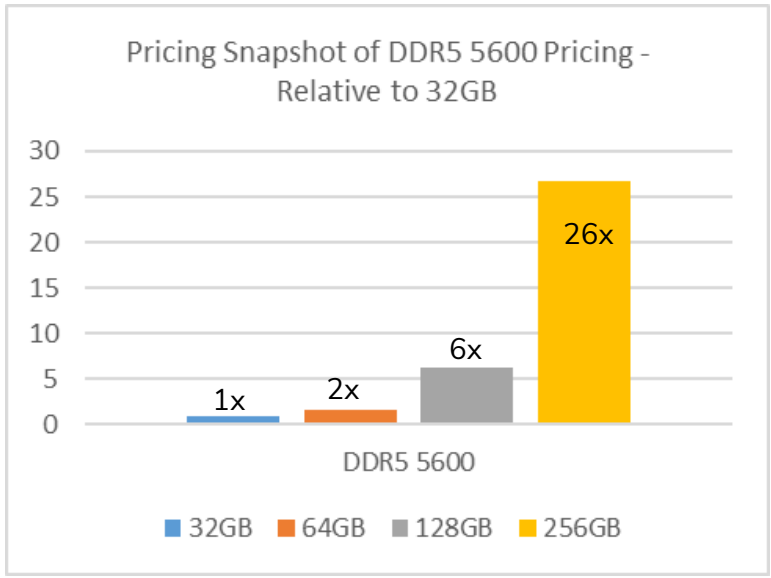
- Industry needs to solve for higher performance, higher capacity memory per CPU and GPU, while minimizing power consumption, heat generation, etc

Scaling Memory Today – Single Server Unit

Typical 2U Server Limits # of RDIMMs



Source: Gigabyte Website www.gigabyte.com



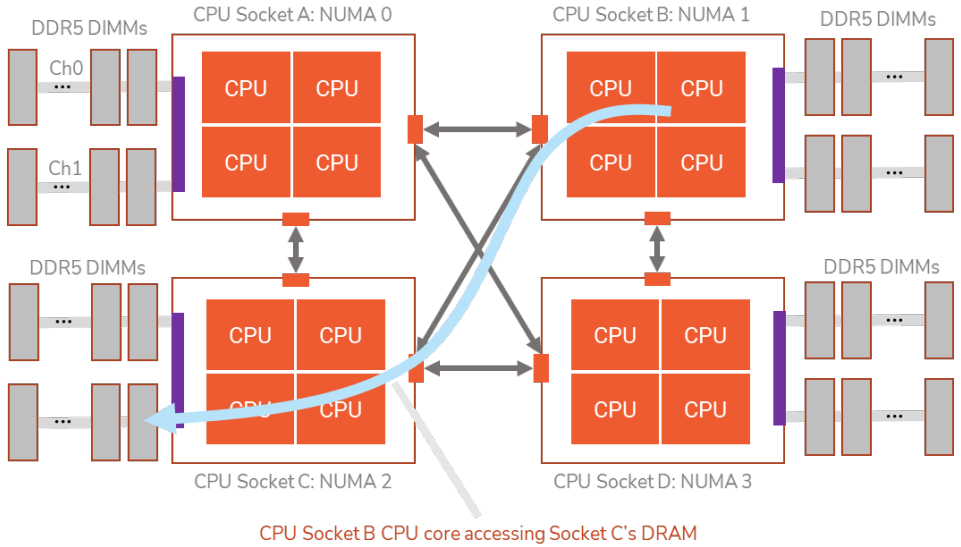
Non-Linear Price Scaling of RDIMMs

Maximum Memory Capacity of Mainstream 2U 2-socket Servers

	AMD TURIN (12 PER CPU, 1DPC)	INTEL GRANITE RAPIDS SP (16 PER CPU, 2DPC)
64GB RDIMMS	1.536TB	2.048TB
96GB RDIMMS	2.304TB	3.072TB
128GB RDIMMS	3.072TB	4.096TB
256GB RDIMMS	6.144TB	8.192TB

Traditional Methods for Overcoming CPU Memory Limits

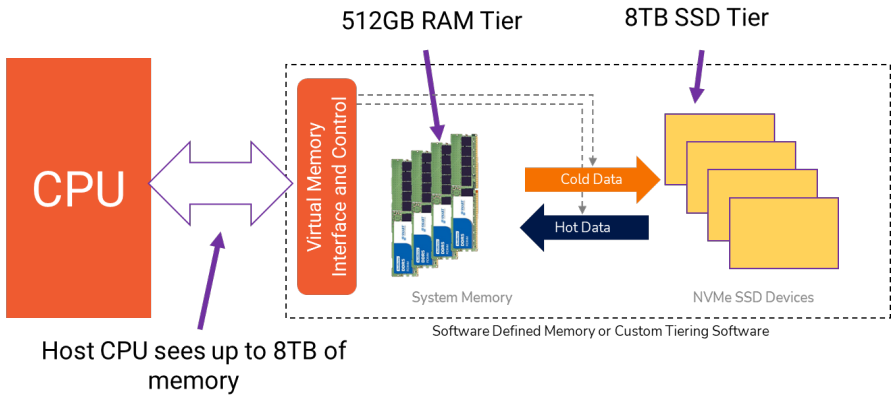
Increase number of CPUs e.g. 4-socket



Pro:
- **Doubles size of memory via CPU expansion**
- **NUMA allows access to all memory via high-speed CPU interconnects**

Con:
- **Expensive if CPUs are simply being used a "memory controllers"**

Utilize Software Defined Memory Tiering with SSDs



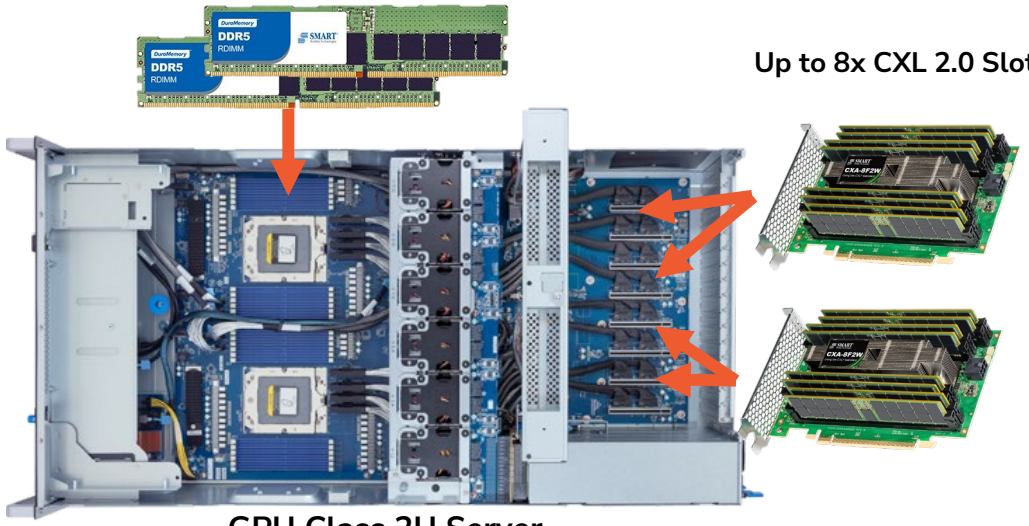
Pro:
- **Significantly increases size of virtual memory**
- **SSDs provide a low-cost expansion method**

Con:
- **Heavy performance tax if/when data hits SSD "memory tier"**
- **SSD tier runs at up to 1000x slower than main memory tier**

SNIA | CMS A SNIA Community

Scaling Server Memory Using CXL AICs

Up to 32x RDIMMs



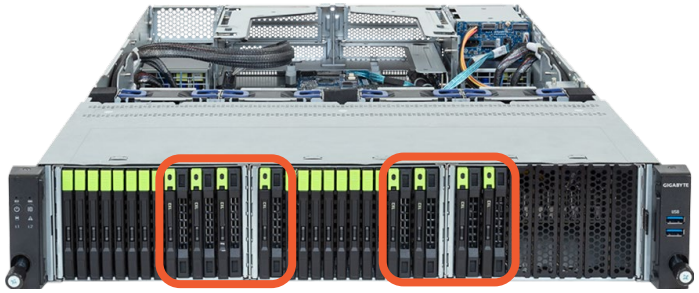
Up to 8x CXL 2.0 Slots

GPU Class 2U Server

Source: Gigabyte Website www.gigabyte.com

E3 CXL 2U Server

Source: Gigabyte Website www.gigabyte.com



Volatile CXL Memory Module

Persistent CXL Memory Module

8x E3.S CXL Memory Modules (CMMs)

	AMD TURIN	INTEL GRANITE RAPIDS SP
	24 RDIMMS + 64 CXL RDIMMS	32 RDIMMS + 64 CXL RDIMMS
64GB RDIMMS	5.632TB	6.144TB
96GB RDIMMS	8.448TB	9.216TB
128GB RDIMMS	11.264TB	12.288TB
256GB RDIMMS	22.538TB	24.576TB

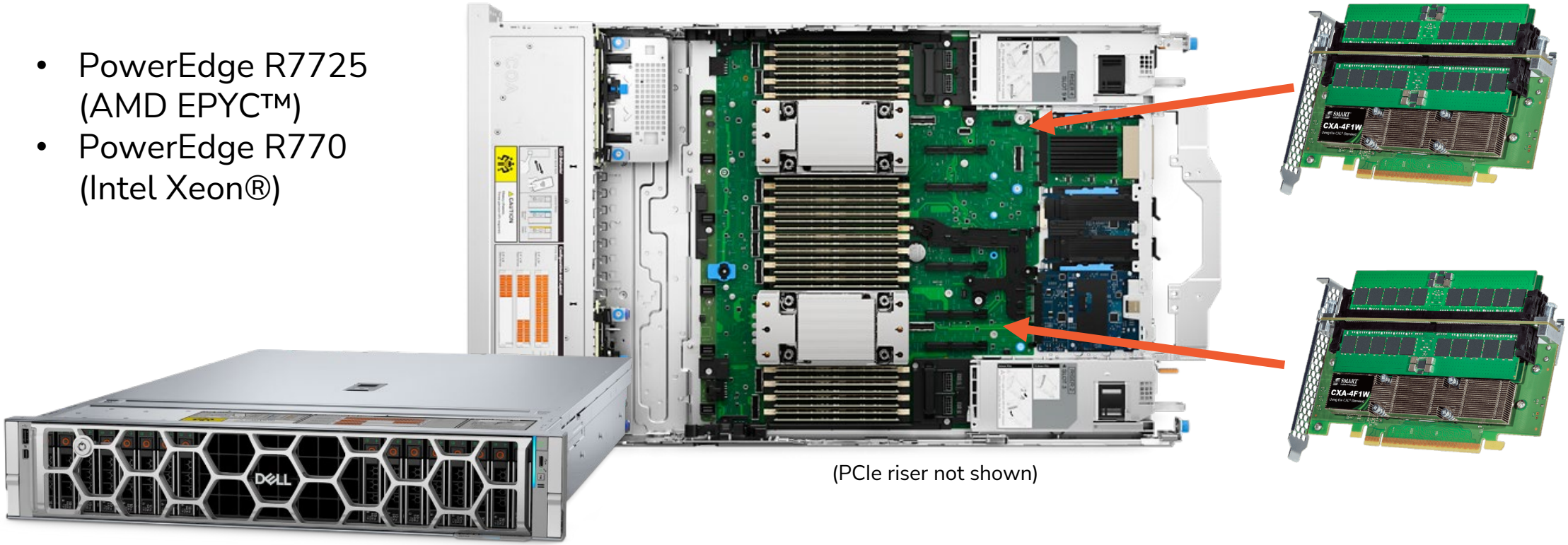
	AMD TURIN	INTEL GRANITE RAPIDS SP
	24 RDIMMS + 8 CXL CMMS	32 RDIMMS + 8 CXL CMMS
64GB RDIMMS	2.048TB	2.560TB
96GB RDIMMS	3.072TB	3.840TB
128GB RDIMMS	4.096TB	5.120TB
256GB RDIMMS	8.192TB	10.240TB

SNIA CMS A SNIA Community

OEM Server Adoption

Dell now offering CXL enabled servers with Dell CXL add-in-cards

- PowerEdge R7725 (AMD EPYC™)
- PowerEdge R770 (Intel Xeon®)

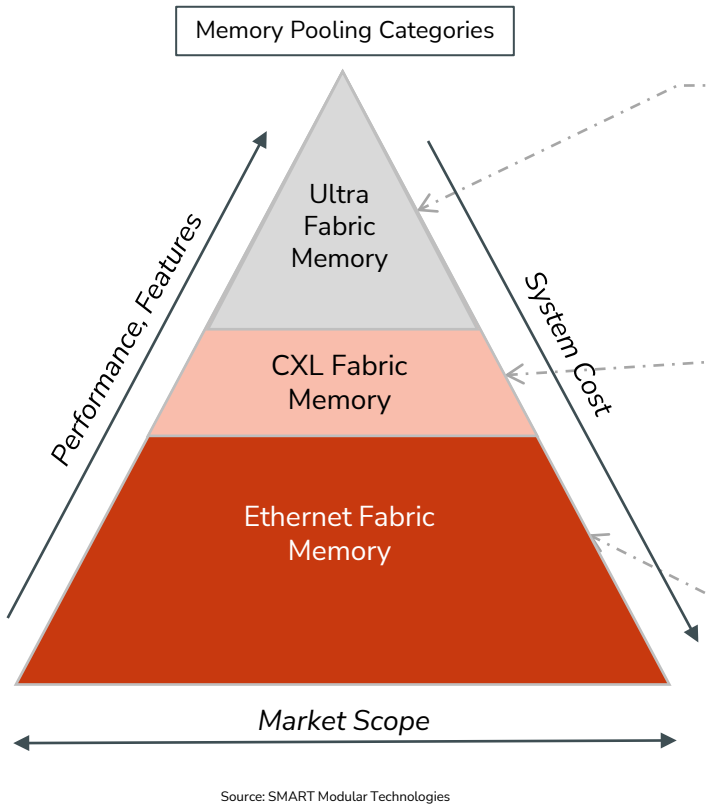


(PCIe riser not shown)

Source: Dell Website (www.dell.com)

SNIA | CMS A SNIA Community

Memory Pooling Product Segments



Source: SMART Modular Technologies

Ultra Fabric Memory (UFM)

- HBM and DDR5 Memory Types
- CPU and XPU/GPU attach options
- CXL 3.1, HBM, memory fabric native protocols
- 800GB/s-1.6TB/s per port, low latency
- Custom all-optics (Silicon Photonics) Cabling and Switches, with Future Standards path to UALink/other Fabrics

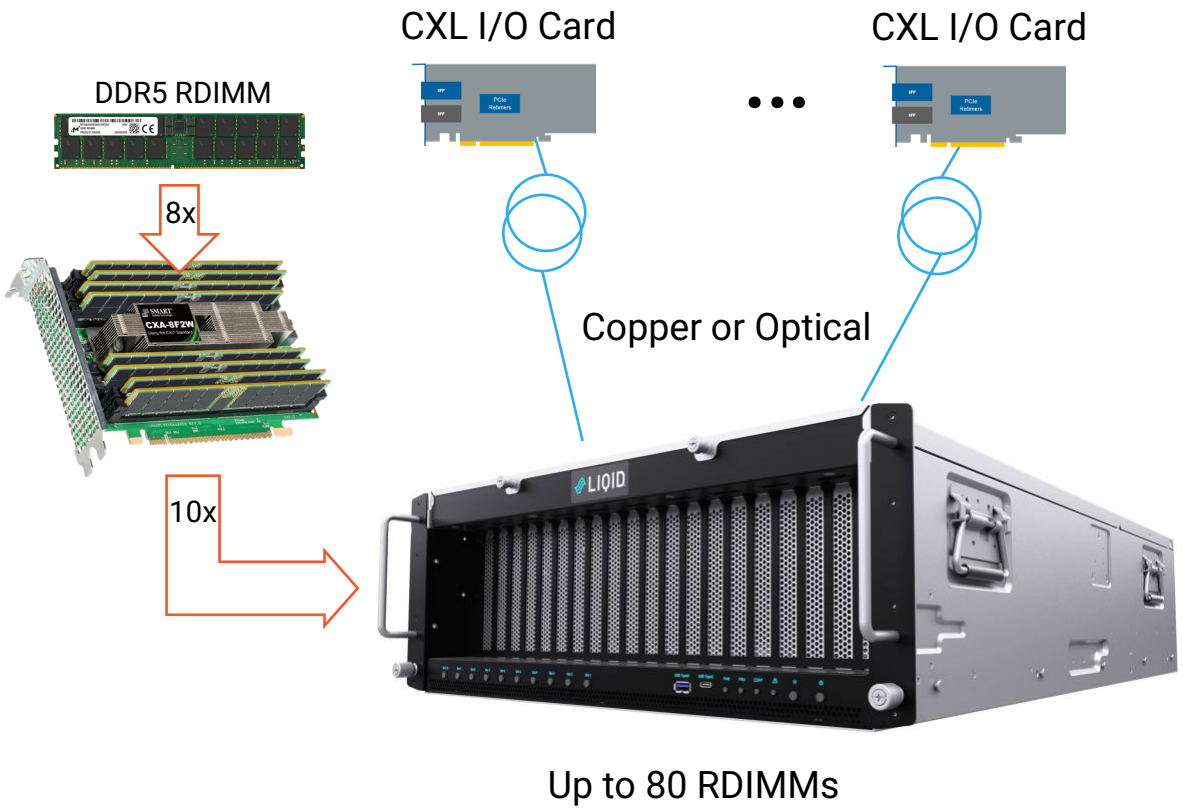
CXL Fabric Memory (CFM)

- DDR5 Pooling
- CPU attach only
- CXL 2.0/3.1 protocol
- 64-128GB/s per port, standard latency
- Standards Based PCIe SIG Cabling and Switches

Ethernet Fabric Memory (EFM)

- DDR5 Pooling
- Ethernet attached
- Memory as a service, CXL Over Ethernet
- Up to 800Gb/s per port, high latency
- Standard Based Ethernet Cabling and Switches

First Generation CXL Fabric Pooling



For RAG Applications, sharing Vector dBases between CPUs helps reduce fragmentation overhead

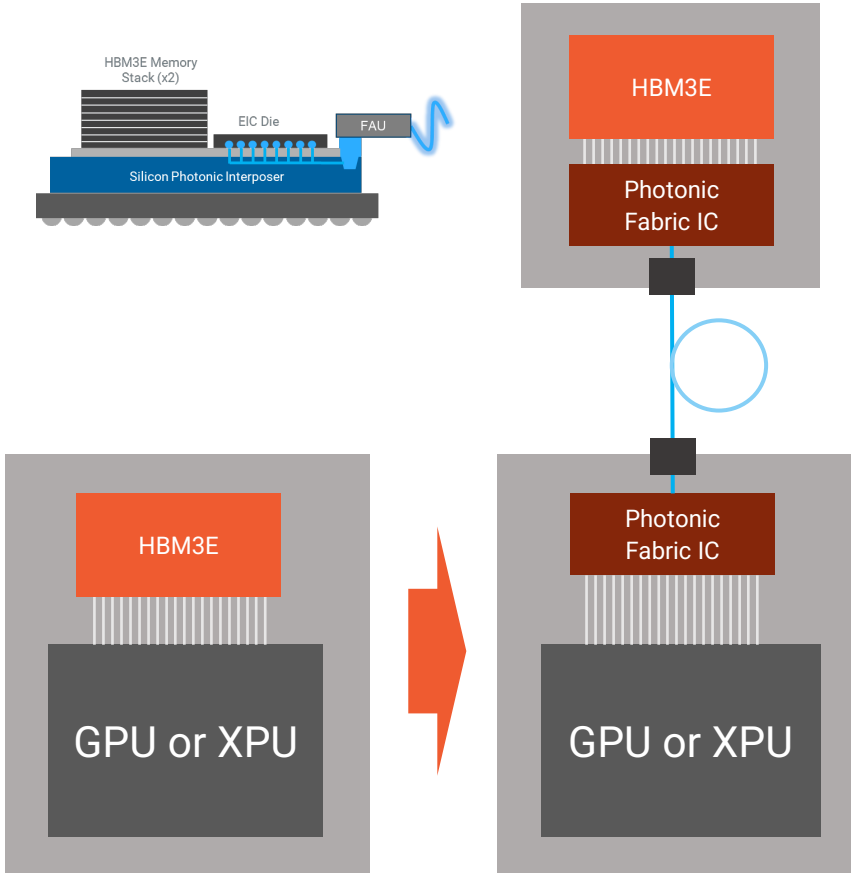
Features

- Memory pooling appliance allows up to 10 CXL AICs to be housed in one unit
- Support up to 80 DDR5 RDIMMs (or 120 DDR4 RDIMMs) in one chassis
- Eventually will support other PCIe peripherals e.g. GPU, FPGA or NIC
- Share memory pool with up to 5 servers via disaggregation management software
- Expandable with dedicated switch units

Benefits

- Host server only needs to use a single PCIe/CXL slot
- Unit may be serviced and memory upgraded/added/removed without powering off servers

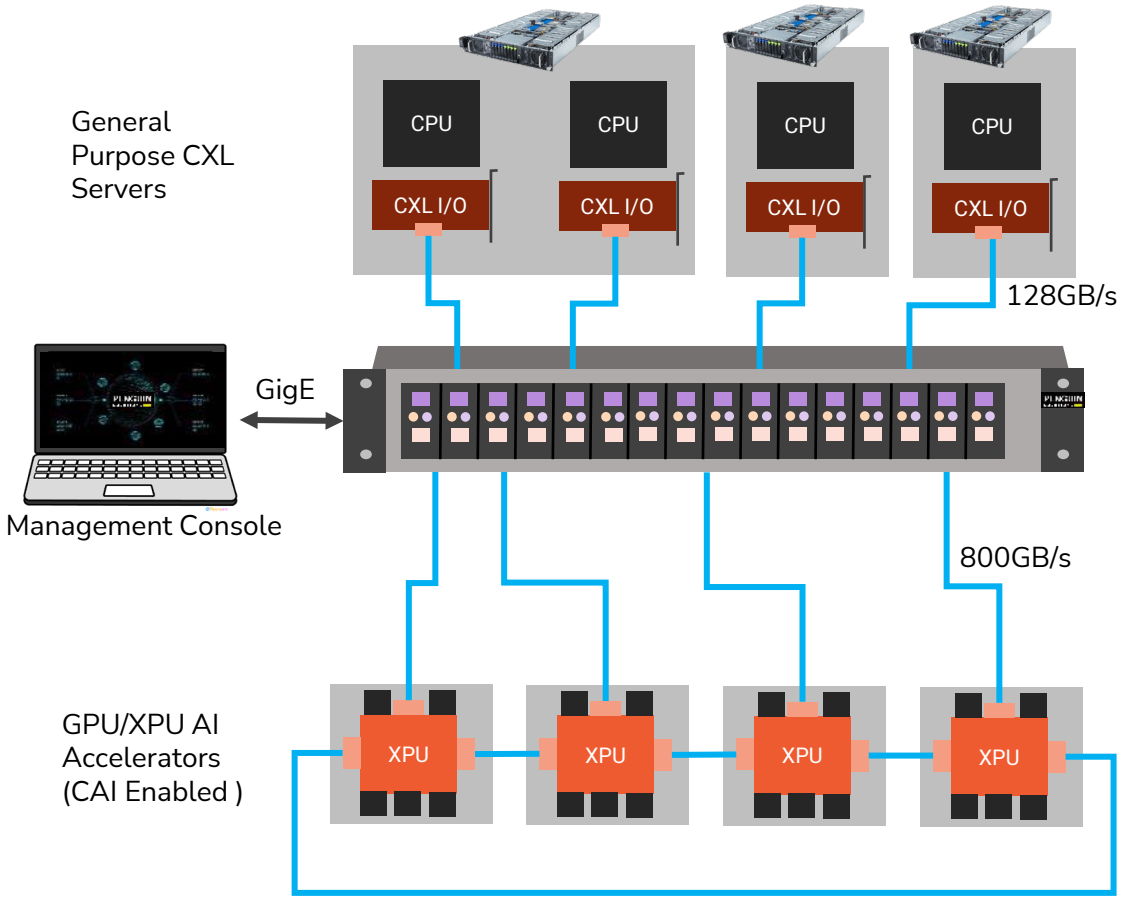
The Rise of Silicon Photonic Memory



- Silicon Photonics Avoids the Beachfront problem and allows serializing of HBM data streams over optical channels and fiber
- HBM3 uses 1024 data signals. HBM4 discussing 2048 data signals which takes up even more silicon “beachfront”
- Moving to serial attached memory with on-chip optics allows HBM memory to be located “off chiplet” and opens the path to GPU-CPU sharing architectures

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
- Memory Supported
  - Up to 32TB DDR5 RDIMMs
  - Up to 1TB HBM3E
- 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

SNIA | CMS A SNIA Community

SNIA | CMS
A SNIA Community

Questions?

Next Steps. . .

- ✓ Please rate this webinar and provide us with feedback
- ✓ Look for a Blog answering questions from this webinar at snia.org/blog
- ✓ Get more SNIA education!
  - **Live**
    - FMS: Future of Memory and Storage, August 4-7, 2025 Santa Clara CA futurememorystorage.com
    - SNIA Developer Conference, September 15-17, 2025 Santa Clara CA sniadeveloper.org
    - SC25, November 16-21, St. Louis MO sc25.supercomputing.org
  - **Online**
    - This webinar and many other videos and presentations on today's topics are in the **SNIA Educational Library** https://snia.org/educational-library
    - **SNIA Video You Tube Channel** https://www.youtube.com/user/sniavideo
- ✓ Join SNIA and the Persistent Memory (PM) State Drive Special Interest Group
  - www.snia.org/join
  - www.snia.org/groups/cms
  - www.snia.org/forums/cmsi/NVDIMM

SNIA | CMS A SNIA Community

SNIA | CMS
A SNIA Community

Thank You