Benefits of CXL for Server Memory Infrastructure

August 6, 2025

Growth of the CXL

CXL Integrators List

Device Type		Feature Set		Spec Revision		Function	
Type 1	7	CXL Core 1.1	41	CXL 1.1	30	Accelerator	3
Type 2	8	CXL Core 2.0	26	CXL 2.0	29	Host	11
Type 3	56					IP	14
						MEM Expander	34

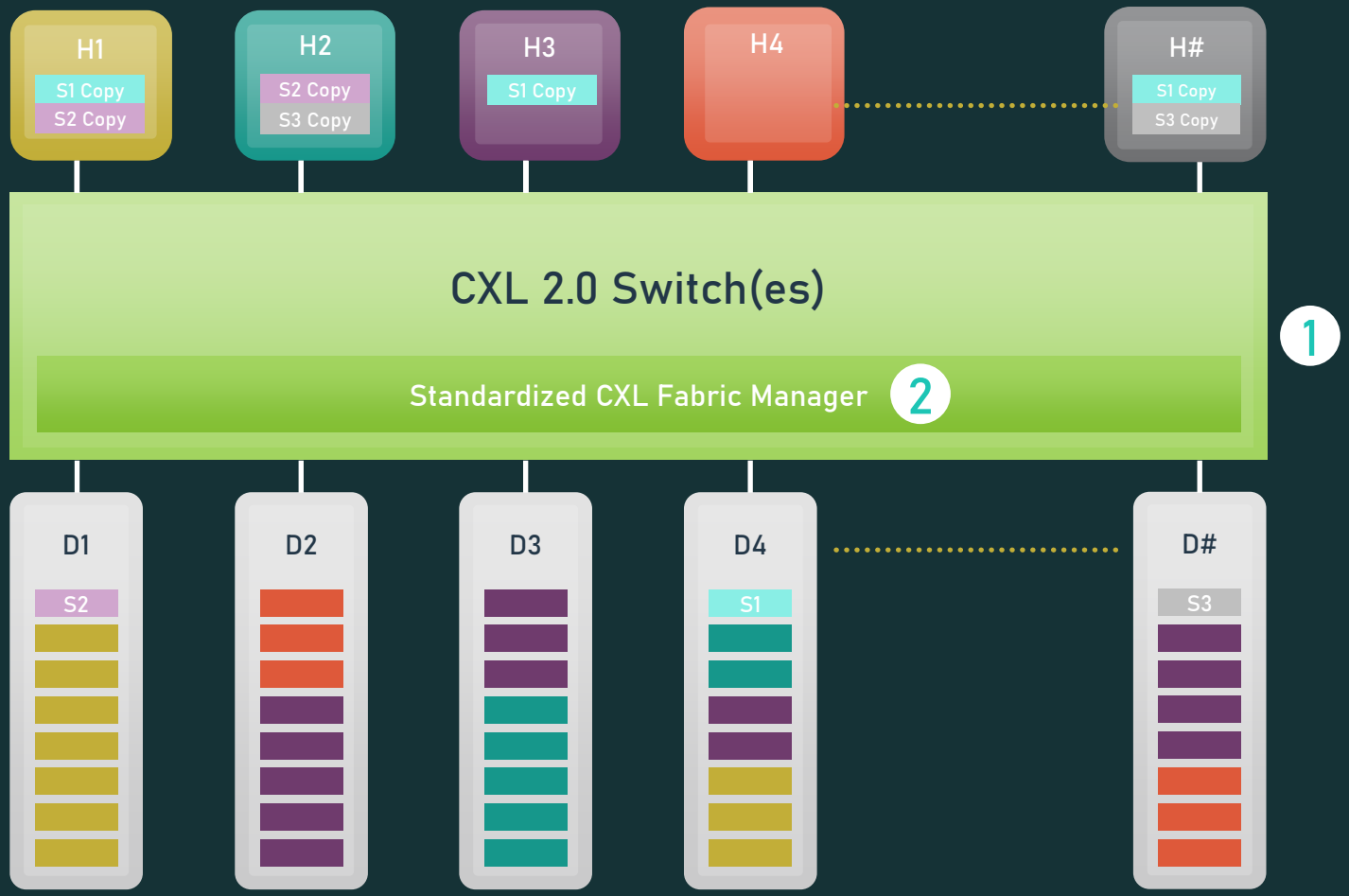
Advanced Micro Devices Inc.	AMD EPYC 9005 Series Processors	Turin	Type 1, Type 2, Type 3	CXL Core 1.1, CXL Core 2.0, MEM 2.0	CXL 2.0	32GT/s	x16	Other - Root Complex	Host	CTE 006
Alphawave Semi	KappaCore32 (PCIe/CXL Controller)	1001	Type 3	CXL Core 1.1, CXL Core 2.0	CXL 2.0	8GT/s	x8	CEM	IP	CTE 007
Astera Labs, Inc.	Astera Labs Aries Gen 5 x8 Retimer	PT5081	Type 3	CXL Core 1.1, CXL Core 2.0, MEM 2.0	CXL 2.0	32GT/s	x8	Other - System on Chip (SoC)	Retimer	CTE 008
Astera Labs, Inc.	Aries Gen 5 Retimer	PT5161	Type 3	CXL Core 2.0, MEM 2.0	CXL 2.0	32GT/s	x16	Other - System on Chip (SoC)	Retimer	CTE 008
Astera Labs, Inc.	Leo Smart Memory Controller	0x01E2	Type 3	MEM 2.0	CXL 2.0	32GT/s	x16	Other - System on Chip (SoC)	MEM Expander	CTE 006
Astera Labs, Inc.	Leo A1000	0x01E2	Type 3	MEM 2.0	CXL 2.0	32GT/s	x16	CEM	MEM Expander	CTE 006
Cadence Design Systems	Cadence CXL Controller IP	0100	Type 3	CXL Core 1.1, CXL Core 2.0	CXL 2.0	8GT/s	x4	CEM	IP	CTE 006
Intel Corporation	Edge-Enhanced Intel Xeon 6 CPU	0DB0, 0DB1, 0DB2, 0DB3, 0DB4,	Type 1, Type 2, Type 3	CXL Core 1.1, CXL Core 2.0, MEM 2.0	CXL 2.0	32GT/s	x16	Other - Root Complex	Host	CTE 006

OEMs offering CXL-capable servers:

- US / EMEA
  - Dell
  - HPE
  - Lenovo
  - Supermicro
- APAC
  - Advantech
  - Giga
  - Quanta
  - AIC

Scan the QR code to view the Integrators List

CXL 3.0: Pooling & Sharing



1. Expanded use case showing memory sharing and pooling

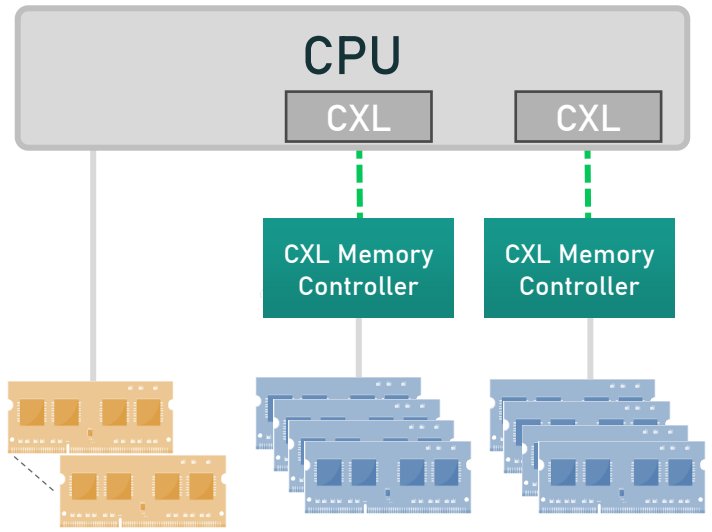
2. CXL Fabric Manager is available to setup, deploy, and modify the environment

Astera Labs

Presented by: Sandeep Dattaprasad

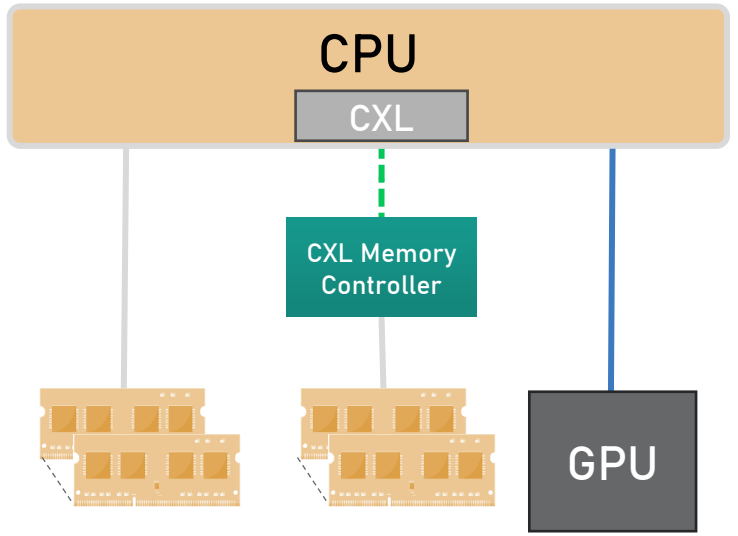
CXL Benefits for AI & Database Infrastructure

In-Memory Databases



- In-Memory DB w/ local + CXL memory
- Increases memory capacity
  - Minimal DB performance impact
- Lower TCO, less power, better thermal mgmt

AI Applications



- Expand host memory for inference/training
- Supports larger batch sizes w/ better performance
  - Higher bandwidth and lower latency
- Improves GPU utilization

Leo Smart Memory Controller Portfolio

Features

Architected to Accelerate AI and Database Infrastructure

- Memory expansion, pooling and sharing
- Performance optimized for memory intensive workloads

Customizable RAS

- Memory testing and repair with vendor defined stress patterns
- Per channel leaky bucket counter for logging DRAM errors
- Enhanced ECC scheme

Datacenter-grade Security

- TEE features compliant w/ leading security standards
- End-to-end security: RoT, Secure boot/update/debug/recovery, anti-rollback

COSMOS framework for Hyperscalers and Enterprise

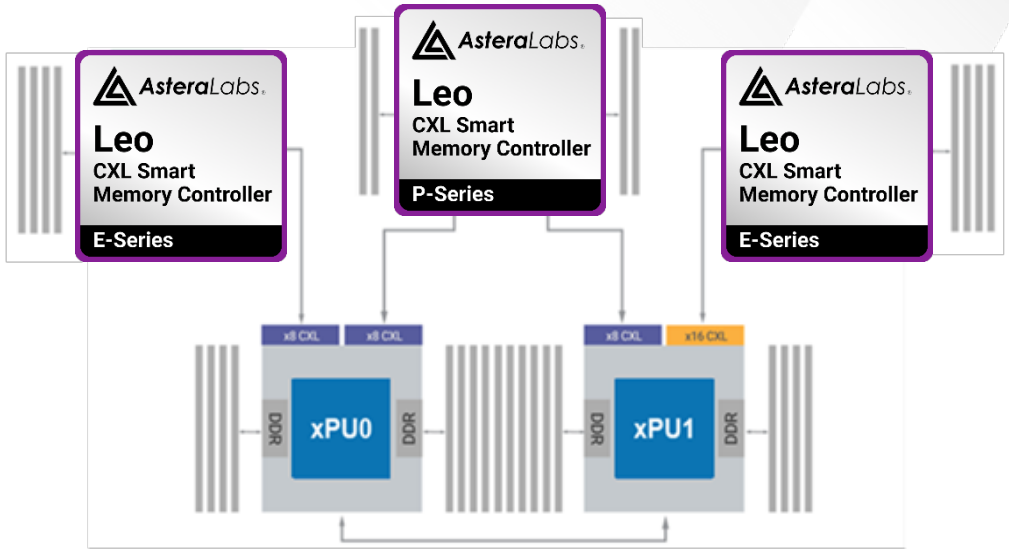
- Tools and infrastructure to assist in Design, Debugging, RAS & Fleet Mgmt
- COSMOS UI, CLI & SDK APIs
- COSMOS Security Manager

Cloud-Scale Interop Lab support

- CXL/PCIe Electrical Testing, Compliance Tests, Reset and Initialization Tests etc.

Parameter	Leo 1
CXL **PCIe**	CXL **1.1/2.0** PCIe 5.0 **32GT/s**
Lane Configuration	1x16, 2x8
DDR Configuration	**2ch DDR5** Up to 5600 (1 DPC) Up to 4800 (2 DPC)
Capacity	Up to **2TB**
Package	27x27mm

System Block Diagram



Product Offering

Leo Smart Memory Controllers

Leo 1 P- Series/E-Series
PCIe 5, CXL 2.0
In Production

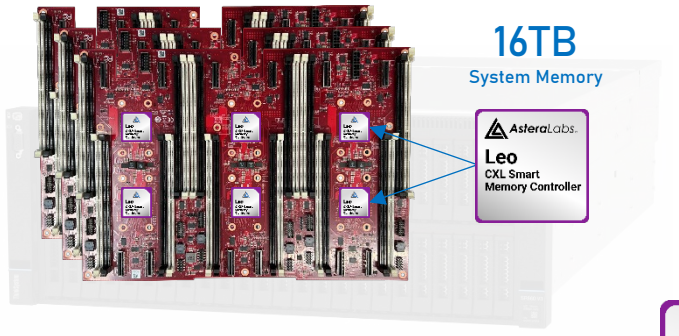
Aurora A-Series Hardware Solutions

4xDDR5 RDIMM
Aurora A1000 Card

Leo Accelerates AI & Database Infrastructure

In-Memory Databases | Transaction Processing

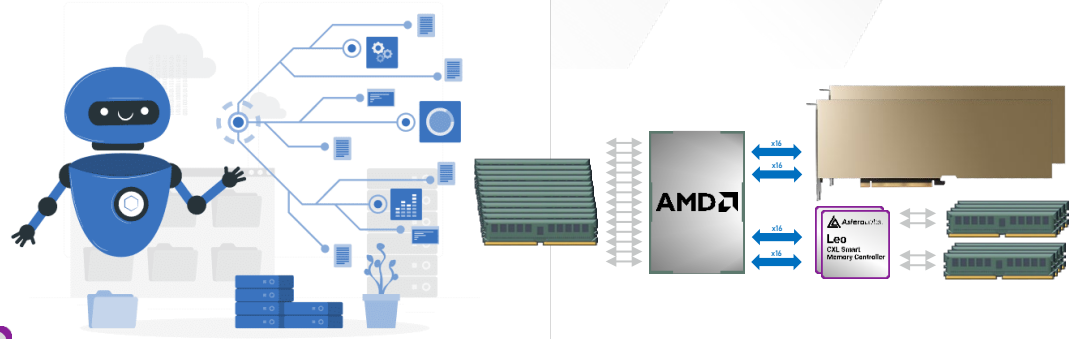
200% more queries per second, **1.5x** memory capacity increase



Based on TPC-H benchmark with 2xLeo CXL Memory Controllers (512GB) and 1xIntel CPUs (1TB)

AI Inferencing | Chatbot Services

40% Faster time to insights with LLMs, w/ **30%** added Memory

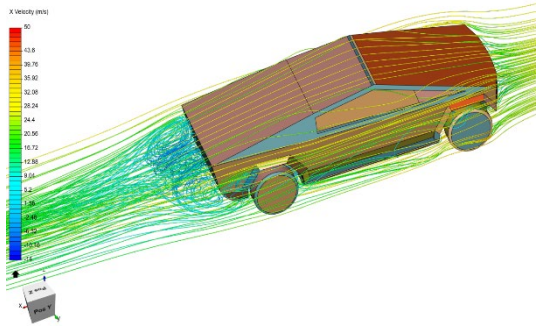


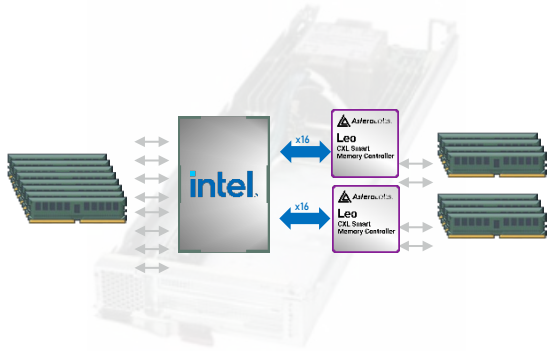
FlexGen (OPT-66B) with 2xLeo CXL Memory Controllers (256GB), 2xNVIDIA GPUs, 1xAMD CPU (768GB)

AsteraLabs.
Leo
CXL Smart Memory Controller

HPC | Computer Aided Engineering

50% more iterations per second, w/ **50%** added Memory

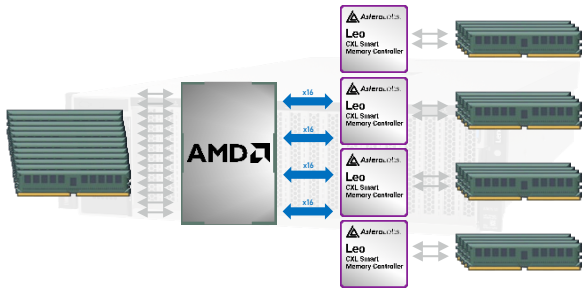




Based on CFD (CPU2017 FP) benchmark with 2x Leo CXL Memory Controllers (256GB) and 1x Intel CPU (512GB)

AI Inferencing | Recommendation System

73% More recommendations per second, **2x** memory capacity increase



Based on DLRM benchmark with 4x Leo CXL Memory Controllers (1TB) and 1x AMD CPU (1.1TB)

Samsung

Presented by: Siamak Tavallaei

What is a RAG Pipeline?

Core components

- Data
- Model
- Embeddings
- Query

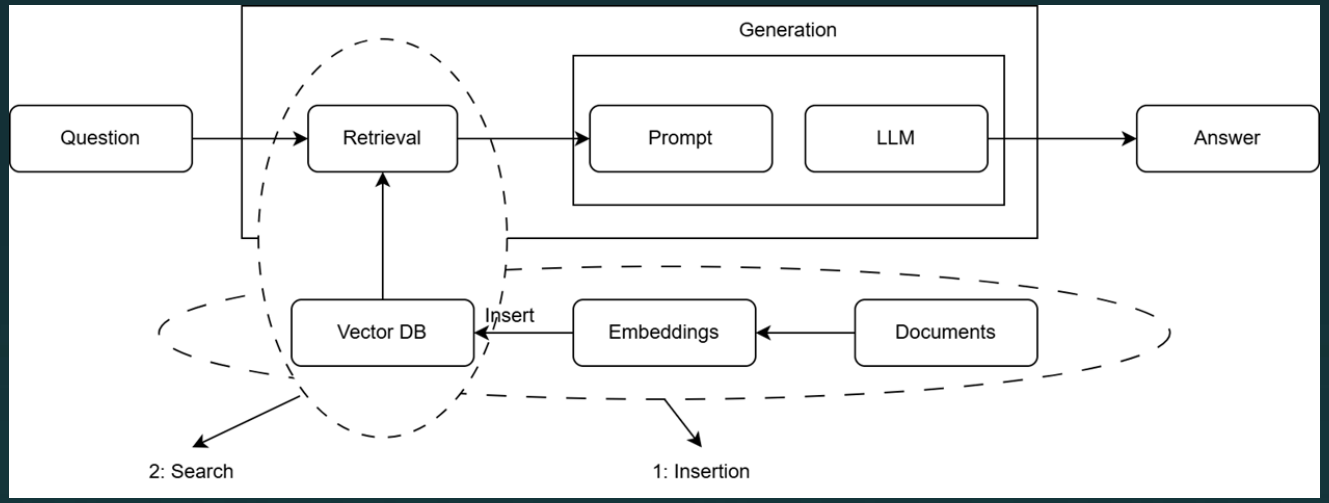
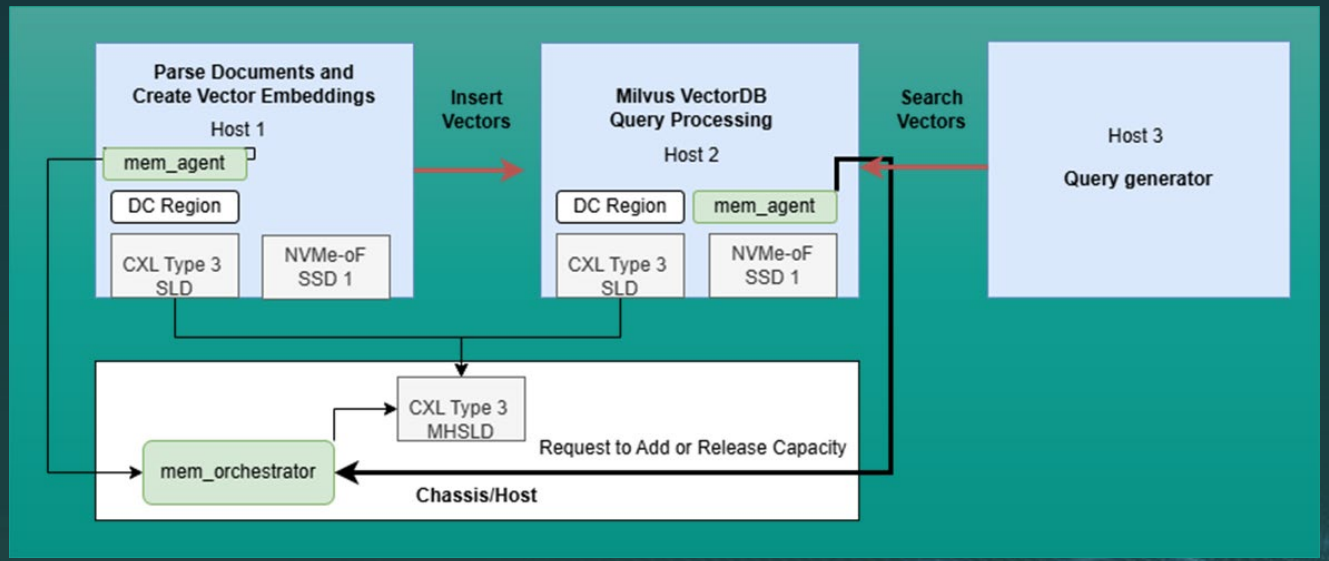


Figure based on[1]

Memory Demand

Phased Approach

- Generate Embeddings
  - Memory demand spikes
- Running the pipeline
  - Based upon the app



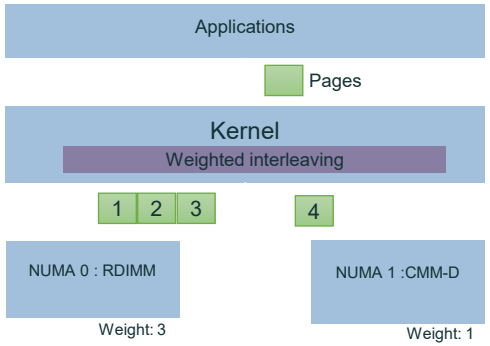
Advantages with CMM-D in RAG Cluster

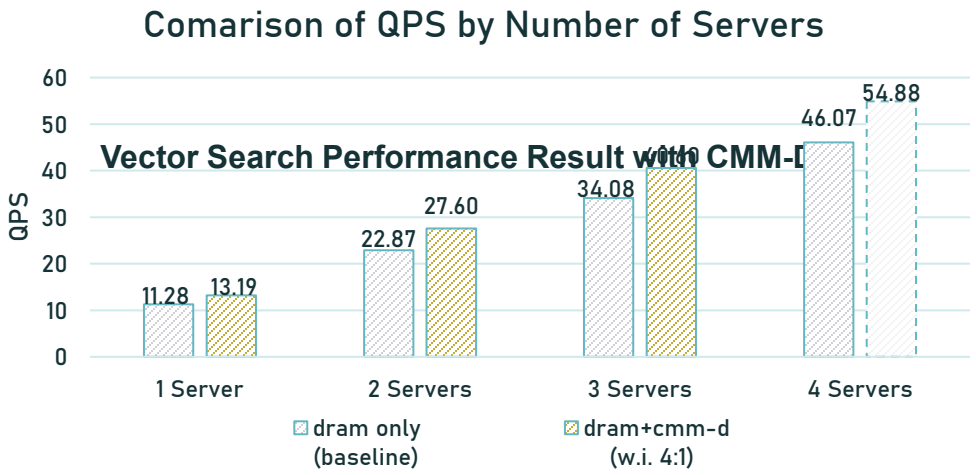
Up to 19% higher performance with CMM-D in VectorDB search compared to DRAM case in Milvus RAG cluster

- Performance gain with bandwidth expansion through the CMM-D in Milvus RAG Cluster
- Using SW interleaving (between DRAM and CMM-D) to achieve optimal CXL bandwidth performance

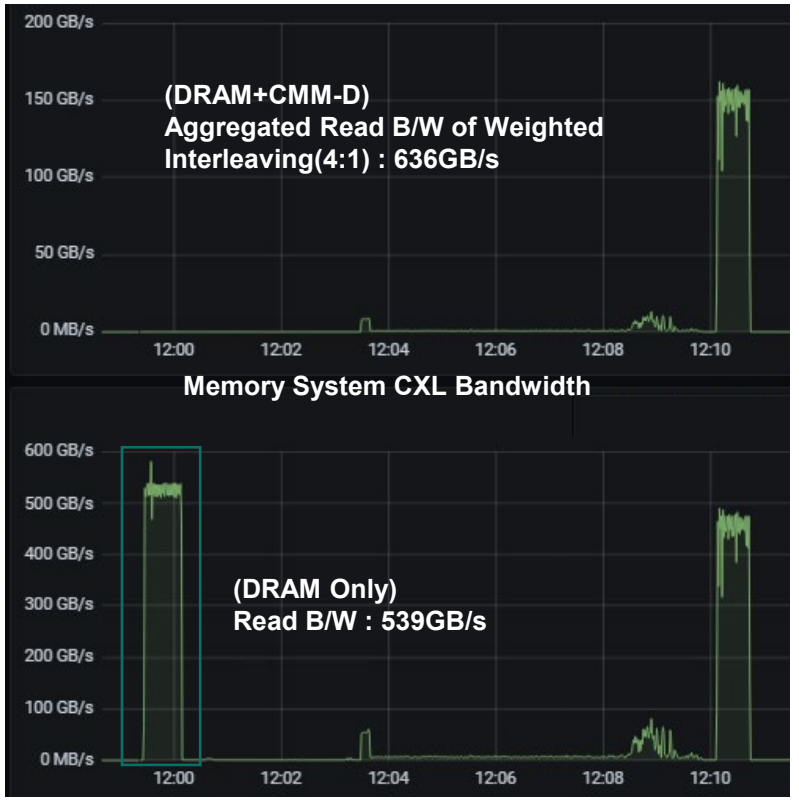
****Weighted Interleaving**

- Linux kernel SW weighted interleaving provides opportunity to define an interleave ratio to best utilize DRAM and CXL memory for optimal performance in a workload
- Included in Kernel Mainline (v6.9)





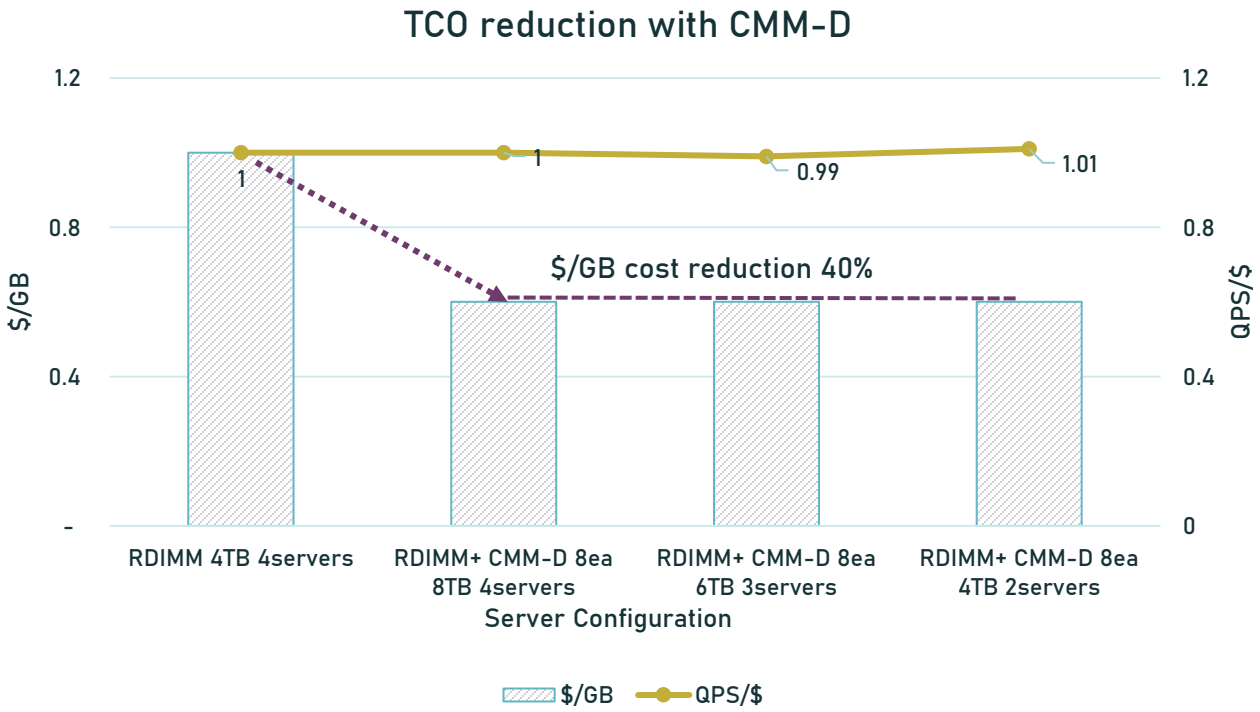
B/W monitoring results in one data server



Advantages with CMM-D in RAG Cluster

TCO reduction effect and memory expansion effect can be secured

- Equivalent QPS/$ and 40% reduction in $/GB cost
- Operating Power reduction through application can reduce operating cost.



Dataset : MSMARCO-V2

Raw Size	Indexing Size(HNSW)	Entity Count	Dimension	Precision	Vector Size
290GB	673GB	138 Million	1024 (cohere)	FP32	4096B

Reference TCO Calculator :https://v0-cxl-tco-2-nvdatd.vercel.app/

Related FMS Activities

- Join the panel discussion on Thursday
  - Title: Driving Interconnects: Memory and storage fabrics for new AI/ML workloads
  - Thursday, August 7, from 1:25 – 2:30 pm PT
  - AI/ML Track (AIML-304-1)
  - Panelists from: Meta, Microsoft, Texas A&M, Cal Poly University, and Samsung
  - Location: Ballroom A

- Stop by Samsung's booth (#407) to learn more about our CXL solutions.

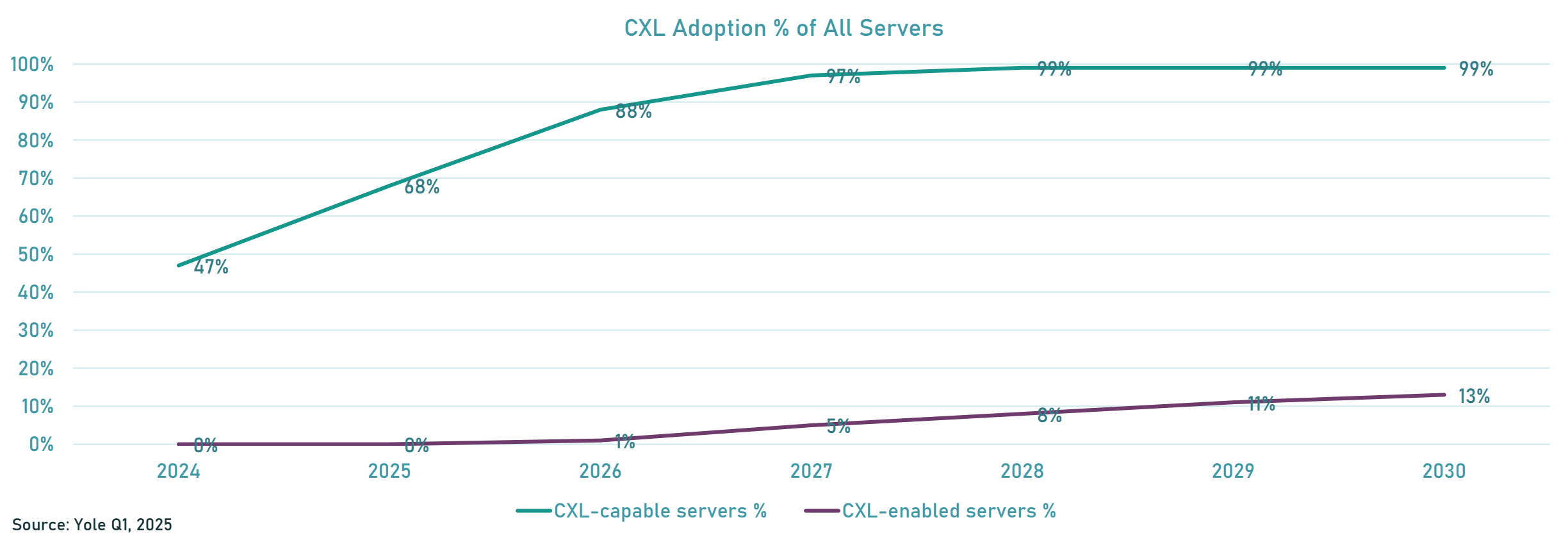
Montage Technology

Presented by: Geof Findley

Montage Technology and CXL Memory Expansion Controller – MXC

- Montage more than 20 years in memory products leading the industry. Second largest PCIe GEN5 and GEN4 Retimer supplier. 1st to ship CXL controllers…Gen 1, 2, and now Gen 3
- MXC newest product based on deep understanding of both DDR and PCIe technologies
- MXC GEN1 supports CXL2.0 and DDR4-3200/DDR5-4800 (In mass production)
- MXC GEN2 supports CXL2.x and DDR5-6400 (In mass production)
- MXC GEN3: Shipping M88MX6852 Type3 CXL® Memory eXpander Controller (Industry 1st)
  - CXL 3.1 compatible
  - PCIe Gen6 speed up to 64GT/s
  - CXL x8 port with bifurcation to 2x4 ports
  - Up to DDR5-8000, with two independent memory controllers
  - Enhanced RAS capability
  - Security with IDE/TSP/DICE
  - Rich management features

CXL Adoption and Mix…has a home!

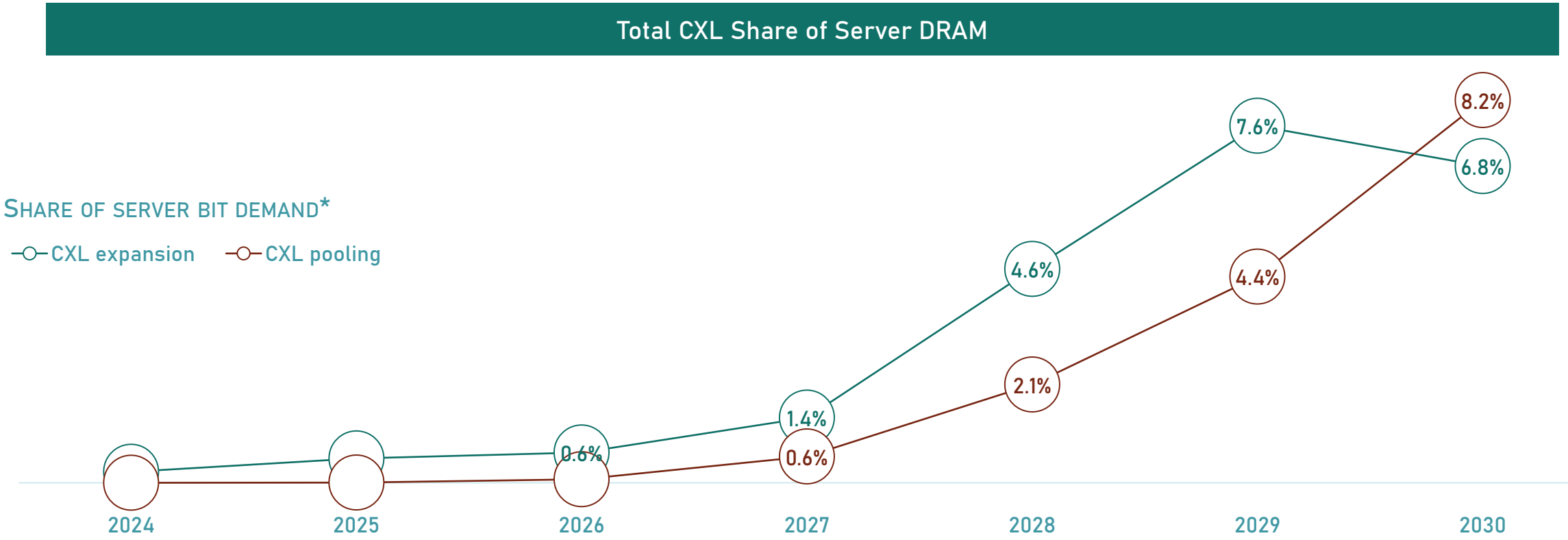


Source: Yole Q1, 2025

Two Thirds of Servers today can support CXL products by end of '26 well over 90%

CXL adoption in the datacenter…its Started

- 32Gb monolithic die (and corresponding 128GB RDIMM) and MRDIMM (with higher bandwidth) are alternatives to CXL expansion. Expansion being considered when high DRAM content is desired
- Memory pooling is getting developed and deployed with CXL 2.0 today, will explode when CXL3.x is available.



Source: Techinsight Q2, 2025

TCO Savings Examples with CXL Memory

Avoid High-Cost DIMMs

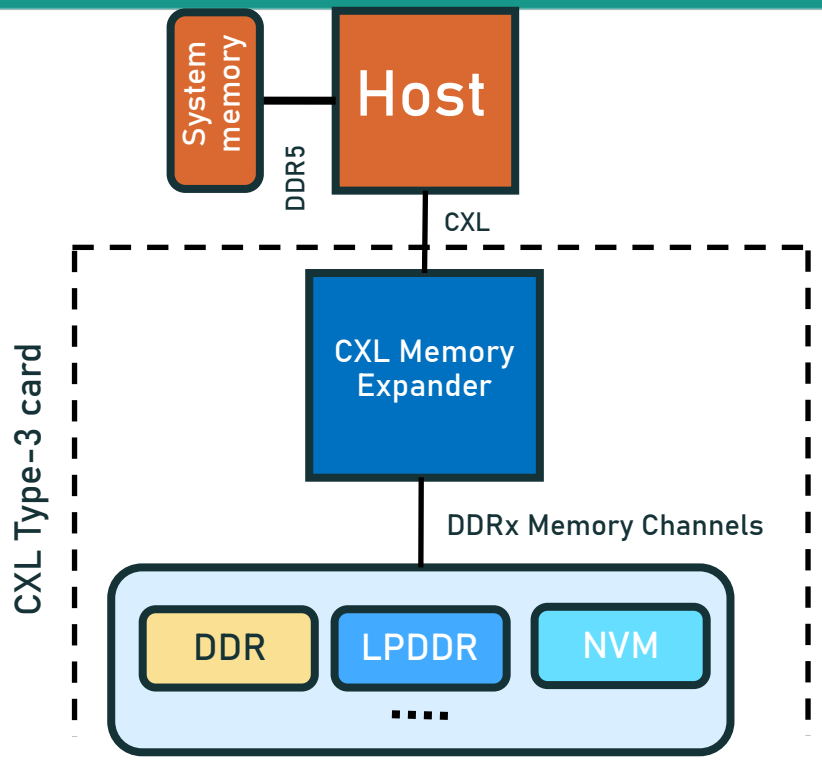
- 128GB and 256GB DIMMs have high price premiums
- CXL add-in cards with DIMM slots provide more total channels per socket → same system capacity with lower priced DIMMs
- Same concept applies regardless of mode: Intel Flat Memory Mode & SW-based tiering

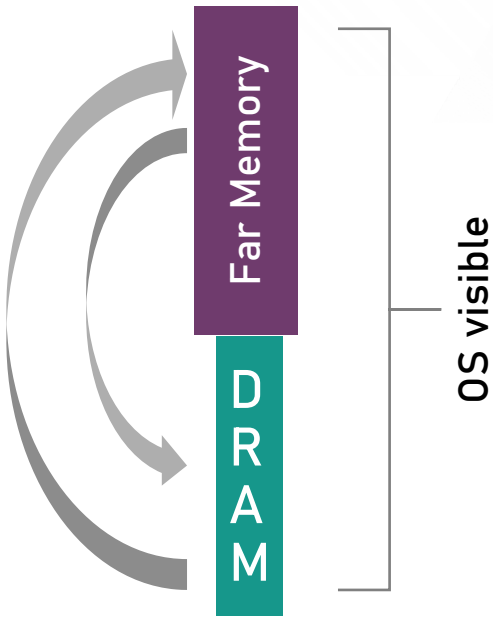
Memory per socket	Socket-attached DIMMs only	With CXL	Memory TCO Savings*
1TB	8x 128GB (1DPC)	8x 64GB + 8x 64GB on CXL	24%
2TB	16x 128GB	16x 64GB + 16x 64GB on CXL	27%
4TB	16x 256GB	16x 128GB + 16x 128GB on CXL	16%

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

Use CXL-attached DIMMs to achieve high system memory capacity and avoid expensive high-capacity DIMMs

Memory Tiering: H/W Based Example, Intel FMM





Flat MM (1:1 ratio)

e.g. 512GB DRAM + 512GB Far Memory

- Both DRAM and far memory exposed to OS as combined physical memory – One memory tier
- Data is 'Tiered': Resides in either DRAM or FM - no replication
- Hot data is swapped into DRAM – one cacheline at a time, not a whole 4KB page
- Performance very good due to 1:1 Near/Far memory ratios
- No software modification needed

Summary and Conclusion

- Throughput Analysis
  - CXL FMM Setup shows performance almost equivalent to Native Setup in terms of throughput
  - Across all tested workloads, the CXL FMM setup consistently delivers performance within ~95–100%

- Latency Analysis
  - Read Latency: CXL FMM Setup tends to be 5–10us higher than Native, a relative increase of 3–5%
  - Update Latency: Generally slightly higher on CXL FMM (10–20us)
  - Latency results are consistent across repetitions
  - Latency with CXL FMM Setup is slightly higher, especially for update operations, but the increase is small

- Stability
  - Each workload repeated 3 times, and results were highly consistent, indicating stable system behavior.
  - No performance degradation or instability was observed due to CXL usage.

Conclusion: CXL FMM Setup demonstrates excellent usability and stability in MongoDB performance testing

Note: CXL Memory Module only add additional hardware latency less than 100ns. However, Overall CXL FMM latency addition is 5-10us. This hints much of the latency savings could come from software side improvement

XConn Technologies

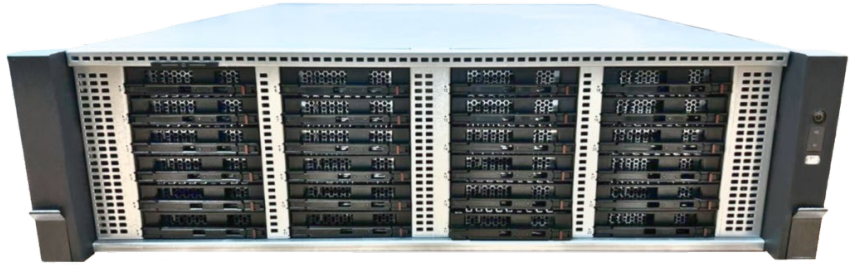
Presented by: JP Jiang

XConn's CXL 2.0 Switch

- World's First CXL 2.0 (XC50256) & PCIe 5.0 (XC51256) switch IC
- 2,048 GB/s total BW with 256 lanes
- Lowest port-to-port latency
- Lowest power consumption/port
- Reduced PCB area, Lower TCO
- Compatible with CXL 1.1 and CXL 2.0
- Supports memory expansion/pooling/sharing
- Works in hybrid mode (CXL/PCIe mixed)
- In production and shipping now

CXL Memory Sharing/Pooling Chassis

H3P



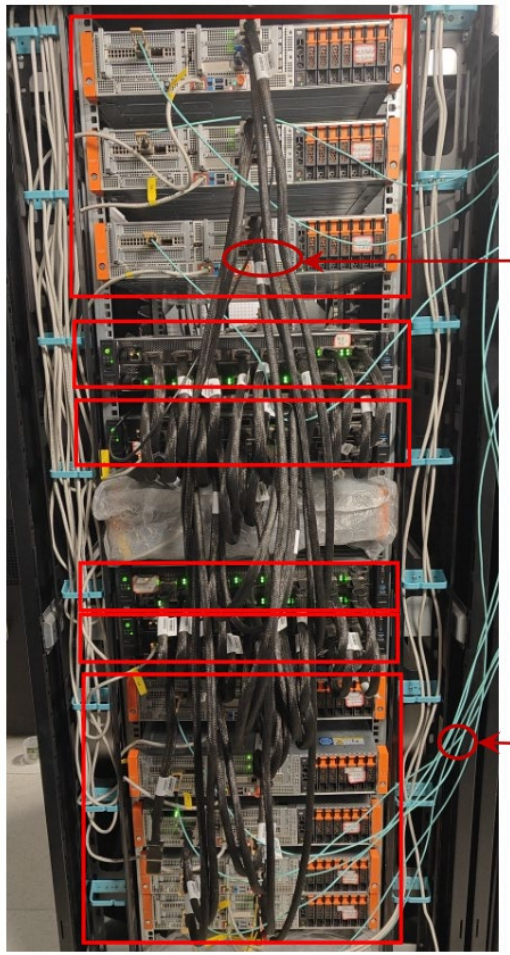
Samsung

Liqid

XConn Technology

Composable Memory Pooling & Sharing

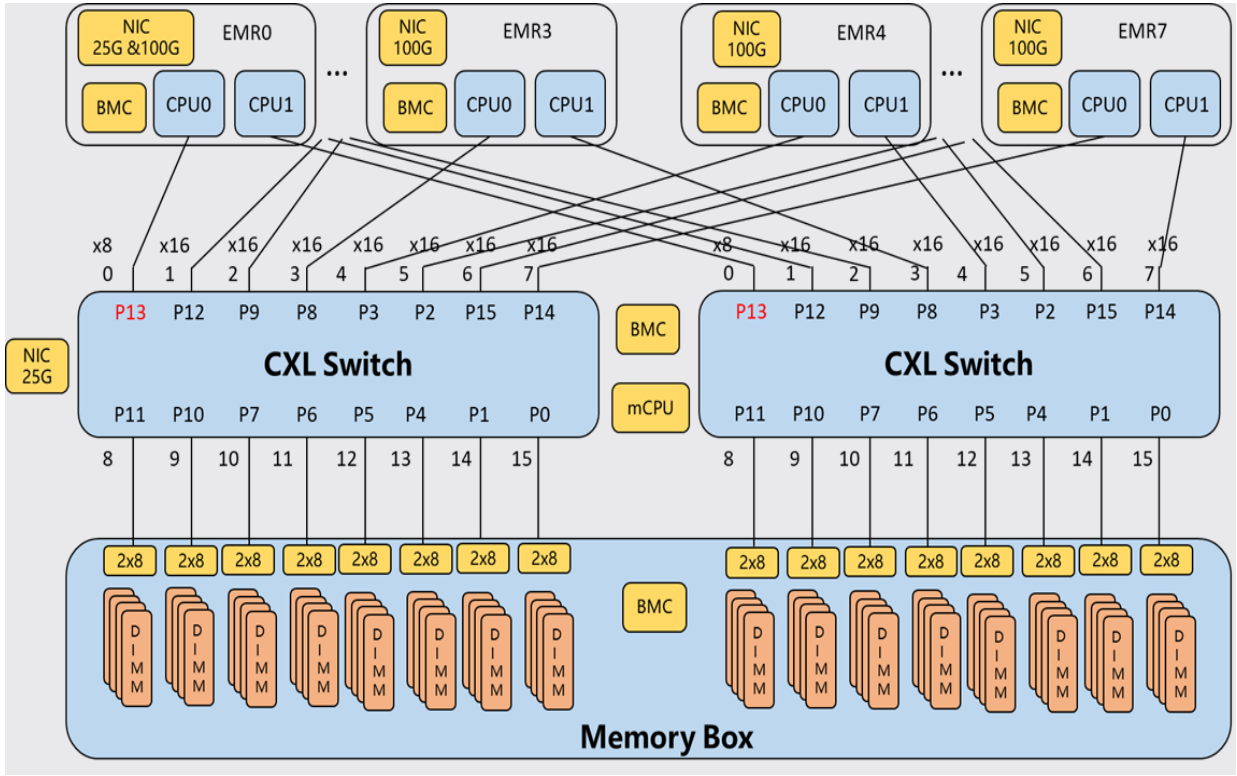
For Cloud Database



ACM 2025 SIGMOD Industrial best paper award

Unlocking the Potential of CXL for Disaggregated Memory in Cloud-Native Databases

https://lnkd.in/gwwB_4Ph



CXL Performance Improvement over RDMA

Compare with DRAM

	DRAM		CXL w/o switch		CXL w. switch	
	Local	Remote	Local	Remote	Local	Remote
Latency (ns)	146	231	265.2	345.9	549	651

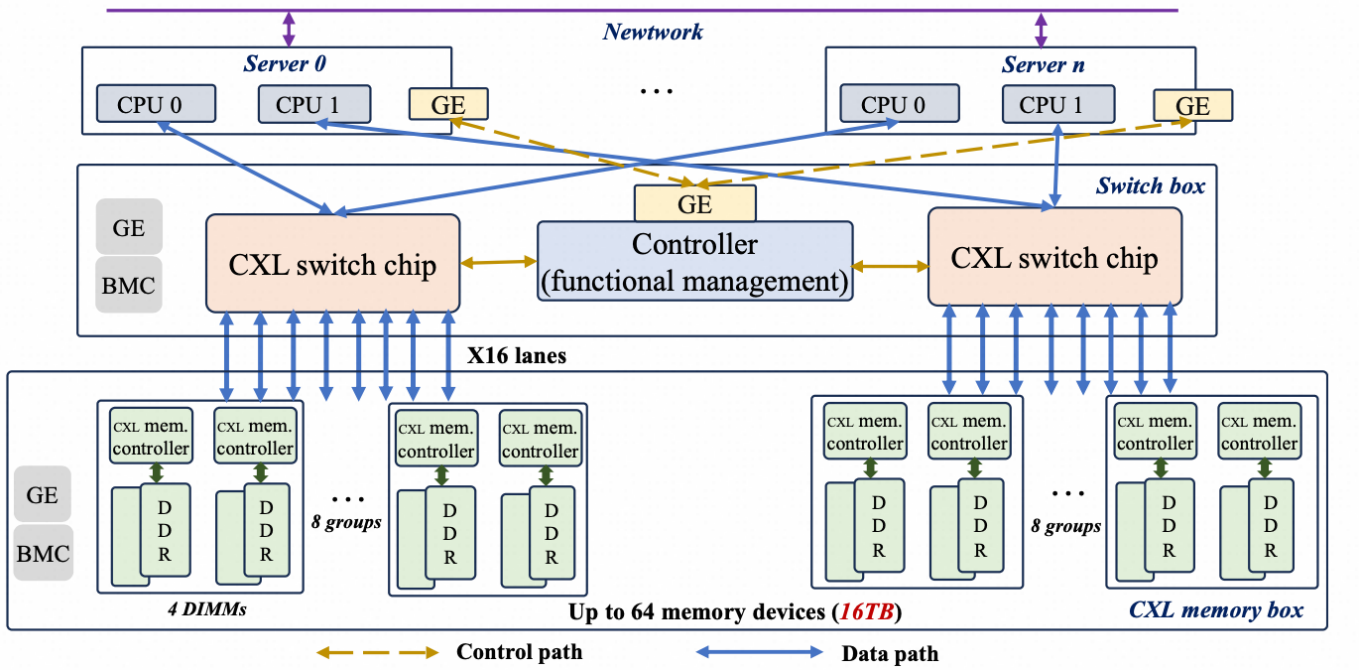
- 3.76× that of local DRAM
- 2.82× that of remote DRAM
- Switch introduces additional latency

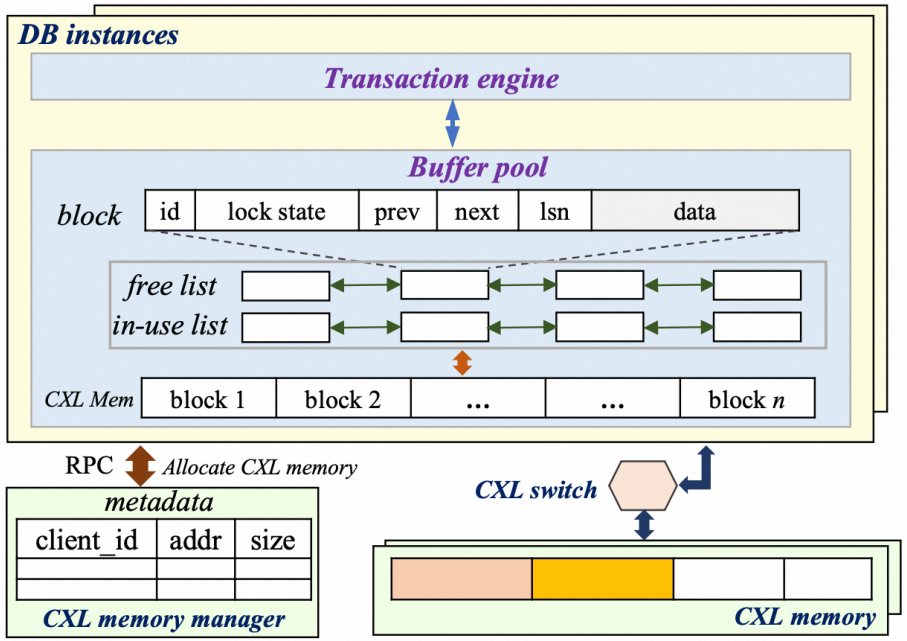
Compare with RDMA

Size	Write latency (μs)		Read latency (μs)	
	RDMA	CXL	RDMA	CXL
64B	4.48	0.78	4.55	0.75
512B	4.69	0.84	4.79	0.85
1KB	4.77	0.88	4.91	1.07
4KB	5.06	1.02	5.58	1.86
16KB	6.12	1.68	7.13	2.46

- Reducing latency by 5.74× for writes and 6.07× for reads at 64B
- CXL latency is more sensitive to data size
- Avoiding page-level copy is beneficial

CXL-based Memory Pool in PolarDB

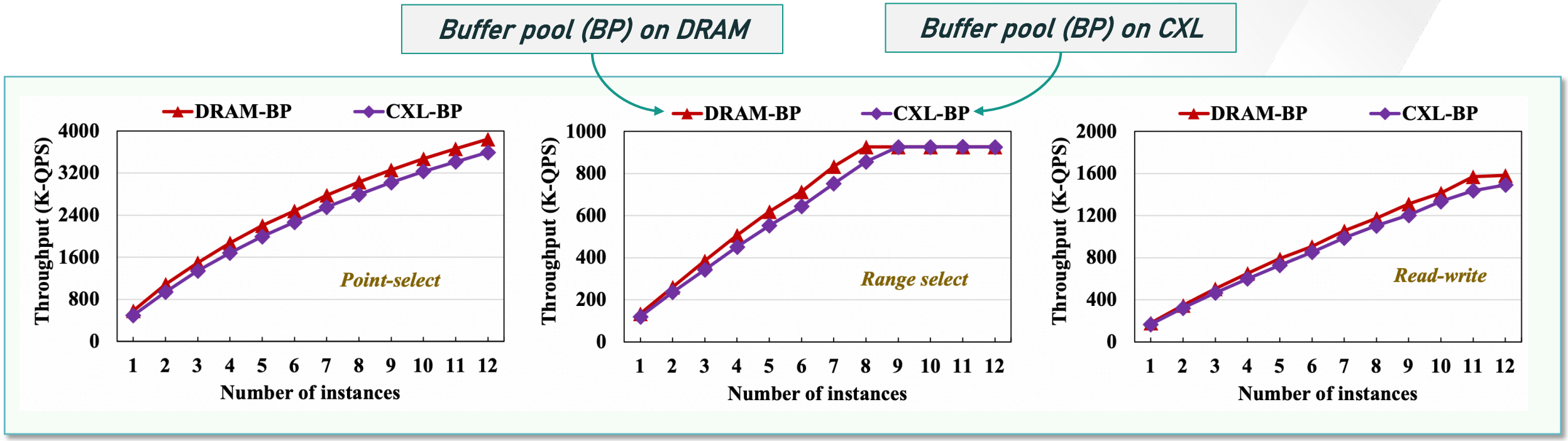




- Servers send control message via Ethernet
- CXL switch is connected via CXL x16 lanes
- Up to 16 TB memory

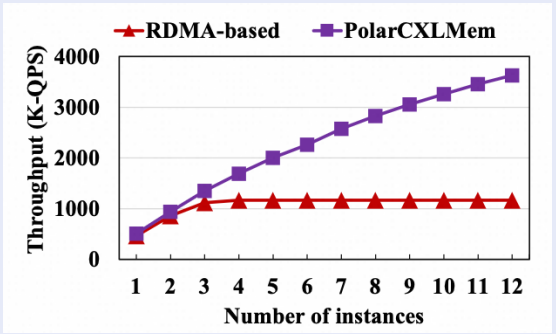
- Avoid tiered memory, deploying BP directly on CXL memory
- A metadata server is dedicated for the CXL memory pool management
- Compute node allocates CXL memory via RPC

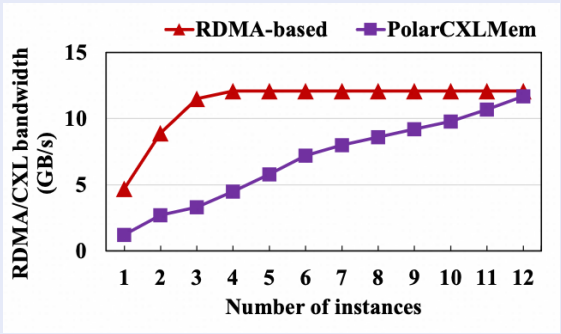
Database Performance on CXL



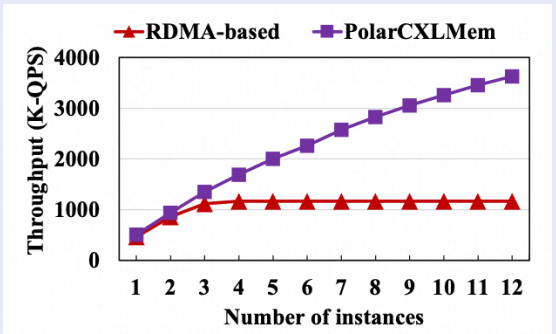
- CXL-BP shows comparable performance with DRAM-BP
- Database buffer pool is bandwidth-sensitive
- Memory tiering is not necessary, saving bandwidth and simplifying design

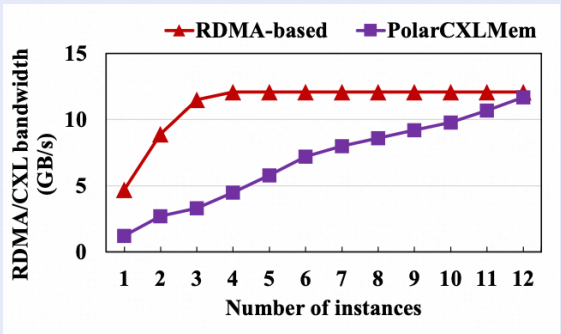
Performance in Pooling Scenarios



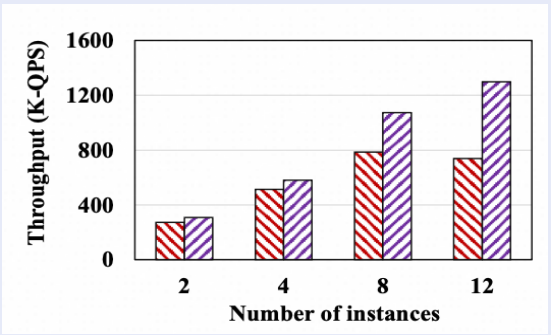


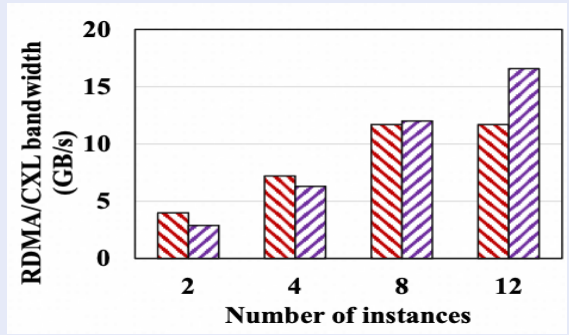
Point-select





Point-select





Read-write

Lower bandwidth usage

⬇ 75%

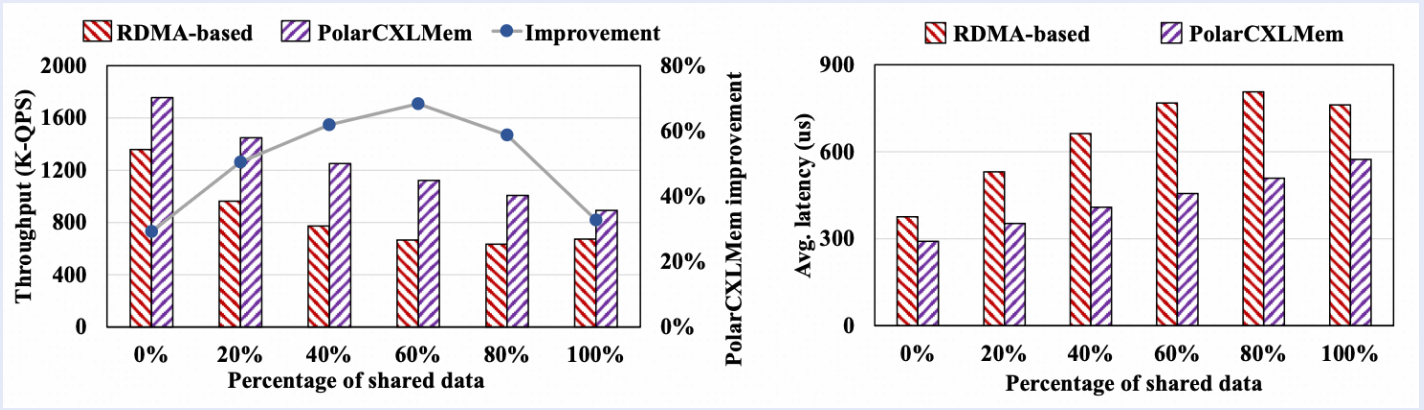
Higher performance

⬆ 3.2x

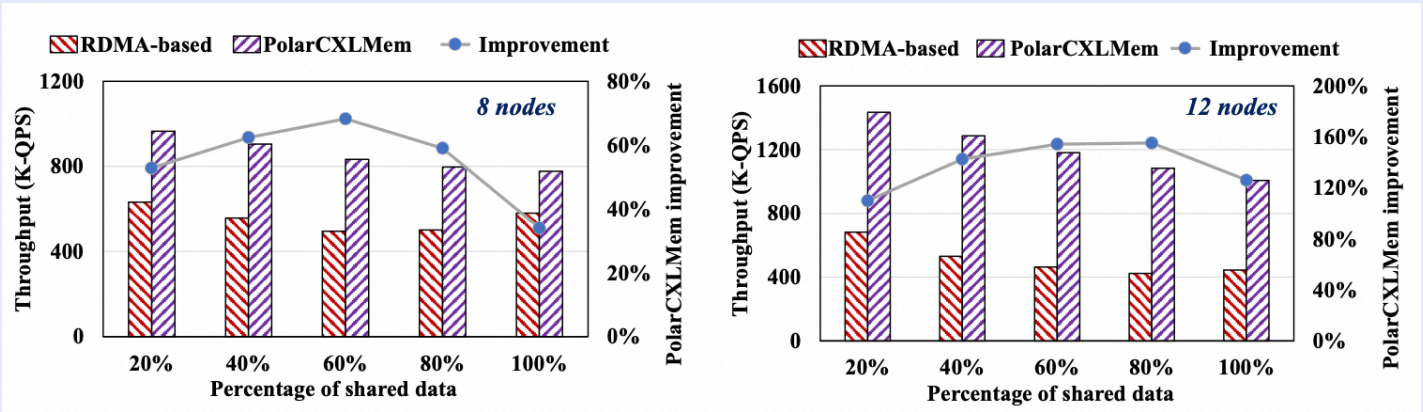
Higher resouces utilization

⬆ 4x

Performance in Sharing Scenarios



Point-update



Read-write

Over 70% improvement
in point-update workload

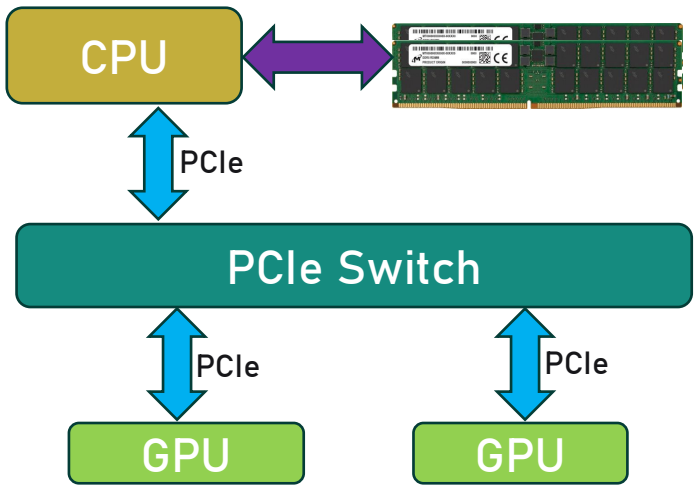
Over 160% improvement
in 12-nodes cluster

Larger cluster,
greater improvement

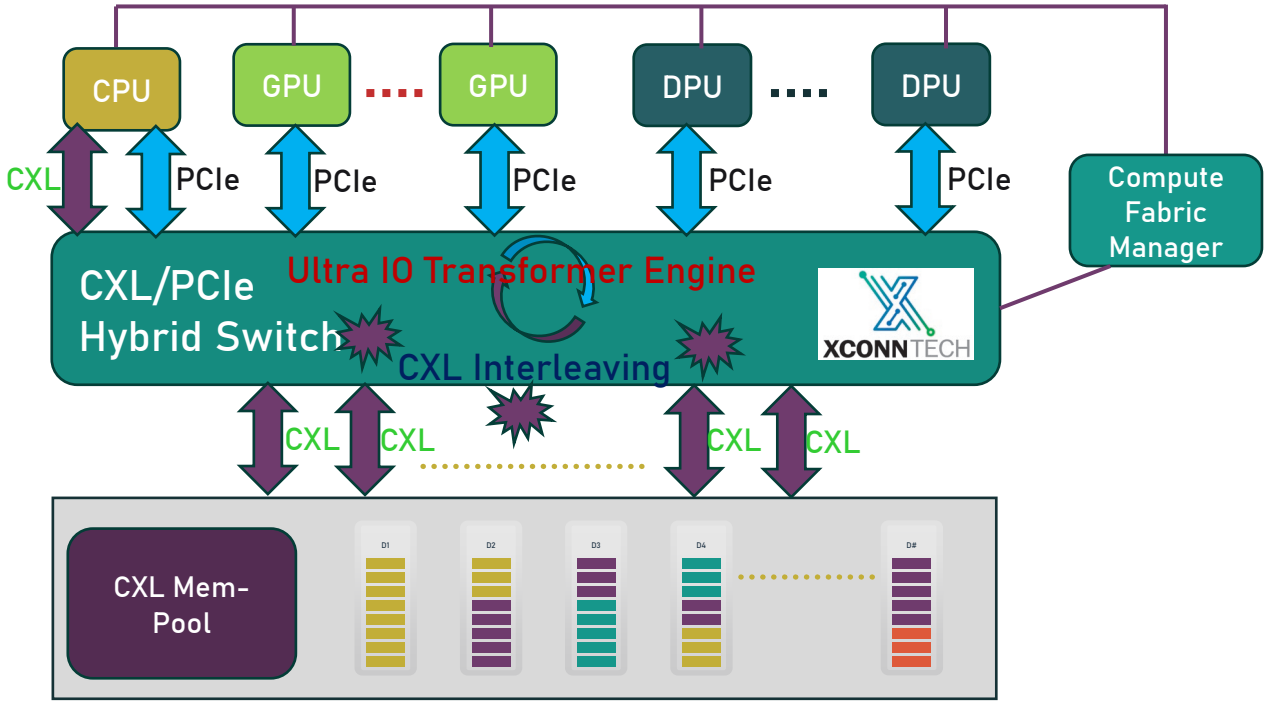
CXL For AI Workloads

Break the Memory Wall

- Majority of GPUs/Accelerators do not support CXL, PCIe is available
- AI “Memory Wall” --Large AI models require multi-TBs or more memory (Tokens, KV caching, etc.)
- XConn’s “Ultra IO Transformer” enables GPUs/DPUs (PCIe devices) to directly access CXL memory pool



CPU-Centric Computing



GPU-Centric Computing With Large Memory Pool Enabled by “Ultra IO Transformer” Technology

Thank You

www.ComputeExpressLink.org