Emerging CXL Uses and Value Proposition

CERN OpenLab Presentation
Mar 5, 2025

Performance
Generative AI and advanced AI applications

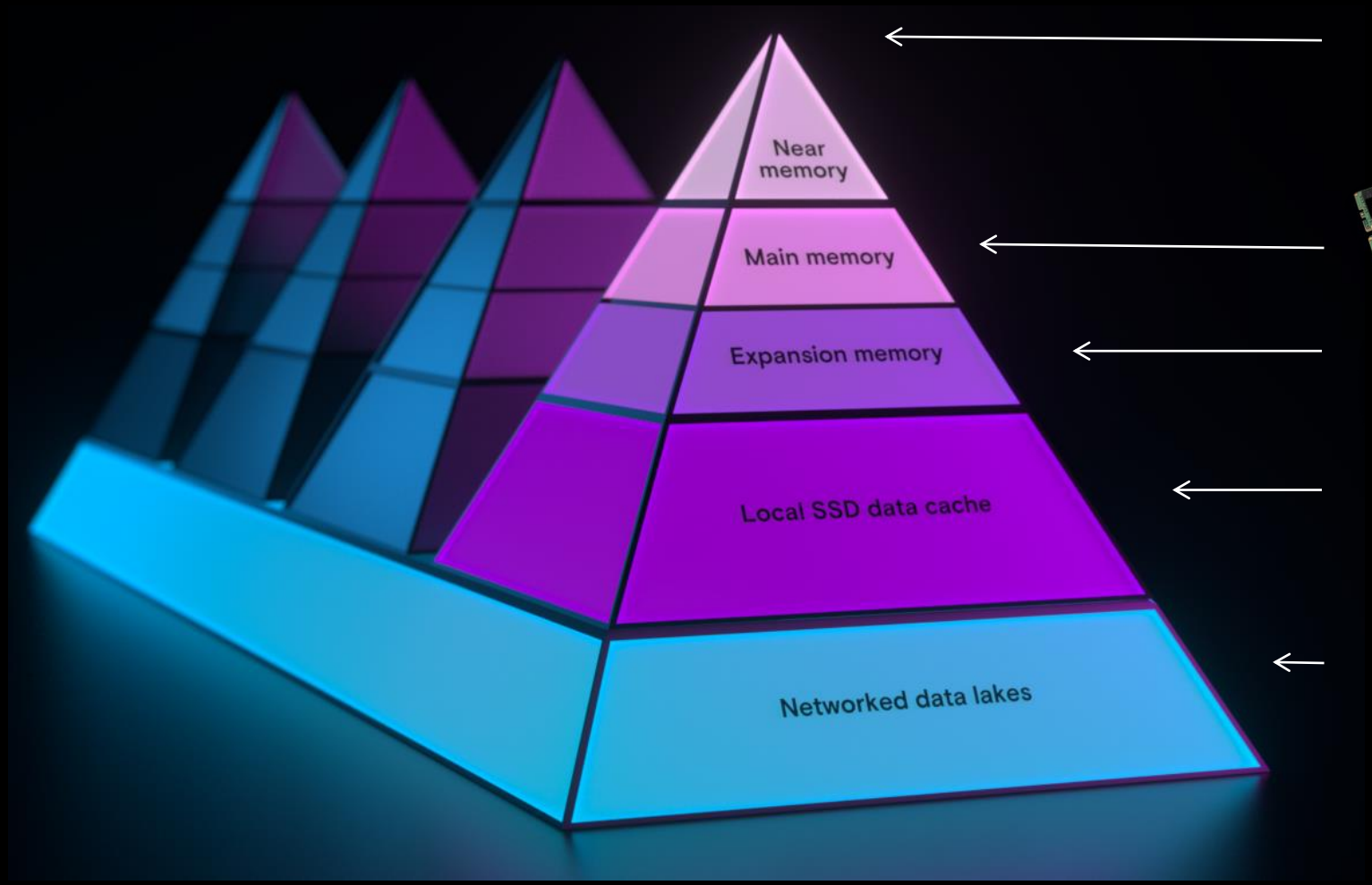
Capacity
Datasets, LLMs, and AI workloads growing capacity exponentially

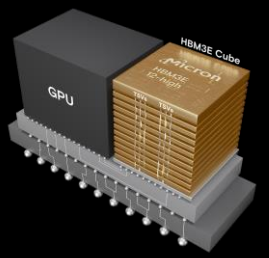
Power
Data centers power consumption almost tripling by 2030

micron

Micron memory and storage hierarchy

Micron has industry-leading products across the datacenter value chain





High-bandwidth in-package memory

HBM3E

High-capacity memory

96GB and 128GB DDR5

Memory expansion module with CXL™ CZ122

High-performance data center NVMe™ SSD

Micron 9000 series

High-capacity data center NVMe SSD

Micron 6000 series

Introducing Micron CZ122 Memory

Delivering capacity, bandwidth, flexibility, RAS [4]

128GB / 256GB

Up to 2TB incremental server capacity supporting CXL 2.0 [1]

Out of Box performance on capacity and BW expansion [3]

37GB/s

Industry-leading low latency and high memory bandwidth per CMM using PCIe® Gen5 x8 [2]

E3.S 2T x8

Industry-standard form factor for broad deployment

1. By adding 8x256GB CZ12x modules, system limitations apply
2. MLC (2R1W) was used to measure BW and latency with Micron and competition
3. Supports heterogeneous Interleave and Meta Data on Major CPU Platforms
4. Supports Internal CVME threshold, Device Initiated hPPR

How CXL Scales Up Server(s) Performance

At Server Level - Attributes Gained via CXL

Additional Direct-Attach memory (up to 2TB)

- Eight E3.S 2T modules * 256GB/module
  - Consuming 4 PCIe lanes
- Enumerated as Interleaved, Warm-tier, or Mem-mapped FS

Additional Memory Bandwidth

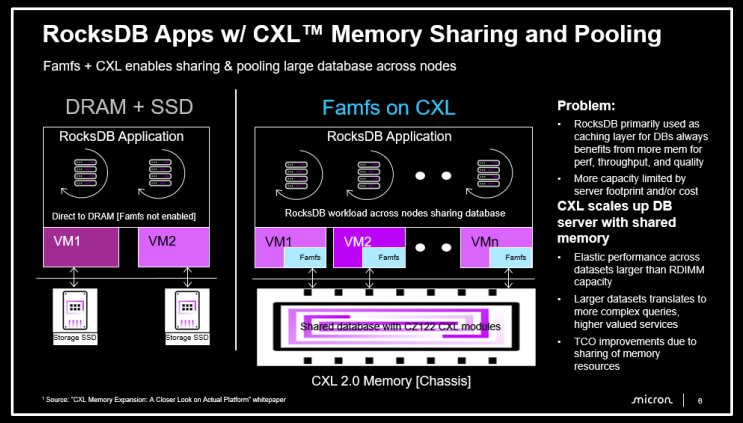
- 37GB/s Dedicated Bandwidth per module
- Up to 296GB/s using 8 E3.S 2T modules
  - Consuming 4 PCIe lanes

Large & Shareable Memory via JBOM (Q1'25)

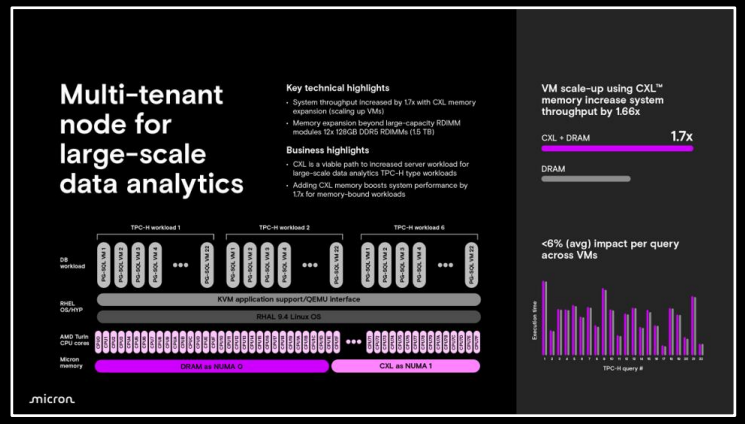
- Up to 5.5TB memory behind JBOM, x8 PCIe lane
- Up to 4 Servers directly connected JBOM for sharing
- Up to 22TB memory for server with 4 JBOMs

CXL™ Playbooks for Partner Engagements

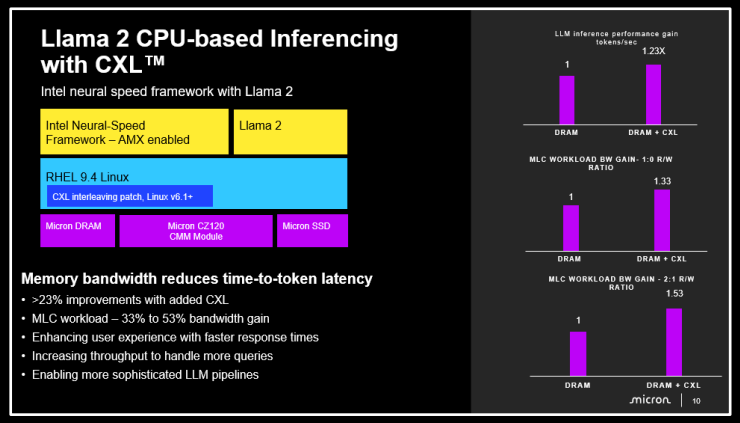
Scaling up DB sizes with large shared expanded CXL memory (up to 20TB)



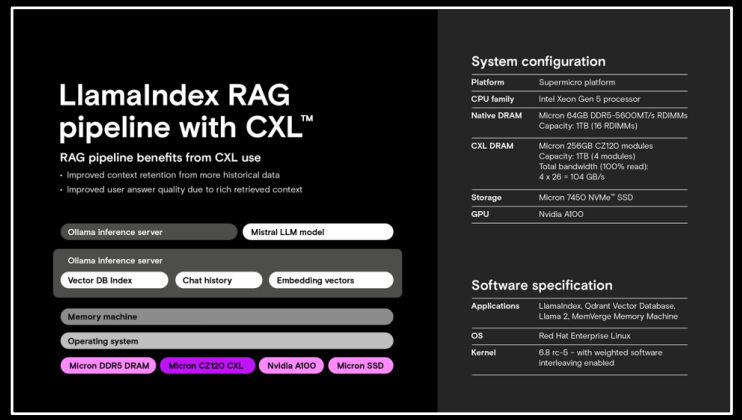
Improving CPU core utilization for multi-tenant workload (by 70%)



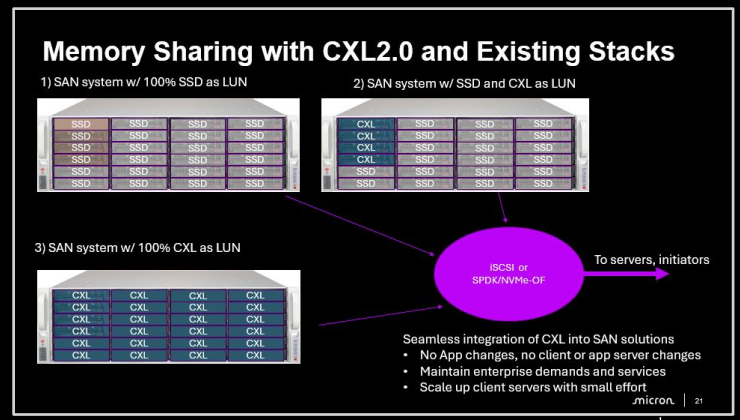
Acceleration Llama 2 CPU only inferencing via additional CXL Bandwidth (by 23%)



Accelerating RAG pipeline with large shared memory and added BW (by 8X!)

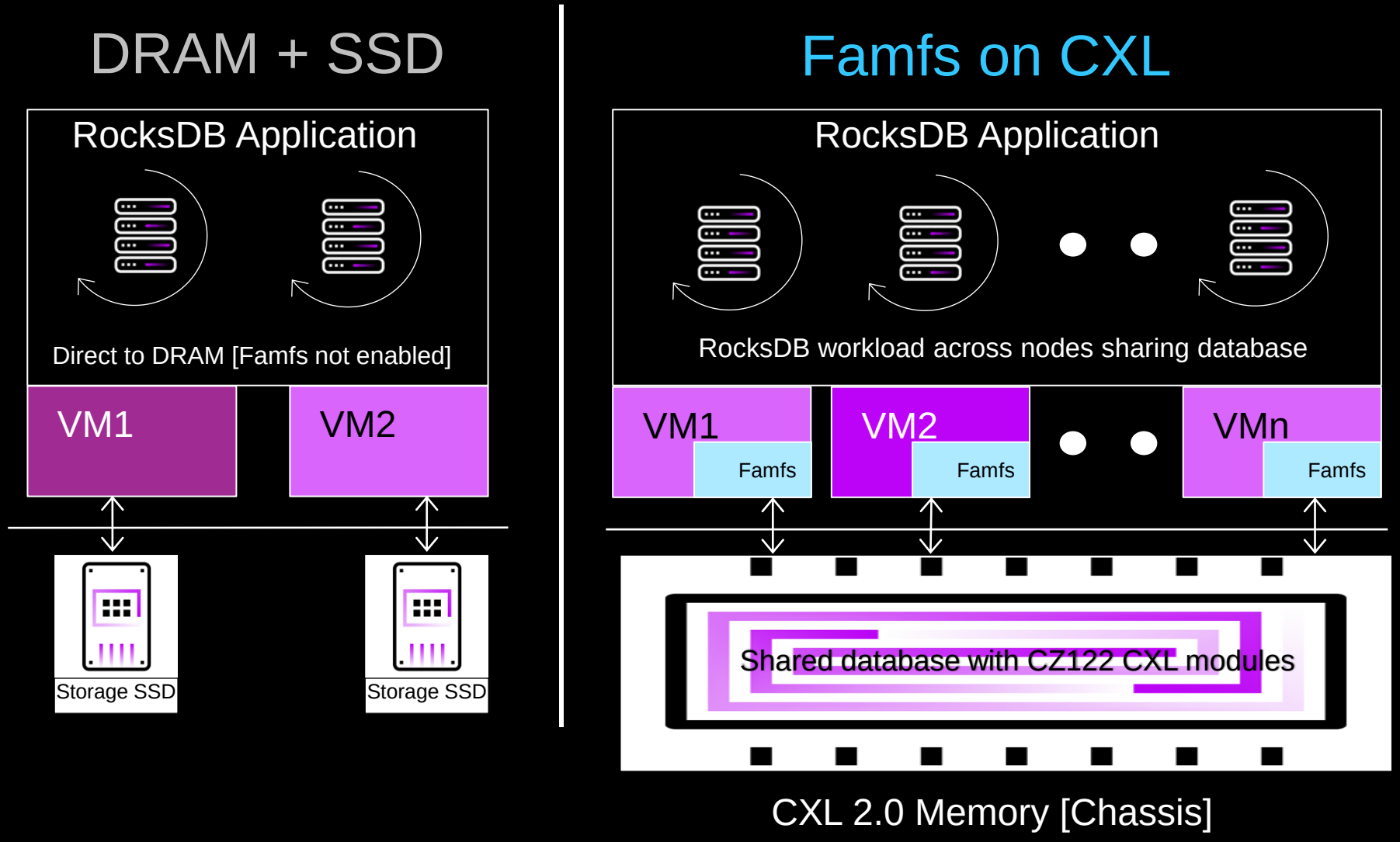


Under exploration: CXL Mem in SAN



RocksDB Apps w/ CXL™ Memory Sharing and Pooling

Famfs + CXL enables sharing & pooling large database across nodes



Problem:

- RocksDB primarily used as caching layer for DBs always benefits from more mem for perf, throughput, and quality
- More capacity limited by server footprint and/or cost

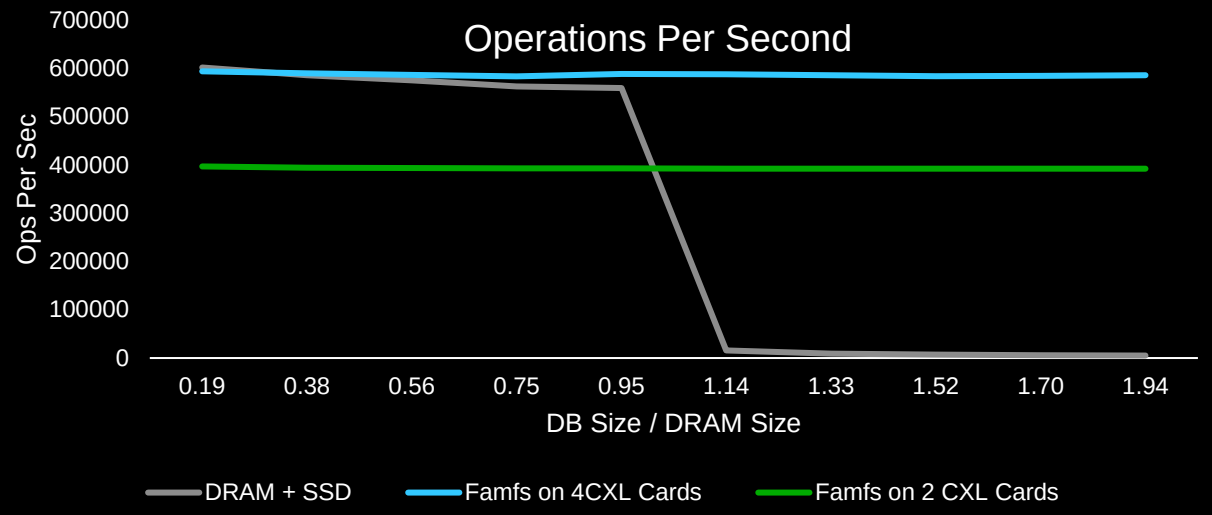
CXL scales up DB server with shared memory

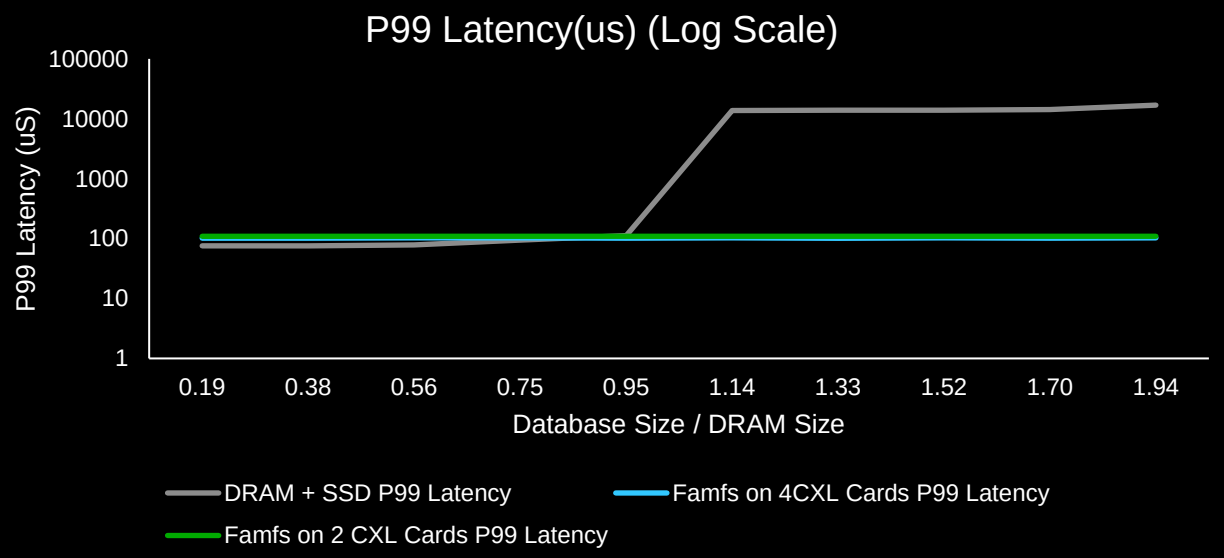
- Elastic performance across datasets larger than RDIMM capacity
- Larger datasets translates to more complex queries, higher valued services
- TCO improvements due to sharing of memory resources

[1] Source: "CXL Memory Expansion: A Closer Look on Actual Platform" whitepaper

RocksDB BenchDB CXL Results (DRAM+SSD, DRAM+CXL™+SSD)

Enables Larger Data Set Sizes for the DB applications





Key Highlights

- CXL enables LARGER Datasets for RocksDB app
  - Avoids steep performance drop shown in chart at 1.0 ratio
- Famfs (shared CXL memory) can still approach system RAM performance by addition of more CXL lanes
- Results based on P99 Latency, Uniform Random Reads and 16 threads

CXL Chassis + Famfs adds TCO benefits

- System RAM allocated for RocksDB execution can be freed for general purpose
- RocksDB data in CXL Mem can be shared and pooled across other RocksDB nodes/VM
- Deduplication saves overall system power

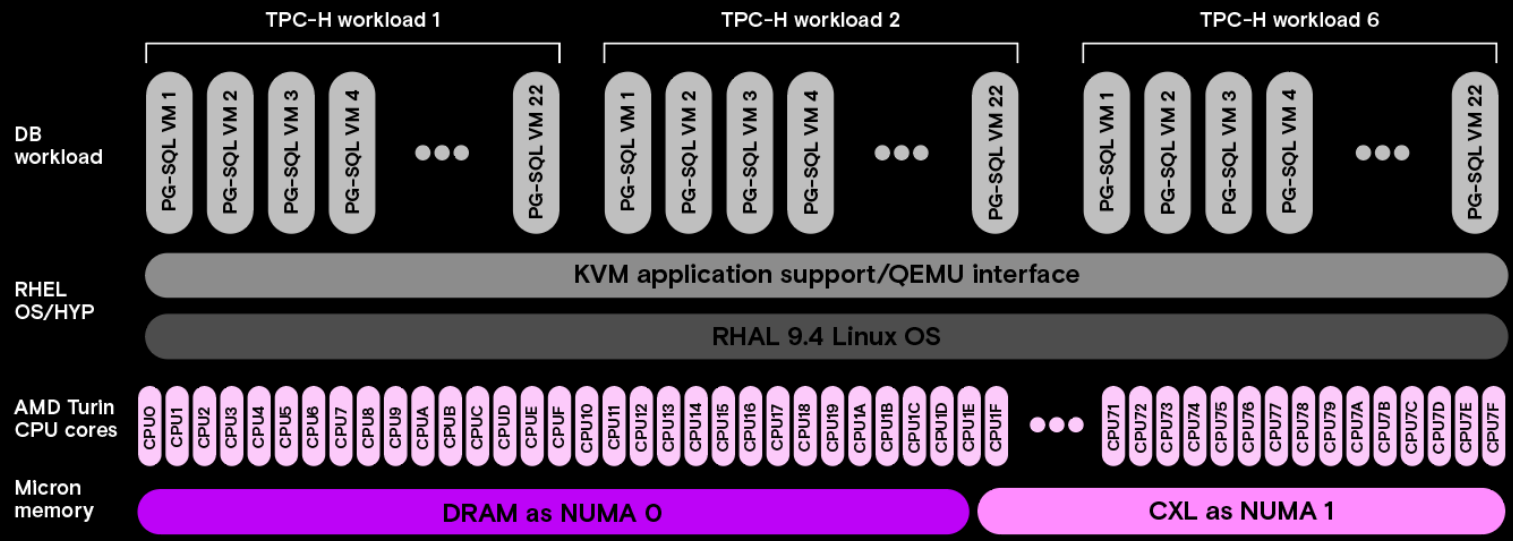
Multi-tenant node for large-scale data analytics

Key technical highlights

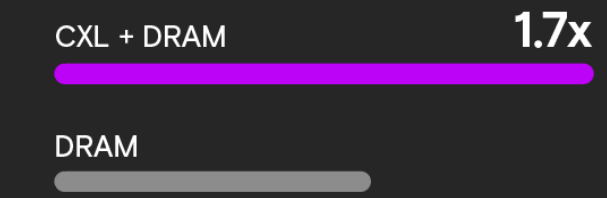
- System throughput increased by 1.7x with CXL memory expansion (scaling up VMs)
- Memory expansion beyond large-capacity RDIMM modules 12x 128GB DDR5 RDIMMs (1.5 TB)

Business highlights

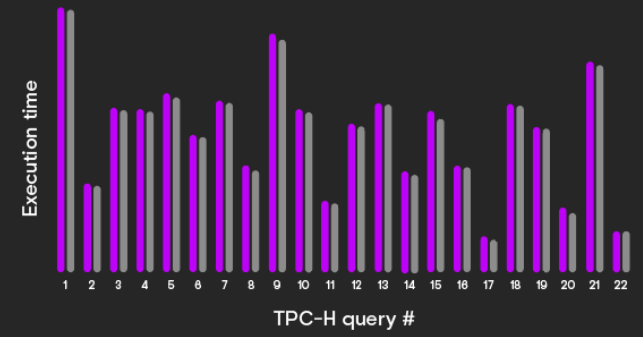
- CXL is a viable path to increased server workload for large-scale data analytics TPC-H type workloads
- Adding CXL memory boosts system performance by 1.7x for memory-bound workloads



VM scale-up using CXL™ memory increase system throughput by 1.66x



<6% (avg) impact per query across VMs



micron

Scaling up with CXL Benefits Summary

Scale out (add more machines)

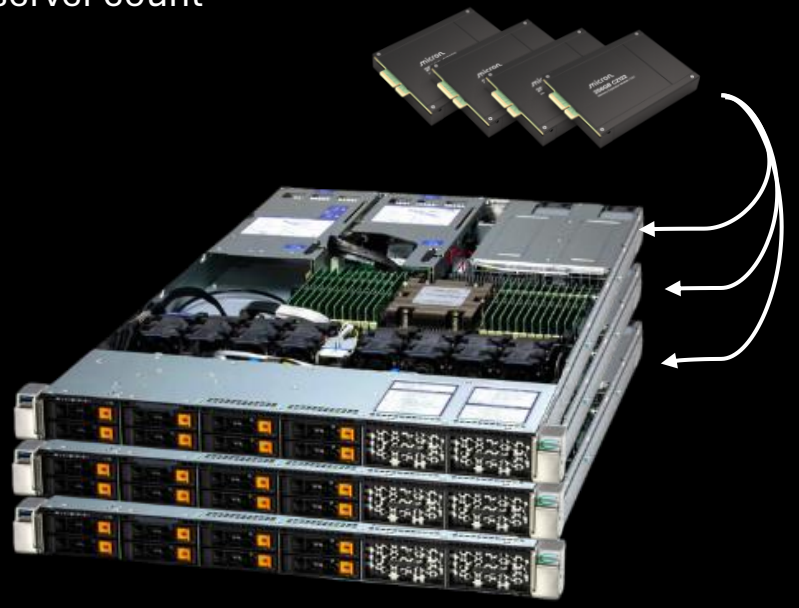
Adds complexity and overhead and can lead to lower core utilization when GB / core is not met for a given workload!

12x128 GB = 1.5 TB (per RDIMM/Server)

Data Sharding + Networking

Scale up (use fewer machines)

Adds costs per server with added CXL, but lowers overall cost with less server count



12x128 GB = 1.5 GB (RDIMM/Server)
4x 256GB = 1TB (CXL/Server)

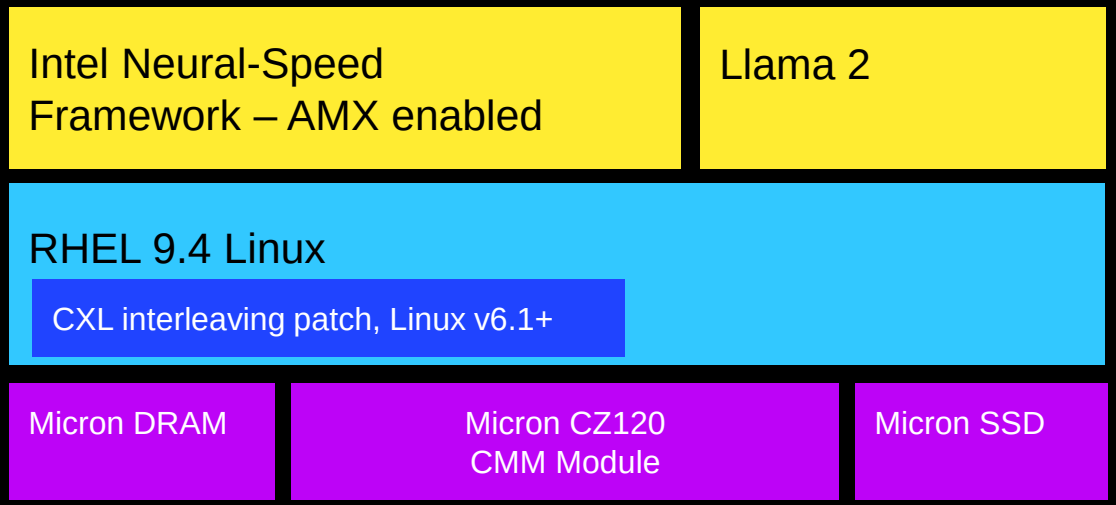
35% TCO saving for customer[3]

1. AMD EPYC, 12 128GB DDR RDIMM, 2 GPU H100, SSD
2. All of 1, plus 4 CZ122 256GB
3. Does not include SW licensing cost

micron

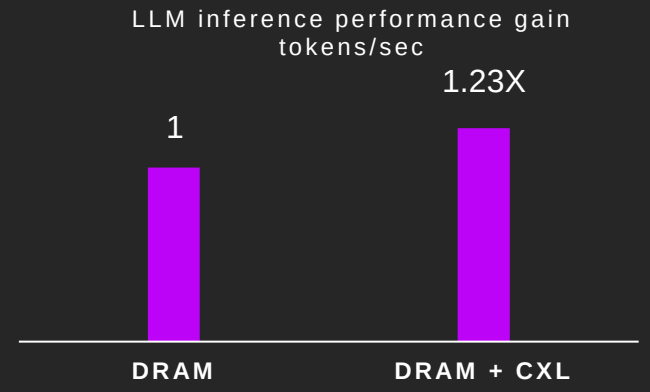
Llama 2 CPU-based Inferencing with CXL™

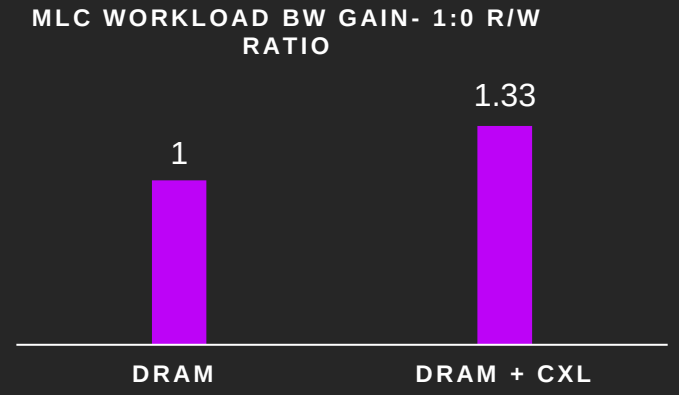
Intel neural speed framework with Llama 2

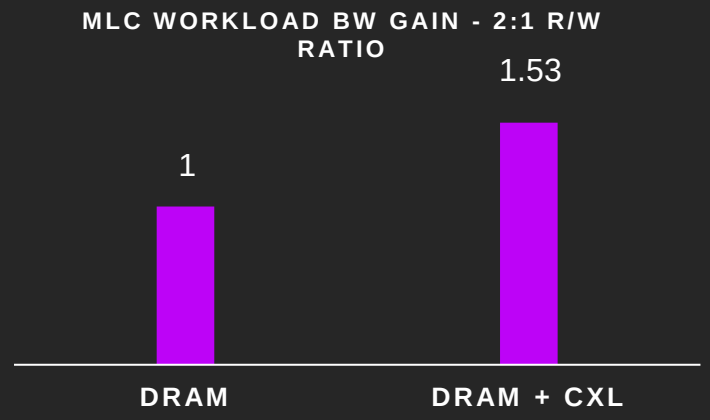


Memory bandwidth reduces time-to-token latency

- >23% improvements with added CXL
- MLC workload – 33% to 53% bandwidth gain
- Enhancing user experience with faster response times
- Increasing throughput to handle more queries
- Enabling more sophisticated LLM pipelines



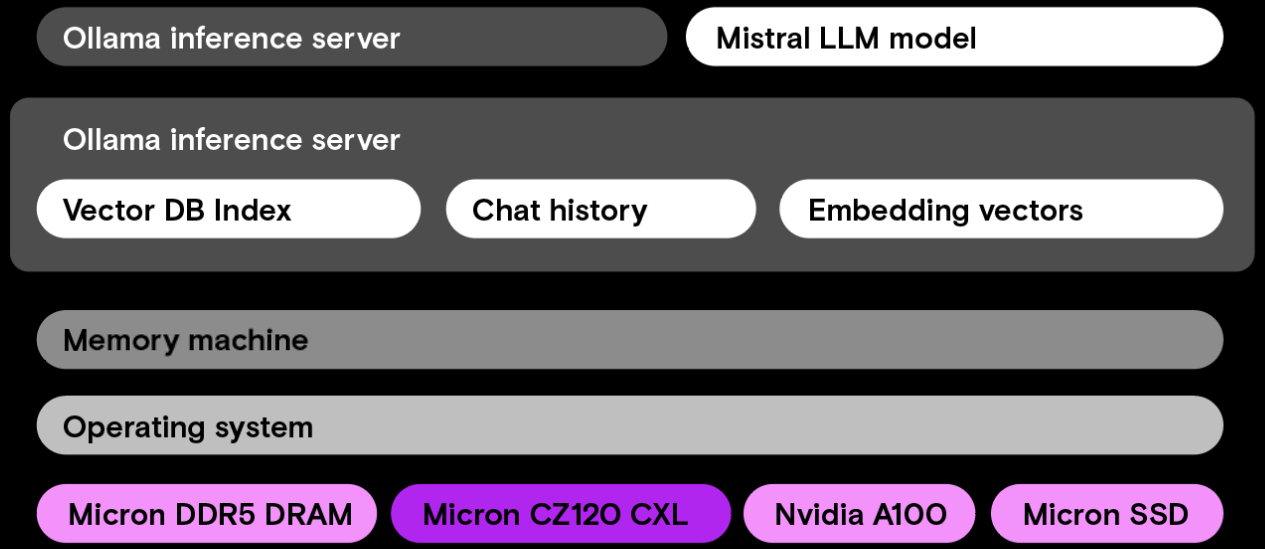




LlamaIndex RAG pipeline with CXL™

RAG pipeline benefits from CXL use

- Improved context retention from more historical data
- Improved user answer quality due to rich retrieved context

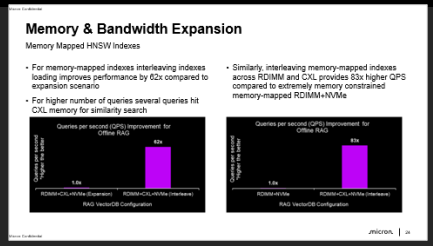


System configuration

Platform	Supermicro platform
CPU family	Intel Xeon Gen 5 processor
Native DRAM	Micron 64GB DDR5-5600MT/s RDIMMs Capacity: 1TB (16 RDIMMs)
CXL DRAM	Micron 256GB CZ120 modules Capacity: 1TB (4 modules) Total bandwidth (100% read): 4 x 26 = 104 GB/s
Storage	Micron 7450 NVMe™ SSD
GPU	Nvidia A100

Software specification

Applications	LlamaIndex, Qdrant Vector Database, Llama 2, MemVerge Memory Machine
OS	Red Hat Enterprise Linux
Kernel	6.8 rc-5 – with weighted software interleaving enabled



AI RAG VectorDB CXL™

High-capacity RDIMM + CXL provides 1.5x performance gain and 30% lower response times for large-scale RAG Vector Databases

RDIMM + NVMe vs. RDIMM + CXL (memory expansion with 4x CZ120)

- RDIMM + CXL improves queries/sec (QPS) by ~8.2x due to higher capacity and faster in-memory computation over RDIMM+NVMe

RDIMM + NVMe vs. RDIMM + CXL + NVMe

- RDIMM + CXL + NVMe improves QPS by ~1.5x, providing a larger memory footprint, enabling faster vector search for large scale databases
- Lowers the response times by ~30% for large scale parallel similarity search computations

Queries per second (QPS) improvement for CXL Memory Expansion[1]
(Higher is better)

RDIMM + CXL (Expansion) **8.2x**

RDIMM + NVMe **1x**

Queries per second (QPS) improvement for large vector DBs[1]
(Higher is better)

RDIMM + CXL + NVMe **1.5x**

RDIMM + NVMe **1x**

Response times for RAG vector search for large vector DBs[1]
(Lower is better)

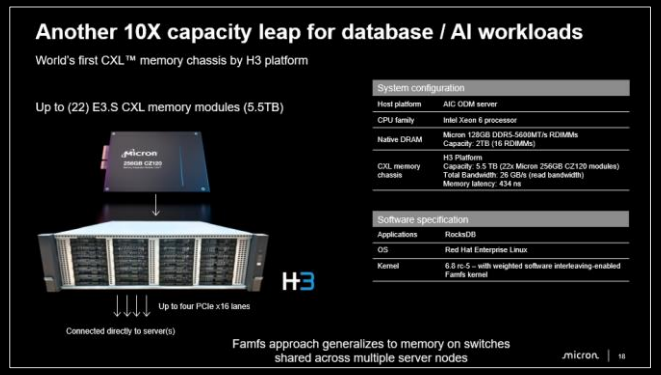
RDIMM + CXL + NVMe **0.7x**

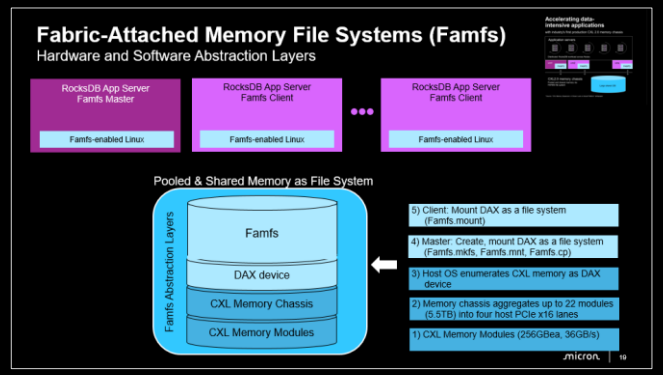
RDIMM + NVMe **1x**

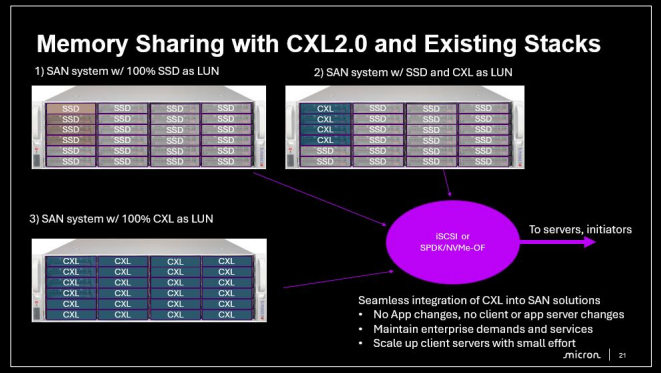
[1] 128GB RDIMM (Total Memory 1.5TB), 4x CZ120 CXL (Total memory: 1TB), 8x 7450 NVMe (Total Storage: 32TB), RAG Footprint: 12TB

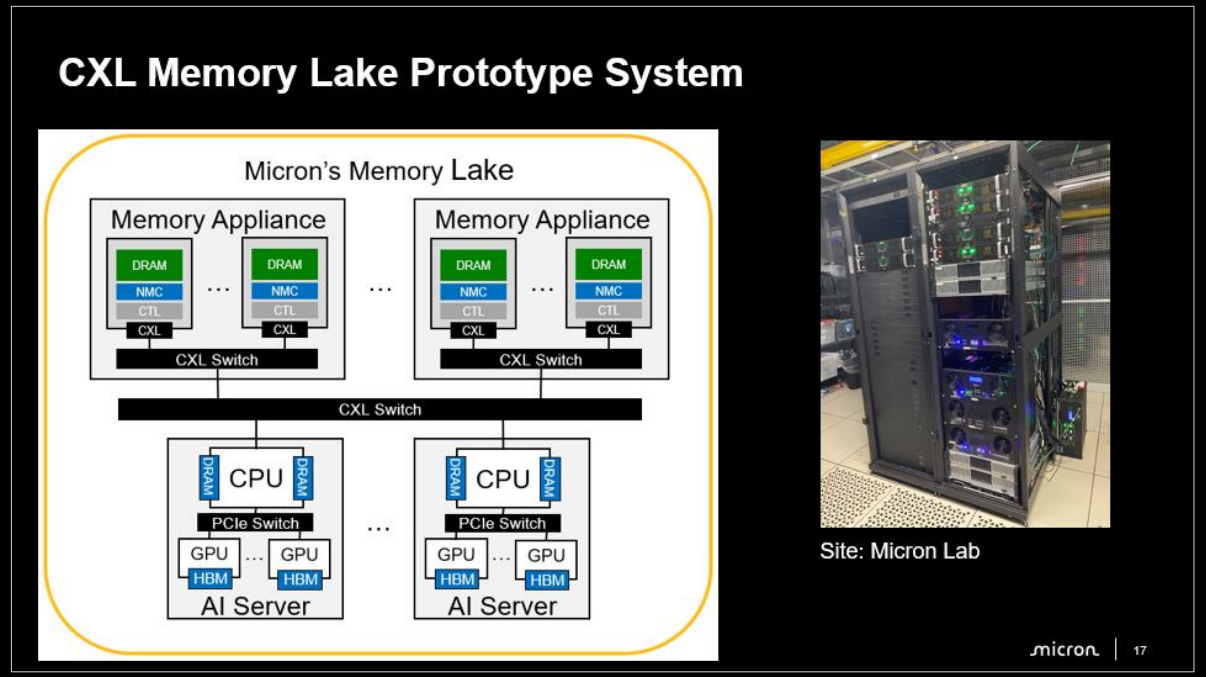
micron

2025 Key Areas for Micron CXL Collaborations









Micron Data Center NVMe™ SSD Portfolio

	High Performance Micron 9000 Series	**Mainstream** Micron 7000 Series	**High Capacity** Micron 6000 Series	**High Endurance** Micron XTR Series
	• Best-in-class performance and efficiency • Designed for AI and performance-critical mixed workloads	• Designed for the broadest range of application workloads • Balanced blend of power, performance, QoS, features	• 20% less power and superior value to competitor QLC • All the advantages of TLC at the price of competitor's QLC	• Extreme endurance (60 SDWPD / 35 RDWPD) • Tier with Micron High Capacity for caching or use independently
Capacities	3.2TB to 30.72TB	400GB to 15.36TB	30.72TB to 122.88TB	960GB to 1.92TB
Interface & Form Factors	**PCIe® Gen5 & Gen6** • U.2 15mm • E1.S 9.5mm/15mm • E3.S-1T 7.5mm	**PCIe Gen4 & Gen5** • U.2/U.3 7mm/15mm • E1.S 5.9mm/15mm • E3.S-1T 7.5mm • M.2 22x80mm & 22x110mm	**PCIe Gen 4 & Gen5** • U.2/U.3 15mm • E1.L 9.5mm • E3.S-1T 7.5mm	**PCIe Gen4** • U.3 15mm