Enhance AI & Database Performance: Experience Micron's CZ122 at OCP

Vishnu Balraj – Product Management
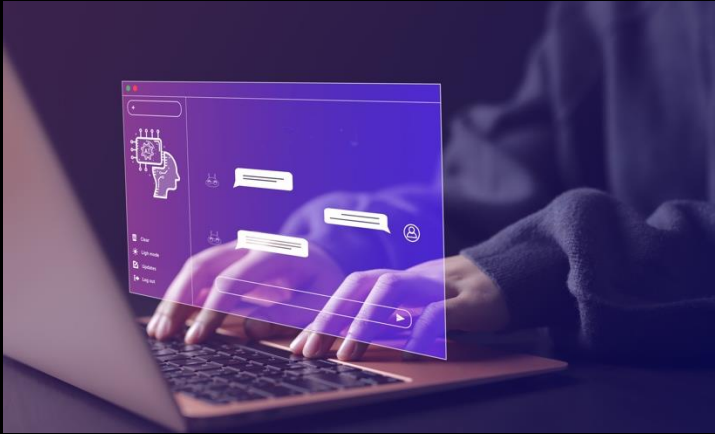Micron Technology, Inc.

Agenda

1. Data center challenges

2. Memory and storage hierarchy

3. Micron CZ120/CZ122 memory modules

4. CZ120/CZ122 use cases & demos

5. Enabling CXL in your company

Data Center Challenges
Scaling for today's data center growth plus tomorrow's AI disruptive needs

Higher Capacity

In-Memory Databases, SaaS, AI training and inference, General Purpose compute continue to grow exponentially, requiring unconstrained memory capacity to enable compute to continue to scale.

More Bandwidth

The explosion of processor core counts and the real-time nature of business is necessitating memory in the data center to provide faster data access to keep compute fed.
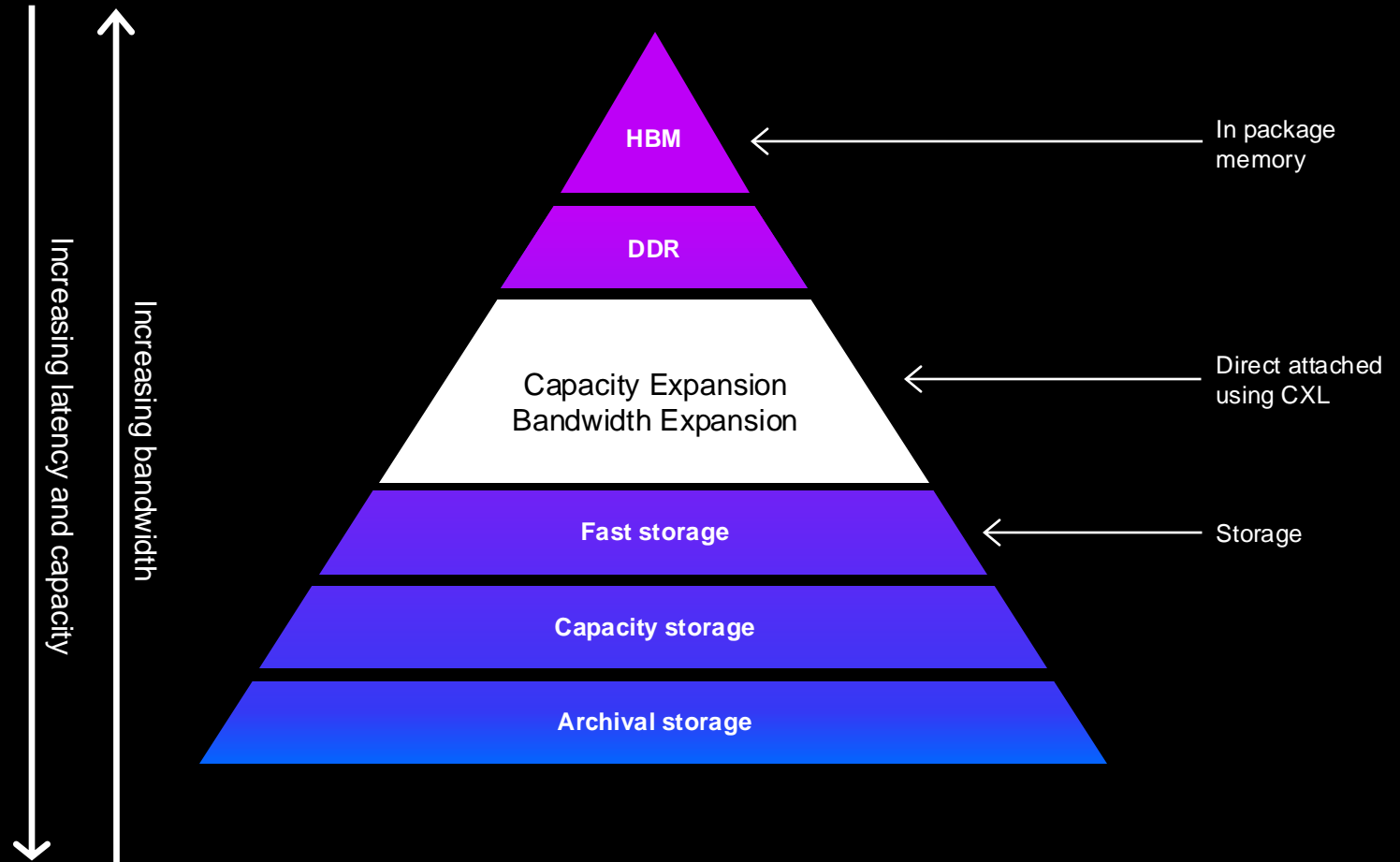

Lower Data Center TCO

Data center operators drive to optimize TCO to deliver the exact server memory capacity and bandwidth that workloads require.

Memory and Storage Hierarchy

Increasingly more workloads are becoming memory bound and the data center needs a solution

Micron Memory Expansion Modules using CXL™ are the solution by adding memory capacity and bandwidth

Increasing latency and capacity

Increasing bandwidth

HBM — In package memory

DDR

Capacity Expansion Bandwidth Expansion — Direct attached using CXL

Fast storage — Storage

Capacity storage

Archival storage

.micron | 4

Introducing Micron CZ120 Memory

Delivering capacity, bandwidth, flexibility

128GB / 256GB

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

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

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

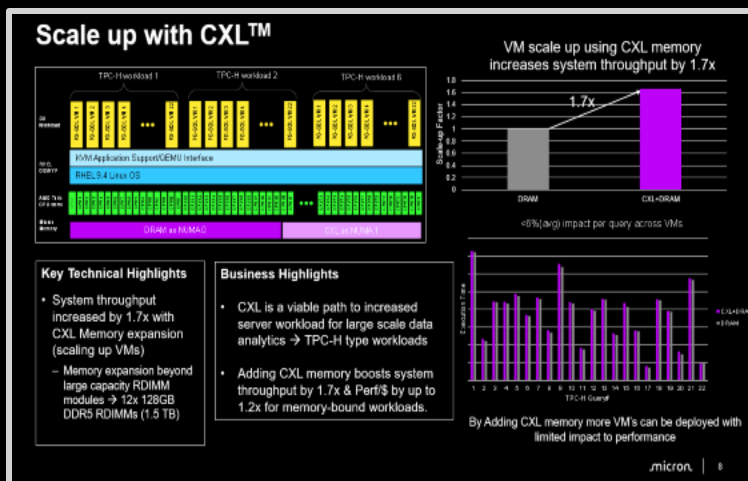1. By adding 8x256GB CZ12x modules, system limitations apply
2. MLC (2R1W) was used to measure BW and latency with Micron and competition
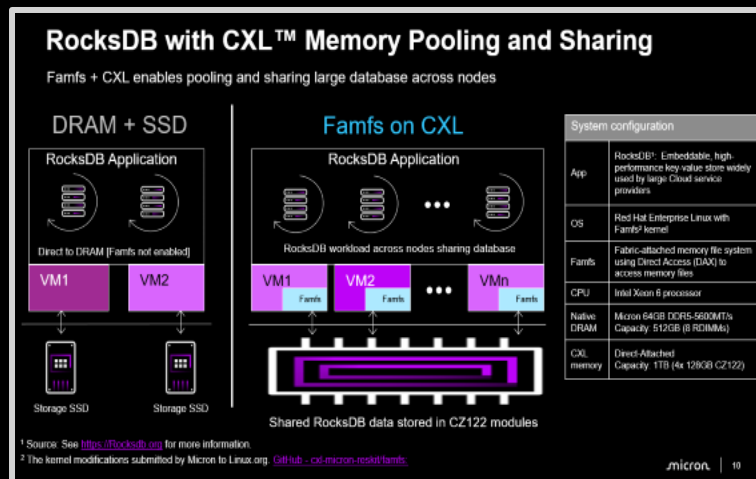3. Supports heterogeneous Interleave and Meta Data on Major CPU Platforms
4. Supports Internal CVME threshold, Device Initiated hPPR
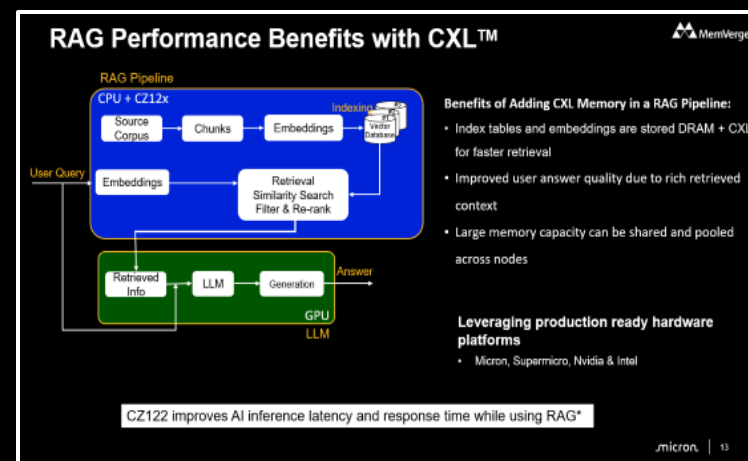
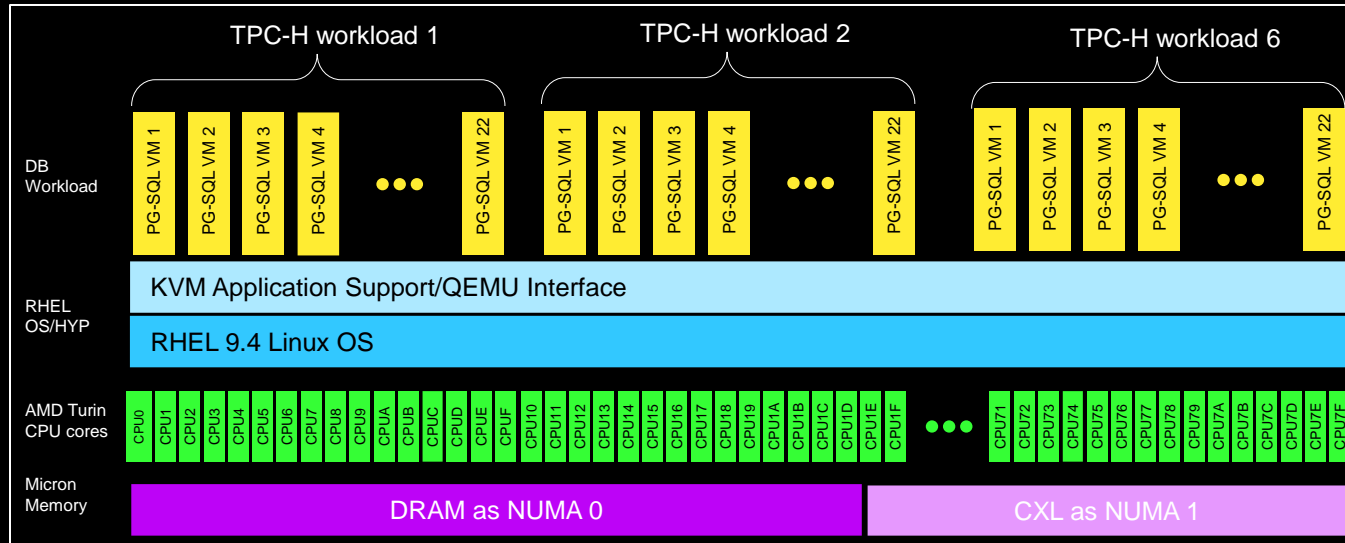# Micron CZ122 Use Cases & Demos

Scale up with CXL

- Contributors : AMD/SMC

- CXL allows more VM instances to be run on a server without adding new servers [Scale-up]

- CXL boosts Perf/$ up to 1.2x for memory-bound workloads.

Improved RocksDB perf with CXL

- Contributor: H3 Platform

- Shared CXL memory via Famfs provides higher sustained OPS, improved p99 latency during for large DB accesses

CXL RAG Inference

- Contributor: MemVerge

- CZ12x improves AI inference latency and response time while using RAG*

*performance improvement is driven via the vector DB being accessed from CXL memory instead of storage

Scale up with CXL™

TPC-H workload 1　　TPC-H workload 2　　TPC-H workload 6

DB Workload: PG-SQL VM 1, PG-SQL VM 2, PG-SQL VM 3, PG-SQL VM 4 ... PG-SQL VM 22 | PG-SQL VM 1, PG-SQL VM 2, PG-SQL VM 3, PG-SQL VM 4 ... PG-SQL VM 22 | PG-SQL VM 1, PG-SQL VM 2, PG-SQL VM 3, PG-SQL VM 4 ... PG-SQL VM 22

RHEL OS/HYP: KVM Application Support/QEMU Interface — RHEL 9.4 Linux OS

AMD Turin CPU cores: CPU0 CPU1 CPU2 CPU3 CPU4 CPU5 CPU6 CPU7 CPU8 CPU9 CPUA CPUB CPUC CPUD CPUE CPUF CPU10 CPU11 CPU12 CPU13 CPU14 CPU15 CPU16 CPU17 CPU18 CPU19 CPU1A CPU1B CPU1C CPU1D CPU1E CPU1F ... CPU71 CPU72 CPU73 CPU74 CPU75 CPU76 CPU77 CPU78 CPU79 CPU7A CPU7B CPU7C CPU7D CPU7E CPU7F

Micron Memory: DRAM as NUMA 0 | CXL as NUMA 1
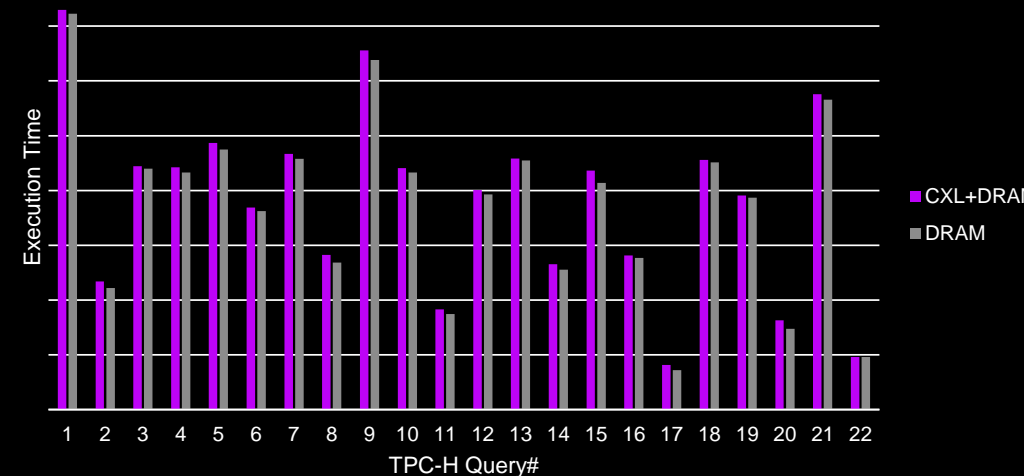
Key Technical Highlights

- 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 throughput by 1.7x & Perf/$ by up to 1.2x for memory-bound workloads.

VM scale up using CXL memory increases system throughput by 1.7x

Scale-up Factor (y-axis: 0 to 1.8)

1.7x

DRAM | CXL+DRAM
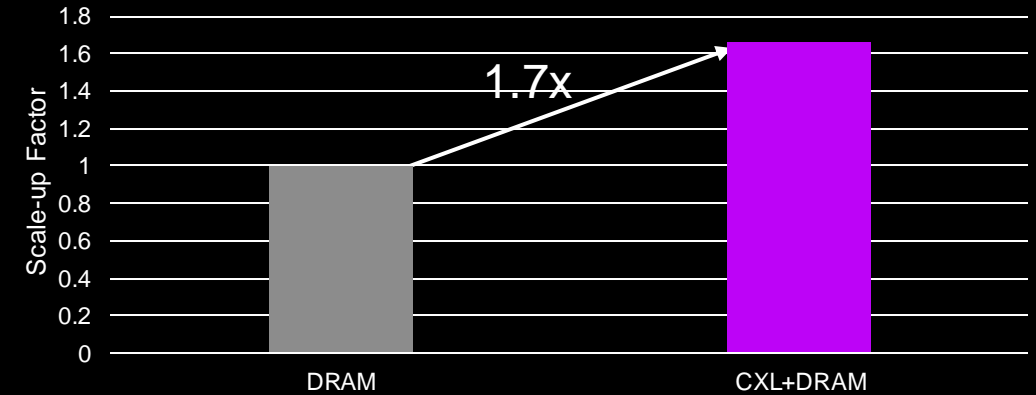
<6%(avg) impact per query across VMs

Execution Time vs TPC-H Query# (1–22)

Legend: CXL+DRAM, DRAM

By Adding CXL memory more VM's can be deployed with limited impact to performance

Multi-Tenant Node for Large Scale Data Analytics
Production Ready System Configuration

AMD 5th Gen EPYC™ 128 core

12x 128GB DDR5-6400MT/s

4x 256GB CZ122 – PCIe Gen 5 x8
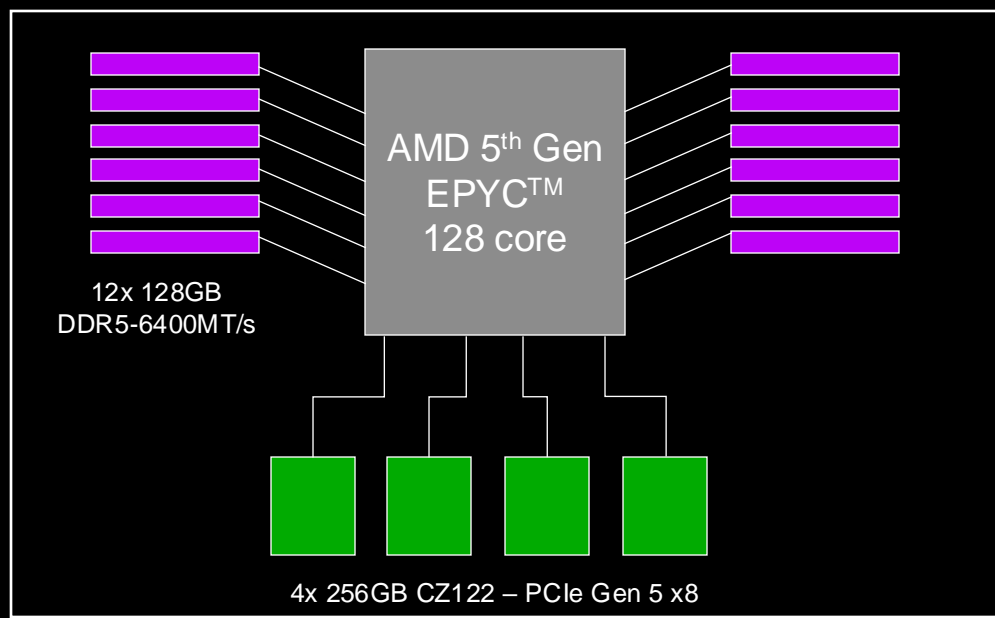
Supermicro H13 Petascale Storage Server

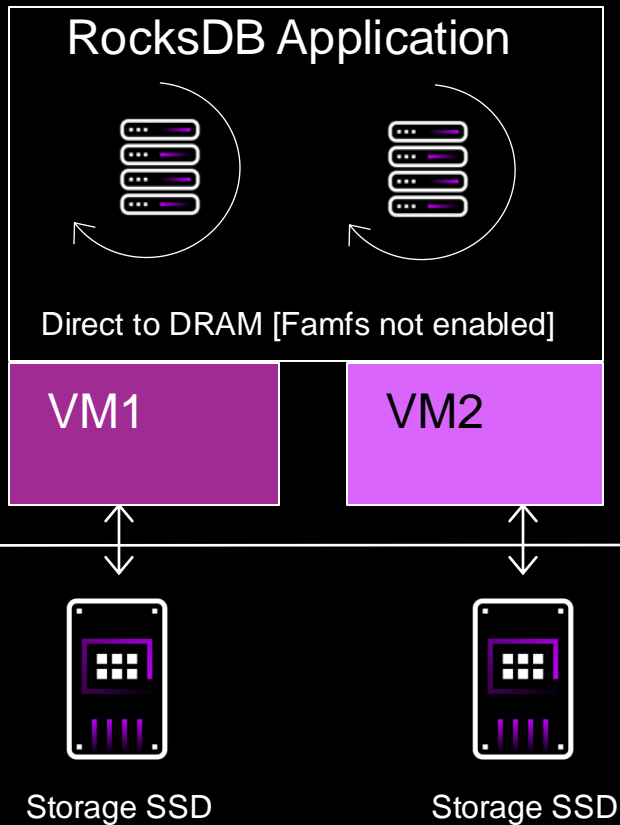| System configuration	
Host platform	Supermicro H13 Petascale Storage Server
CPU family	AMD 5th Gen EPYC™
Native DRAM	Micron 128GB DDR5-6400MT/s RDIMM Capacity: 1.5TB (12 RDIMMs)
CXL memory	Micron 256GB CZ122 memory module Capacity: 1TB (4x E3.S modules) Total Bandwidth: 148GB/s (37GB/s)
Storage	Micron 7450 SSDs

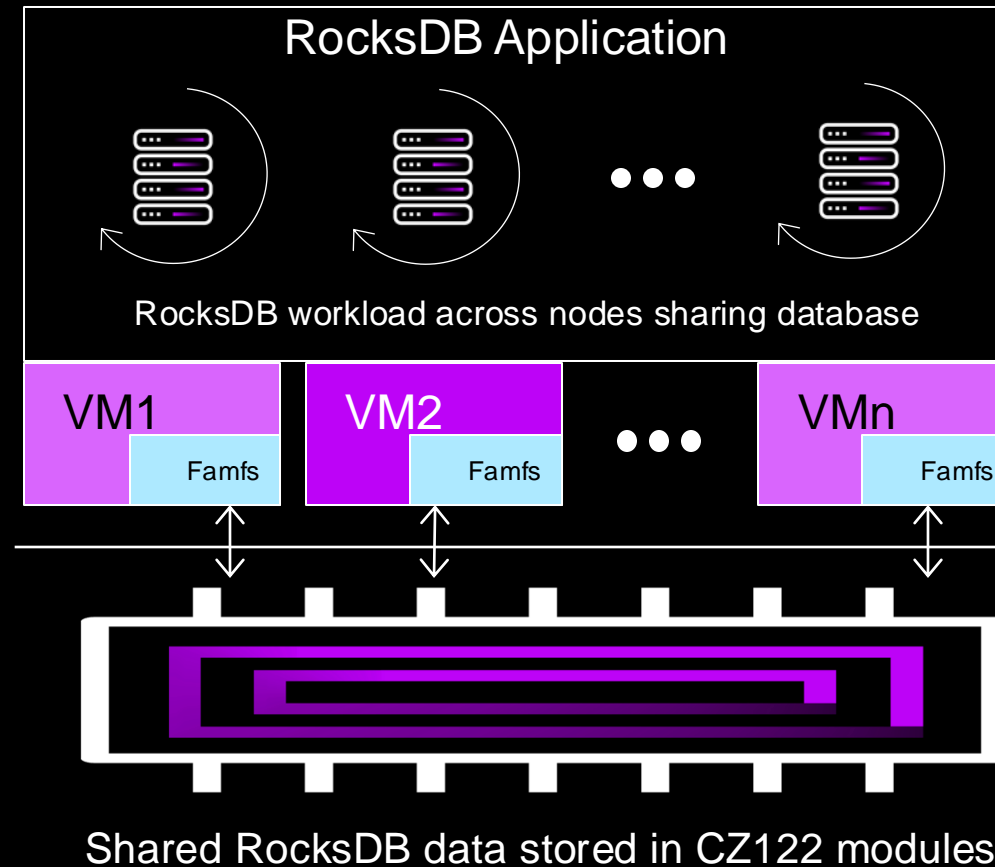| Software specification	
Applications	PostgreSQL Database, QEMU/KVM
OS	Red Hat Enterprise Linux v9.4
Kernel	6.11.0-1 el9.elrepo.x86_64

RocksDB with CXL™ Memory Pooling and Sharing

Famfs + CXL enables pooling and sharing large database across nodes

DRAM + SSD

RocksDB Application

Direct to DRAM [Famfs not enabled]

VM1 VM2

Storage SSD   Storage SSD

Famfs on CXL

RocksDB Application

RocksDB workload across nodes sharing database

VM1 — Famfs   VM2 — Famfs   • • •   VMn — Famfs

Shared RocksDB data stored in CZ122 modules

System configuration	
App	RocksDB[1]: Embeddable, high-performance key-value store widely used by large Cloud service providers
OS	Red Hat Enterprise Linux with Famfs[2] kernel
Famfs	Fabric-attached memory file system using Direct Access (DAX) to access memory files
CPU	Intel Xeon 6 processor
Native DRAM	Micron 64GB DDR5-5600MT/s Capacity: 512GB (8 RDIMMs)
CXL memory	Direct-Attached Capacity: 1TB (4x 128GB CZ122)

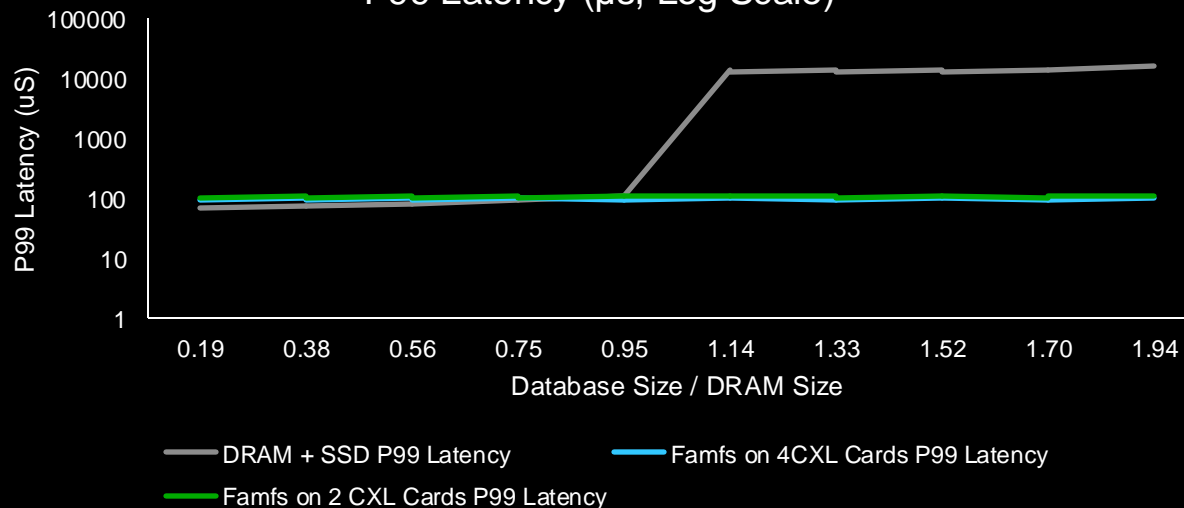[1] Source: See https://Rocksdb.org for more information.

[2] The kernel modifications submitted by Micron to Linux.org. GitHub - cxl-micron-reskit/famfs:

RocksDB BenchDB Results (DRAM+SSD vs. CXL™)

Operations Per Second
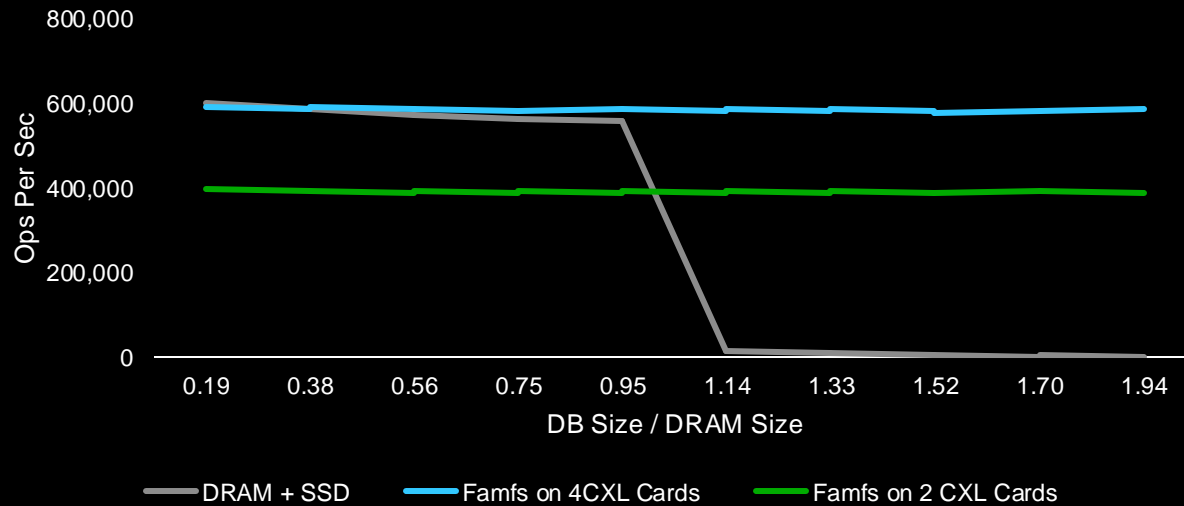
P99 Latency (μs, Log Scale)

Benchmark Highlights

CXL enables high throughput for larger DB sizes that exceed memory capacity

– Avoids performance drop when DB size exceeds DRAM

– RocksDB data stored in CXL instead of DRAM+SSD

– Results based on P99 Latency, Uniform Random Reads and 16 threads

Additional TCO benefits

Large Pool of Memory can be shared across nodes/VMs

– With direct attached: Up to 1TB with 4x CZ122 256GB

– With Chassis: Up to 5.5TB with 22x CZ122 256GB

• System RAM can be freed for other compute purposes

• Deduplication saves overall system power

World's First CXL™ Memory Chassis by H3 Platform

Famfs approach generalizes to memory on switches shared across multiple server nodes

Up to 22x 256GB CZ122 memory modules (5.5TB)

micron
256GB CZ122
Memory Expansion Module | CXL™
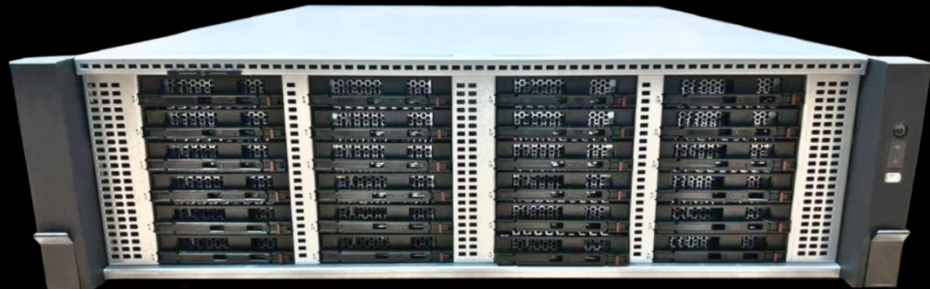
H3

Up to four PCIe x16 lanes

Connected directly to server(s)
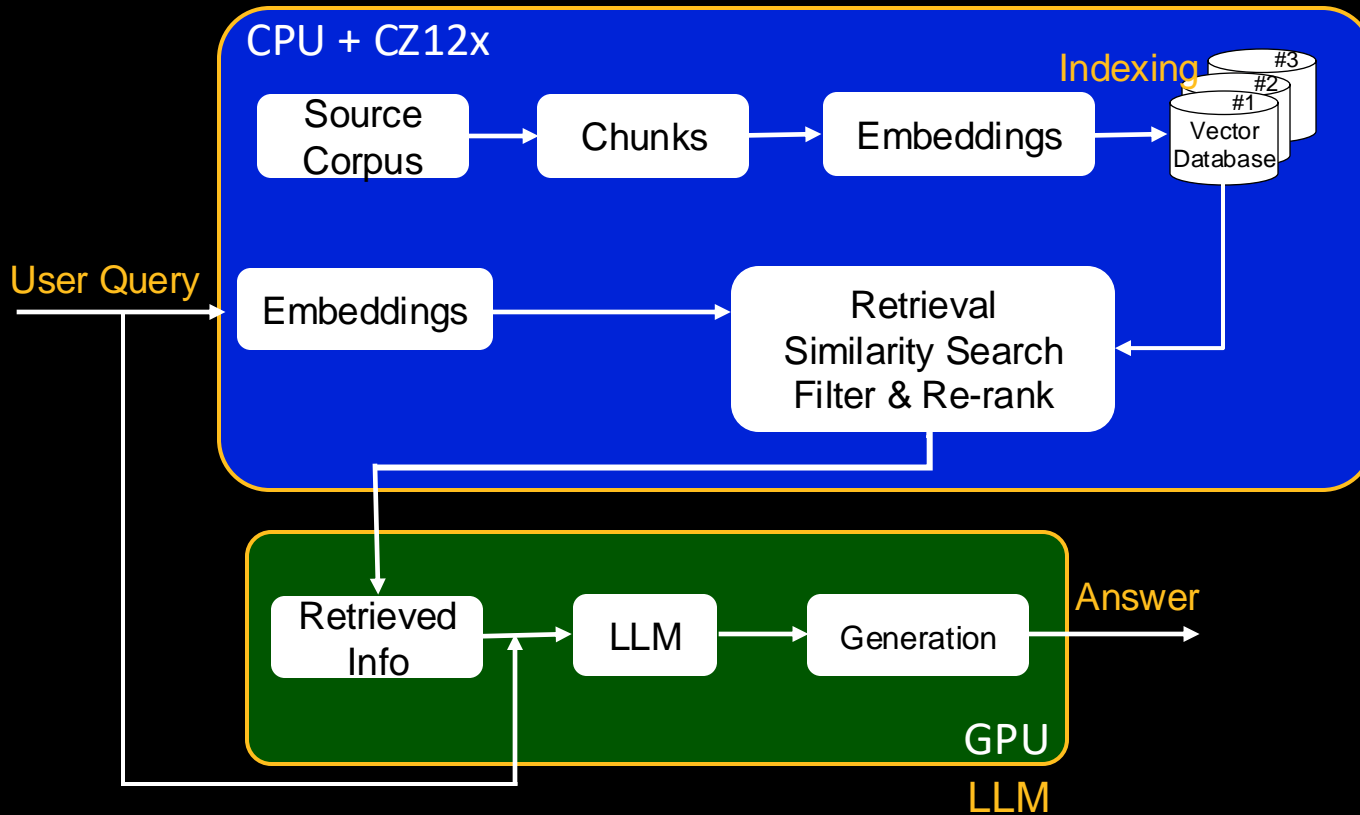
System configuration	
Host platform	AIC ODM server
CPU family	Intel Xeon 6 processor
Native DRAM	Micron 128GB DDR5-5600MT/s RDIMM Capacity: 2TB (16 RDIMMs)
CXL memory chassis	H3 Platform Capacity: 5.5TB (22x Micron 256GB CZ122 modules) Total Bandwidth: 26 GB/s (100% read bandwidth) Memory latency: 434 ns

Software specification	
Applications	RocksDB
OS	Red Hat Enterprise Linux
Famfs	Fabric-attached memory file system using Direct Access (DAX) to access memory files
Kernel	6.8 rc-5 – with weighted software interleaving-enabled Famfs kernel

RAG Performance Benefits with CXL™

RAG Pipeline

CPU + CZ12x

```
Source          Chunks          Embeddings          Indexing
Corpus                                               Vector
                                                     Database #1 #2 #3

Embeddings                      Retrieval
                                Similarity Search
                                Filter & Re-rank
```

User Query

```
Retrieved       LLM             Generation          Answer
Info
                                                     GPU
```

LLM

Benefits of Adding CXL Memory in a RAG Pipeline:

- Index tables and embeddings are stored DRAM + CXL for faster retrieval

- Improved user answer quality due to rich retrieved context

- Large memory capacity can be shared and pooled across nodes

Leveraging production ready hardware platforms

- Micron, Supermicro, Nvidia & Intel

CZ122 improves AI inference latency and response time while using RAG*
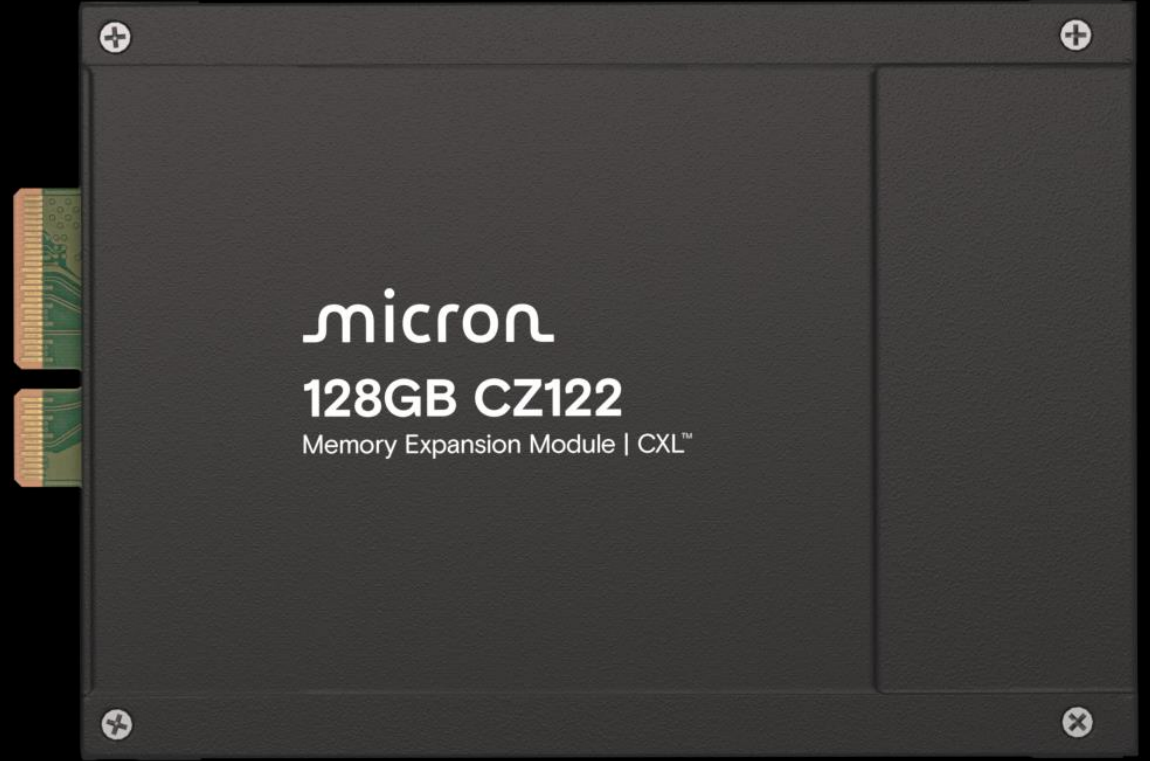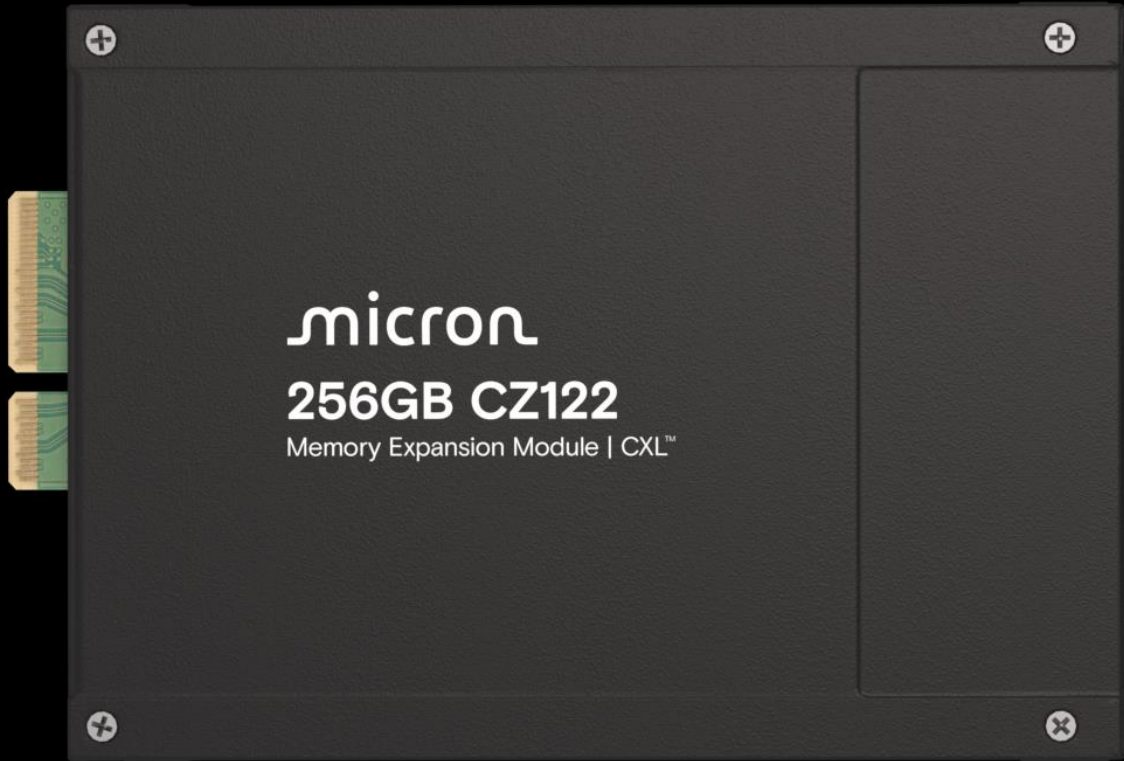
Learn more about CXL™ and Micron CZ122

Qualify your server platforms by enrolling in the Micron Technology Enablement Program (TEP) and leverage our world-class collaboration

- Hands-on support to aid in the development of CXL-enabled designs

- Technical resources including data sheets, electrical and thermal models to aid in product development and evaluation, and engineering consultation related to signal integrity and other technical support topics

- Access to other ecosystem partners who can aid in system-level design

Learn more www.micron.com/CXL

Micron

RocksDB BenchDB Results (DRAM+SSD, CXL)

P99 Latency(us)