



CXL and the New AI Memory Hierarchy

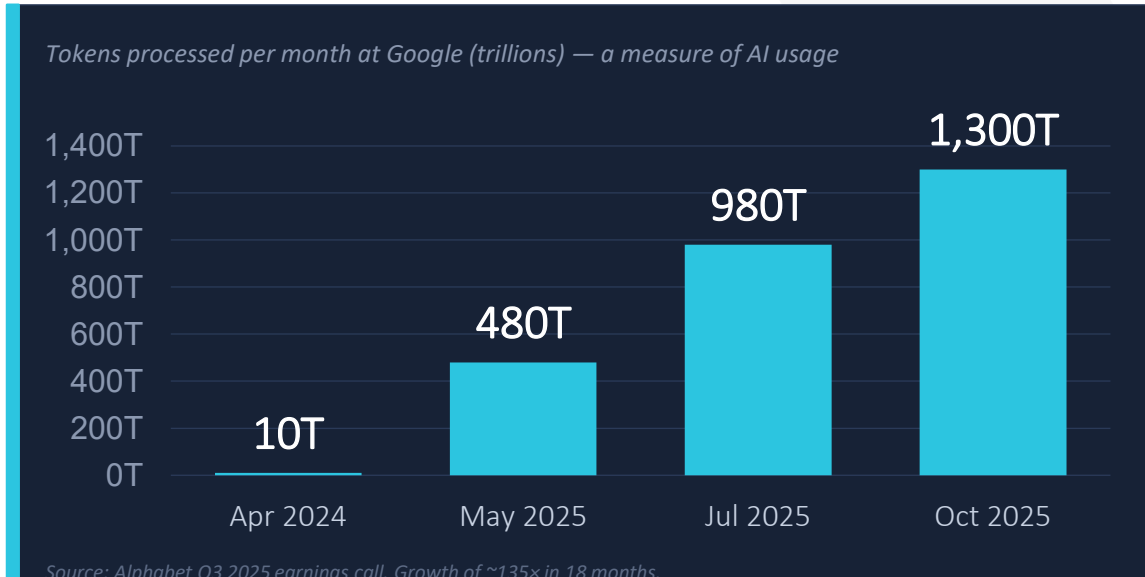
Endless Memory with CXL | Episode 1 | September 30, 2026

AI Demand Grows and Shift to Inference

As AI usage explodes and drives massive capital expenditures

MASSIVE AI DEMAND

AI Use Growing Exponentially



\$30B

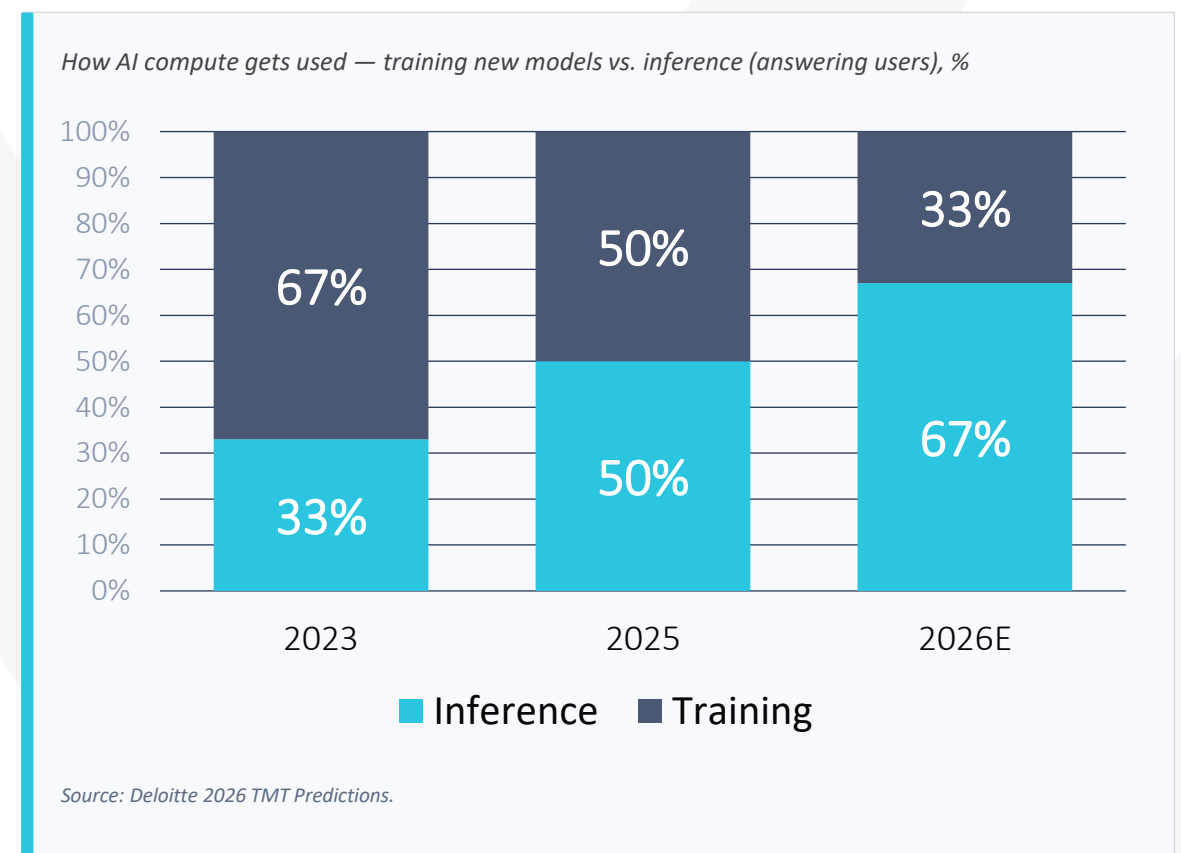
ANTHROPIC REVENUE (30x IN 15 MONTHS)
Anthropic disclosure, April 2026

\$725B

BIG TECH AI INFRASTRUCTURE SPEND, 2026
Amazon, Microsoft, Google, Meta combined

THE SHIFT TO INFERENCE

Using AI will outweigh building AI



AI Inference Drives Memory Consumption

Rapid growth of AI inference is driving massive demand for larger memory footprint across the AI stack

Limited Memory Supply

3x

Global Suppliers

**Memory Supply Is
Structurally Constrained**

**Memory is scarce, expensive, unable to
scale with AI inference demand**

Bigger Context Windows

250x

Context Growth

**Context Windows Are
Exploding Key-Value (KV) Cache Demand**

**KV Cache scales with context length,
driving exponential memory demand**

Memory Drives Cost

80%

HBM → KV cache

**Inference Economics
Are Breaking Down**

**Expensive HBM for KV cache is driving
unsustainable inference costs**

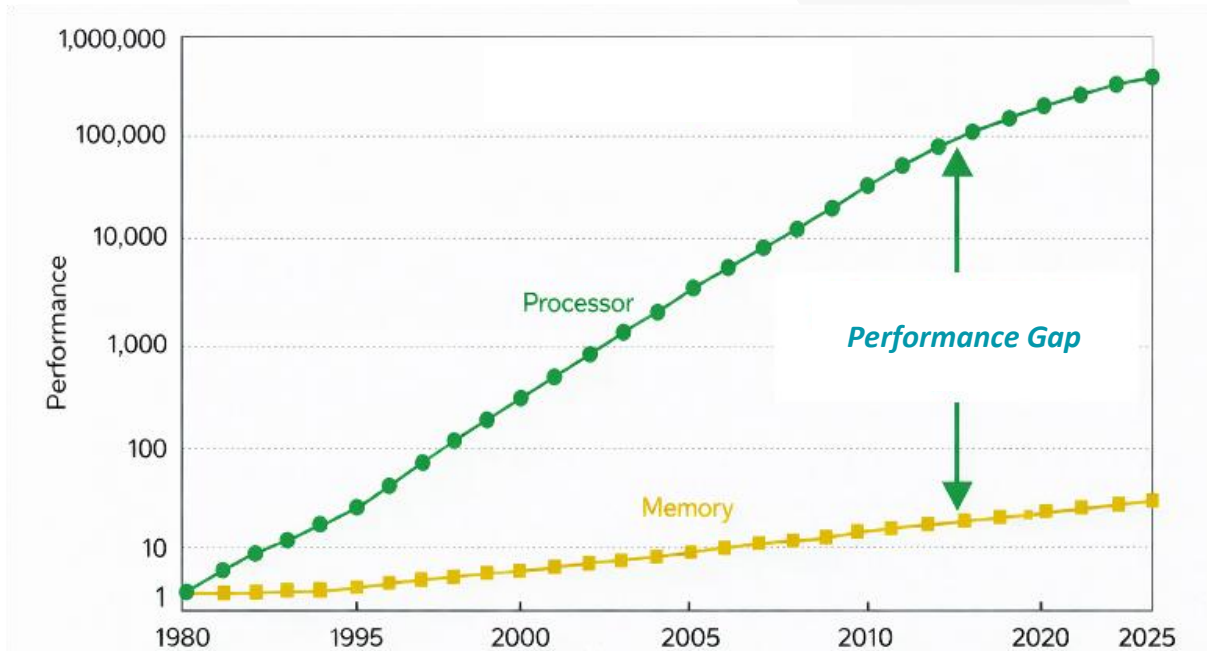
Larger Context Window Drives Higher Memory Demand for KV Cache Performance

Breaking the “Memory Wall”

Liquid’s resource pooling eliminates GPU and memory bottlenecks, enabling efficient scaling and higher performance

THE MEMORY WALL

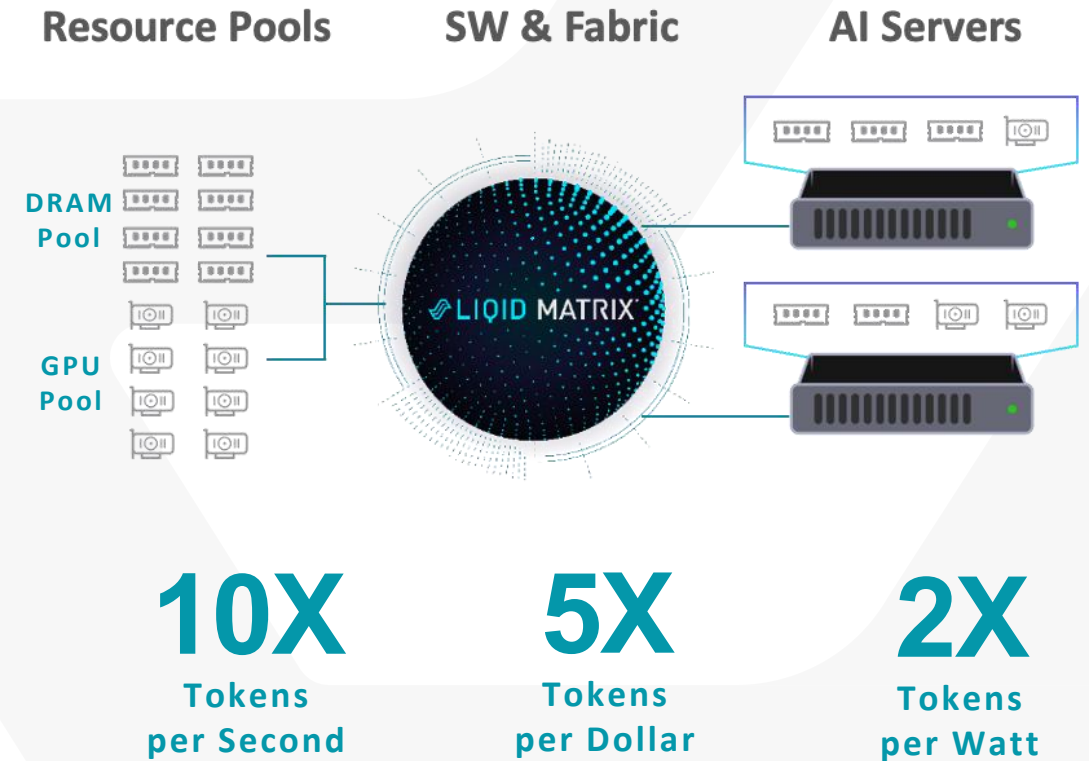
Processor (GPU) and Memory Performance Gap



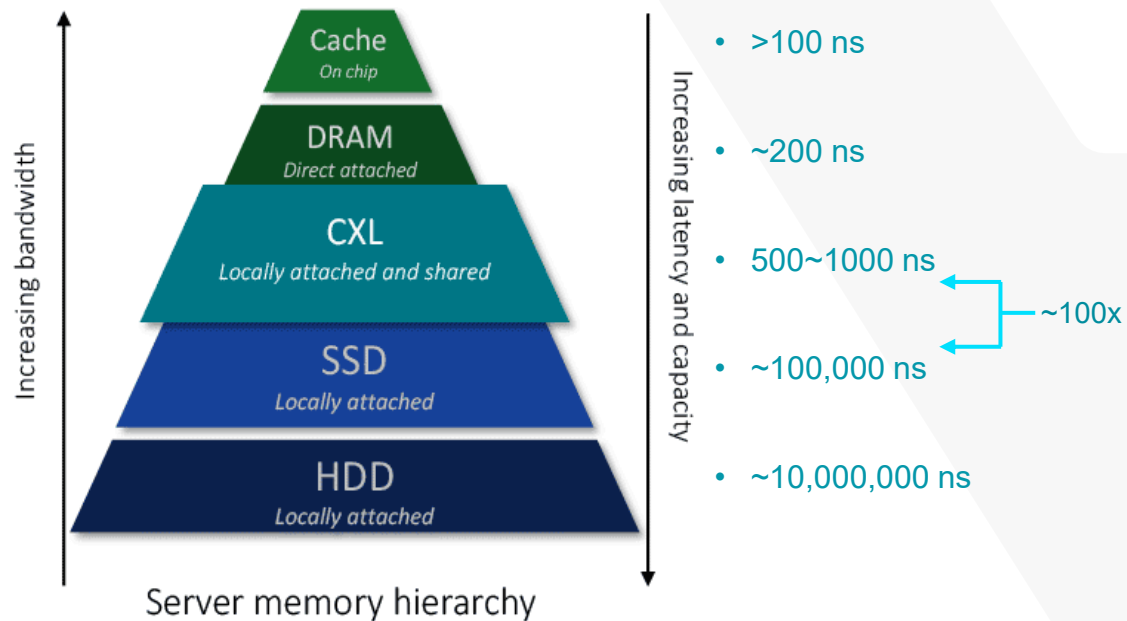
GPU innovation is outpacing memory performance, creating a gap that can be closed through intelligent software

LIQID MEMORY FABRIC

Enable GPU and Memory Pooling and Scaling



Transition to Compute Express Link (CXL) for Lower Latency Applications



CXL Provides 100x-200x Lower Latency Over NVMe

~100x Lower Latency

~100us
NVMe

<1us
CXL



Workload Shift from NVMe to CXL for High-Performance Database Applications

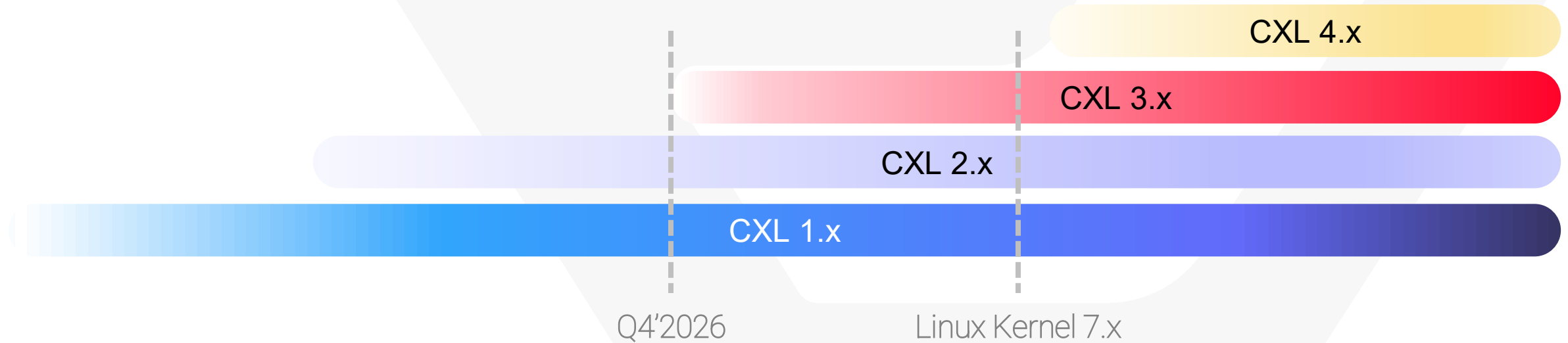


CXL Features by Specification Revision

CXL Specification Feature Summary

Features	Not Supported ✓ Supported			
	CXL 1.0 / 1.1	CXL 2.0	CXL 3.x	CXL 4.0
Release date	2019	2020	2022-2024	2025
Max link rate	32GTs	32GTs	64GTs	128GTs
Flit 68 byte (up to 32 GTs)	✓	✓	✓	✓
Flit 256 byte (up to 64 GTs)			✓	✓
Type 1, Type 2 and Type 3 Devices	✓	✓	✓	✓
Memory Pooling w/ MLDs		✓	✓	✓
Global Persistent Flush		✓	✓	✓
CXL IDE		✓	✓	✓
Switching (Single-level)		✓	✓	✓
Switching (Multi-level)			✓	✓
Direct memory access for peer-to-peer			✓	✓
Enhanced coherency (256-byte flit)			✓	✓
Memory sharing (256-byte flit)			✓	✓
Multiple Type 1/Type 2 devices per root port			✓	✓
Fabric capabilities (256-byte flit)			✓	✓
Back invalidate capabilities on Type 3 devices (HDM-DB)			✓	✓
Fabric Manager API definition for PBR Switch			✓	✓
Host-to-Host communication with Global Integrated Memory (GIM) concept			✓	✓
Trusted-Execution-Environment (TEE) Security Protocol			✓	✓
Memory expander enhancements (up to 32-bit of meta data, RAS capability enhancements)			✓	✓
Security, compliance, and CXL Memory Device enhancements			✓	✓
CXL Bundled Port				✓
Memory RAS enhancements (granular event reporting, Post Package Repair (PPR), and flexible memory sparing operations)				✓

CXL: Timeline and Market Adoption



CXL Market Timeline

- 2023: CXL 1.x @ PCIe-4 Initial products released to market
- 2024: CXL 2.x @ PCIe-5 First OEMs release memory expansion
- 2025: CXL 2.x @ PCIe-5 Memory Pooling and Sharing released
- 2026: CXL 3.x @ PCIe-5 CXL 3.x enhanced features developed
- 2027: CXL 3.x @ PCIe-6 CXL 3.x with advanced user applications
- 2028: CXL 4.x @ PCIe-6 CXL 4.x initial product introduction

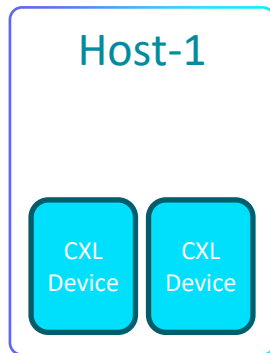
Liquid CXL Timeline

- 2023: Advanced CXL technology demo (hot plug)
- 2024: Development of CXL production system
- 2025: Product launch of CXL 2.x memory appliance
- 2026: Development of advanced CXL 3.x solution

Liquid leads the market in CXL technology, delivering complete solutions for dynamic resource allocation with disaggregated infrastructure

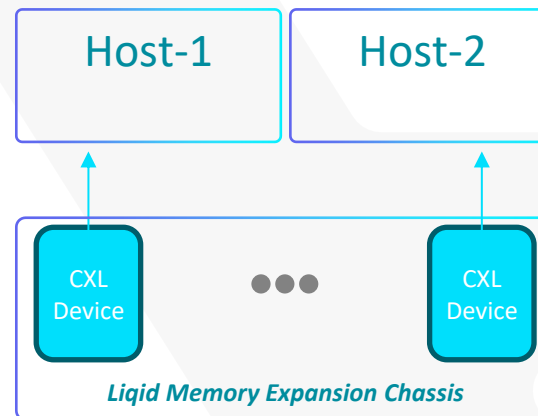
CXL Memory Deployment Options

Memory Expansion



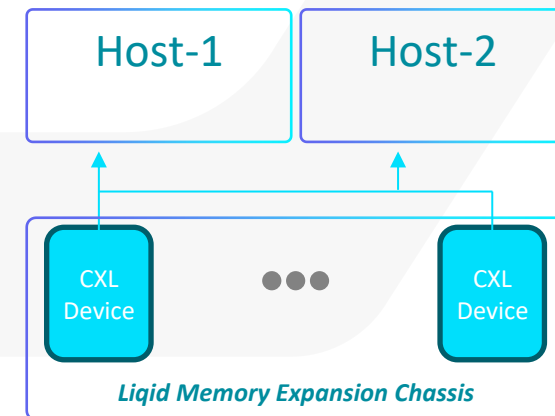
- ✓ Single Host Expansion
- ✓ Internally Deployed
- ✓ Fixed Memory Capacity

Memory Pooling



- ✓ Multi-Host Memory Pooling
- ✓ External Expansion Chassis
- ✓ Dynamic Provisioning Per Host

Memory Sharing



- ✓ Simultaneous Memory Sharing
- ✓ Enable Application Coherency
- ✓ Performance, Features, TCO

Liquid Memory Pooling Portfolio



ToR Switch

- ✓ 15x Ports CXL 2.0 (x16)
- ✓ Support PCIe Gen5
- ✓ 1RU Form Factor



Expansion Chassis

- ✓ 10x AIC Mem Exp Chassis
- ✓ 5x CXL 2.0 (x16) Uplinks
- ✓ 4RU Form Factor



HBA

- ✓ Host Bus Adapter
- ✓ CXL x16 Uplink
- ✓ HHHL Form Factor



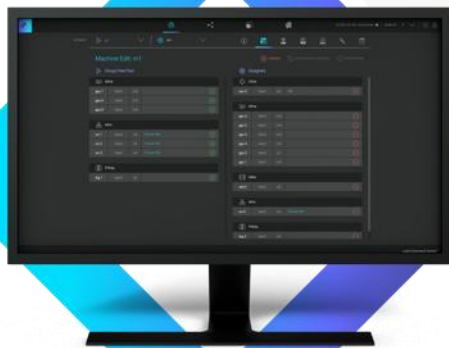
Optical Cables

- ✓ CDFP Connector
- ✓ Single Cable for x16
- ✓ 30m Length Supported



Memory Module

- ✓ 2 TB (256 GB DIMM)
- ✓ CXL 2.0 x16
- ✓ AIC Form Factor

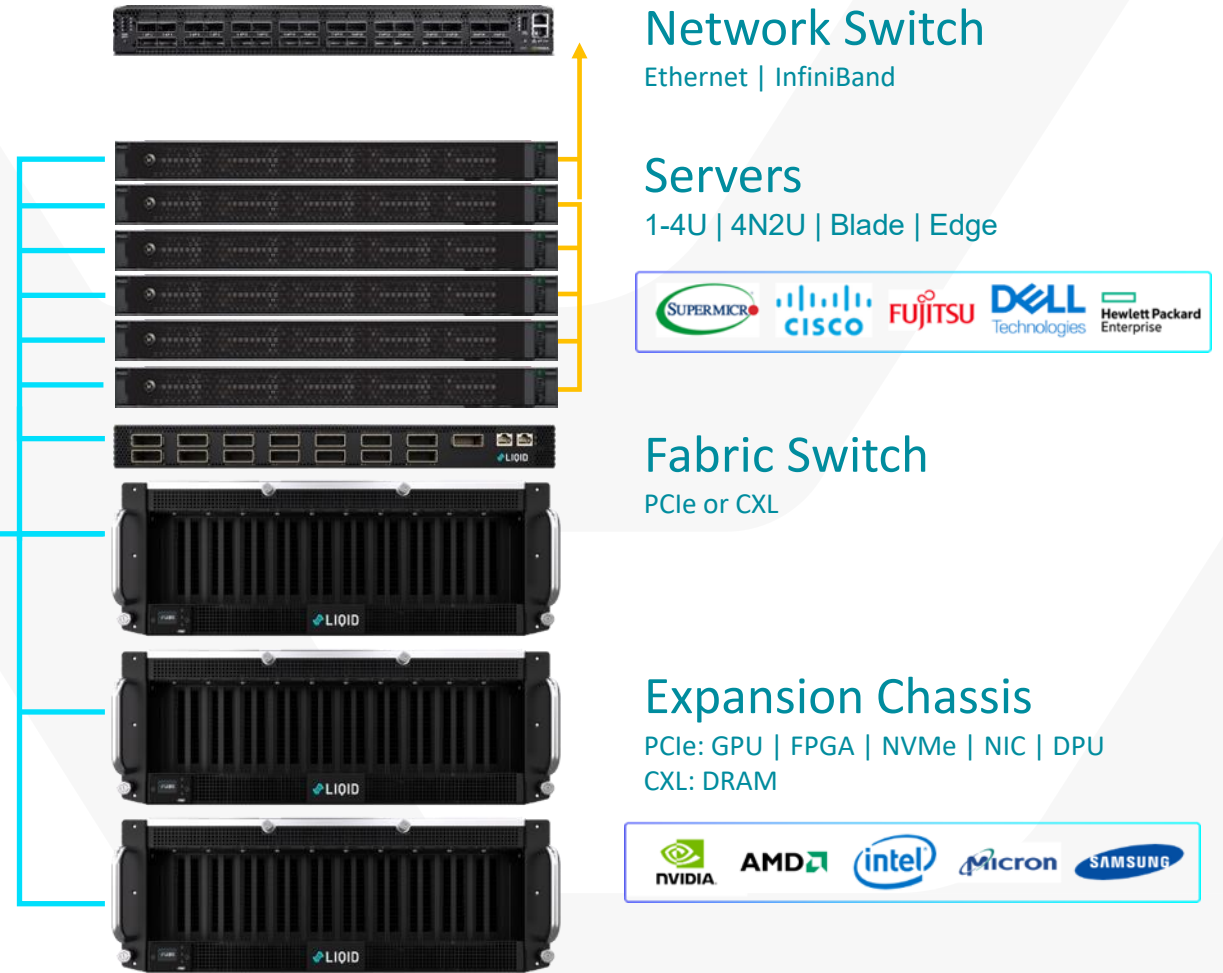
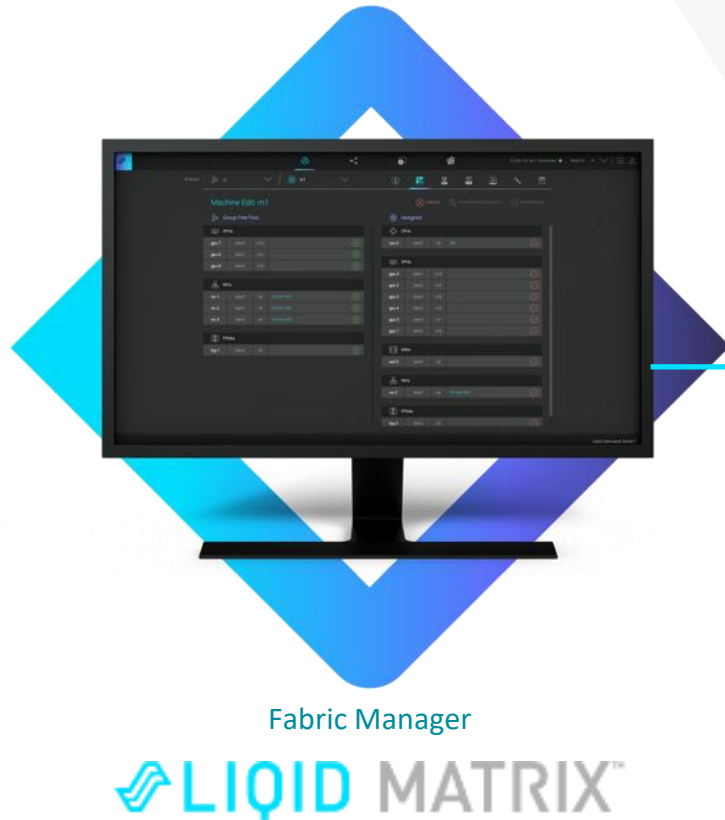


Liquid Matrix Software
CXL Fabric Manager

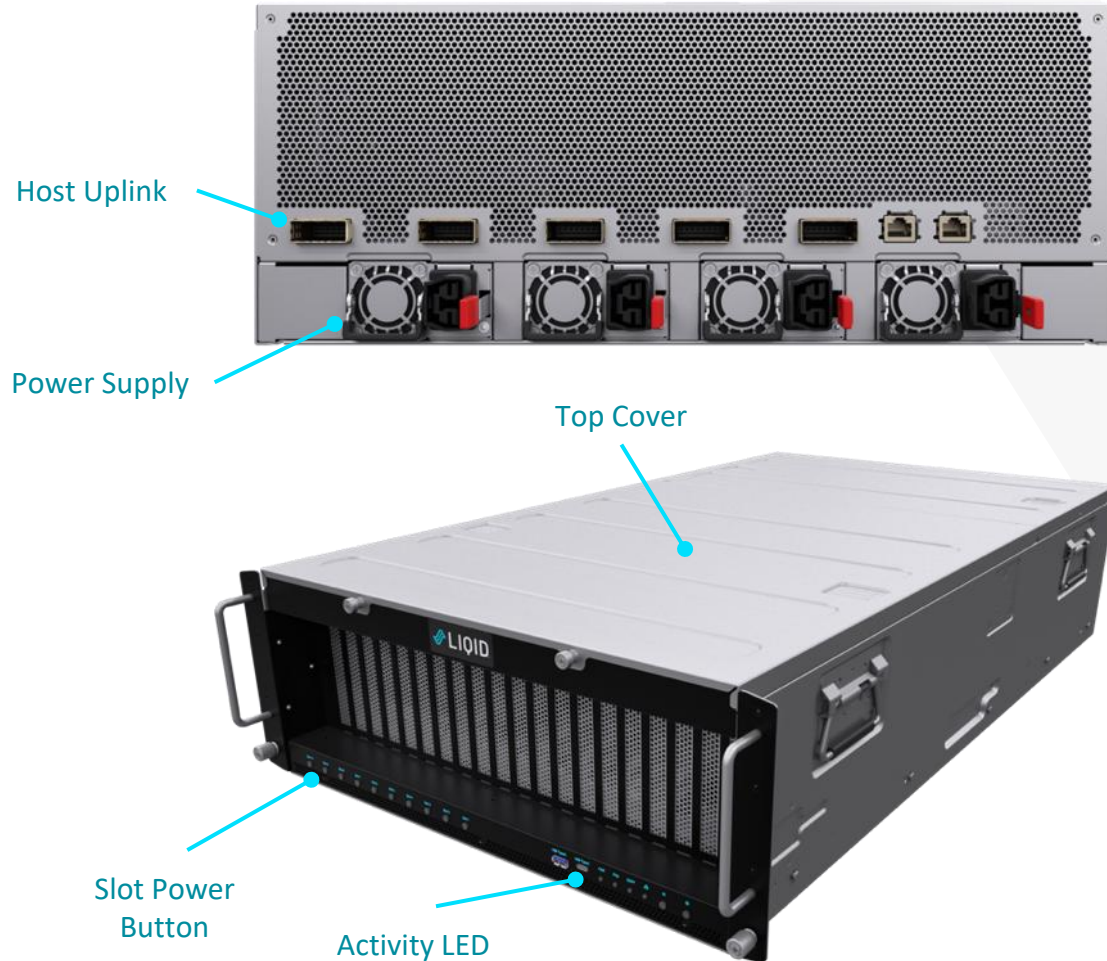
	Production Schedule
Matrix SW	Available
Expansion Chassis	Available
ToR Switch	Available
Host Adapter (HBA)	Available
Optical Cable	Available
Memory Module	Available

Software Defined Memory and GPU Pooling Infrastructure

Orchestration Software



EX-5410C: CXL Memory Expansion Chassis



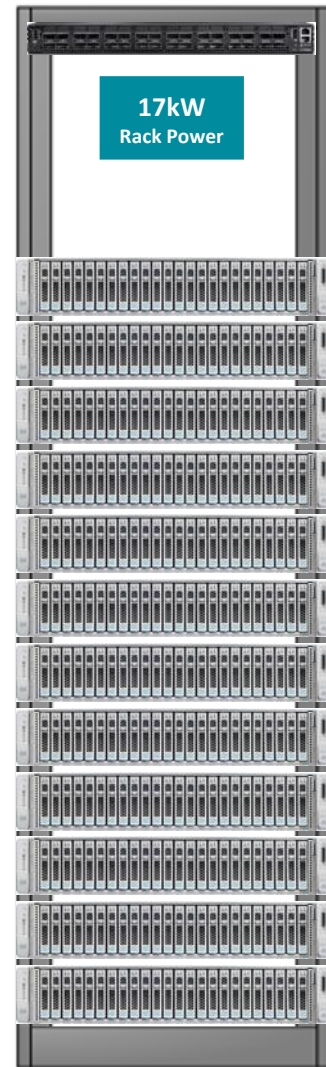
Up to **20TB** of DRAM per Chassis
Scalable to **100TB** per Pod

~ 640 GBs of Bandwidth	~ 500nS Latency
10x CXL Memory Modules Supporting 80x DIMMs	Composable Memory Sharable Across up to 30x Hosts
4RU Form Factor	5x Uplinks Gen5 x16
Liquid Matrix Software - CXL Fabric Manager	

Optimize Memory Power and Placement

Enable DRAM Deployments
in Neighboring Racks

Server
Rack



Optical Interconnect
Up to 30M
Per Hop

17kW
Rack Power



Memory
Rack



Abaco Rack-Scale Design: Built for On-Prem AI

The infrastructure foundation to deploy open-source models at any scale — with the automation, economics, and data security.

FLEXIBLE GPU + MEMORY POOLS

Any Model, Any Size

Right-sized GPU and memory pools deploy everything from 250M to 1T+ parameter models

ON-PREM INFRASTRUCTURE

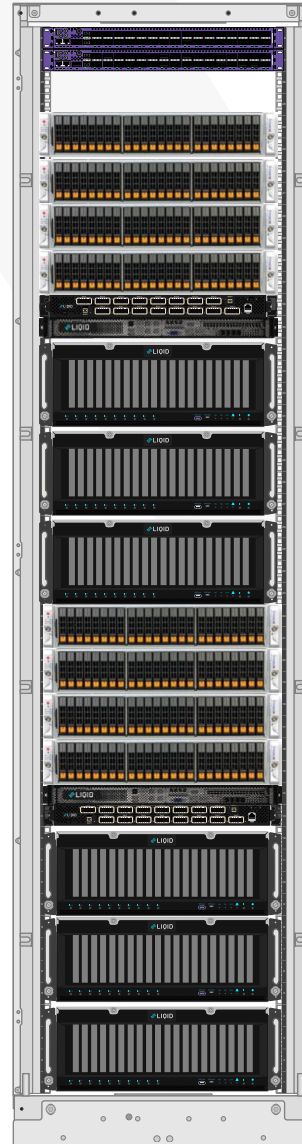
Best-in-Class Tokenomics

Fixed infrastructure cost and shared pooled resources cut token cost by more than 90%

KUBERNETES-NATIVE

Flexible Automation

Deploy, scale, and orchestrate models like any cloud-native workload — automated model deployments



Server Pool

Up to 16x Servers Per Pod

Supports Intel or AMD

GPU Pool

Up to 30x GPU Per Server

Supports Nvidia and AMD

Memory Pool

Up to 160 TB Per Server

Supports Micron and others

Rack-Scale Design

Validated HW and SW

Abaco Ecosystem

Strategic Partnerships



Low Latency Full Stack Solution

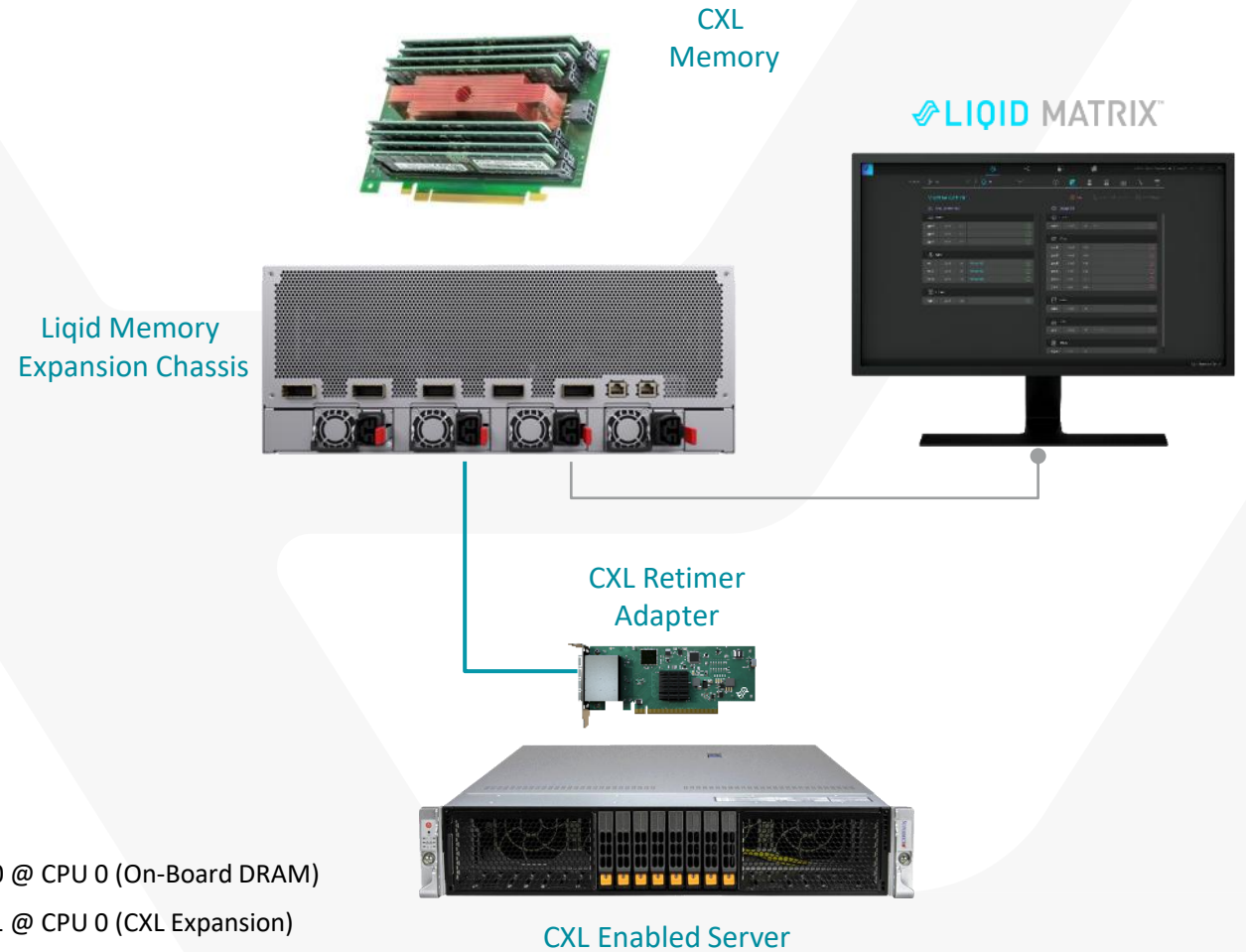
```

root@ftiturin:/home/liquid# numactl -H
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12
          26 27 28 29 30 31 32 33 34 35 36 37 38 39
          52 53 54 55 56 57 58 59 60 61 62 63 64 65 66
          79 80 81 82 83 84 85 86 87 88 89 90 91 92
          104 105 106 107 108 109 110 111 112 113 114
          124 125 126 127
node 0 size: 31413 MB
node 0 free: 28674 MB
node 1 cpus:
node 1 size: 10321863 MB
node 1 free: 10320788 MB
node distances:
node  0  1
  0:  10  18
  1: 255  10
root@ftiturin:/home/liquid#

```

Latency (ns)	NUMA-0	NUMA-1
CPU 0	147	623
CPU 1	235	722

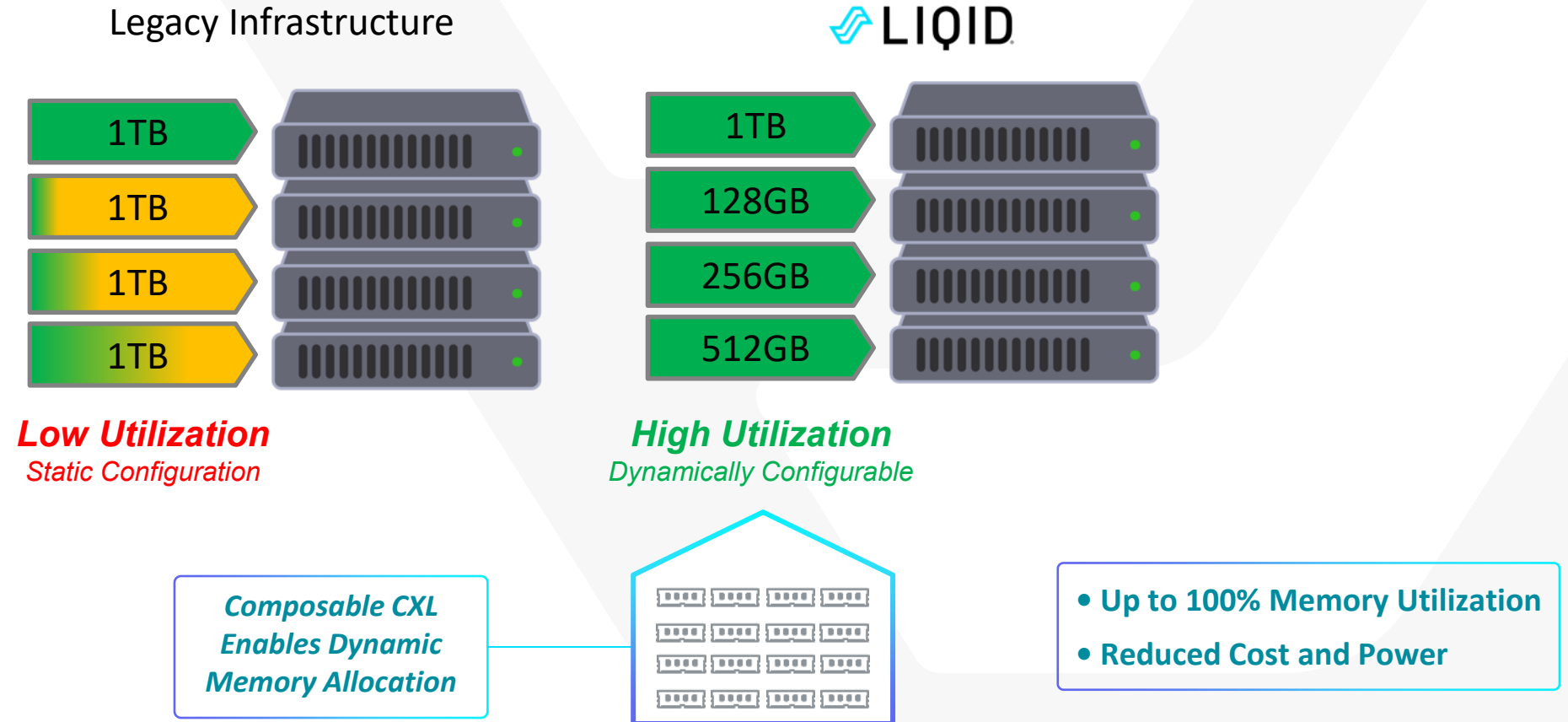
* Round trip latency measurement (ns)



NUMA-0 @ CPU 0 (On-Board DRAM)

NUMA-1 @ CPU 0 (CXL Expansion)

Optimize Memory Utilization with Pooling



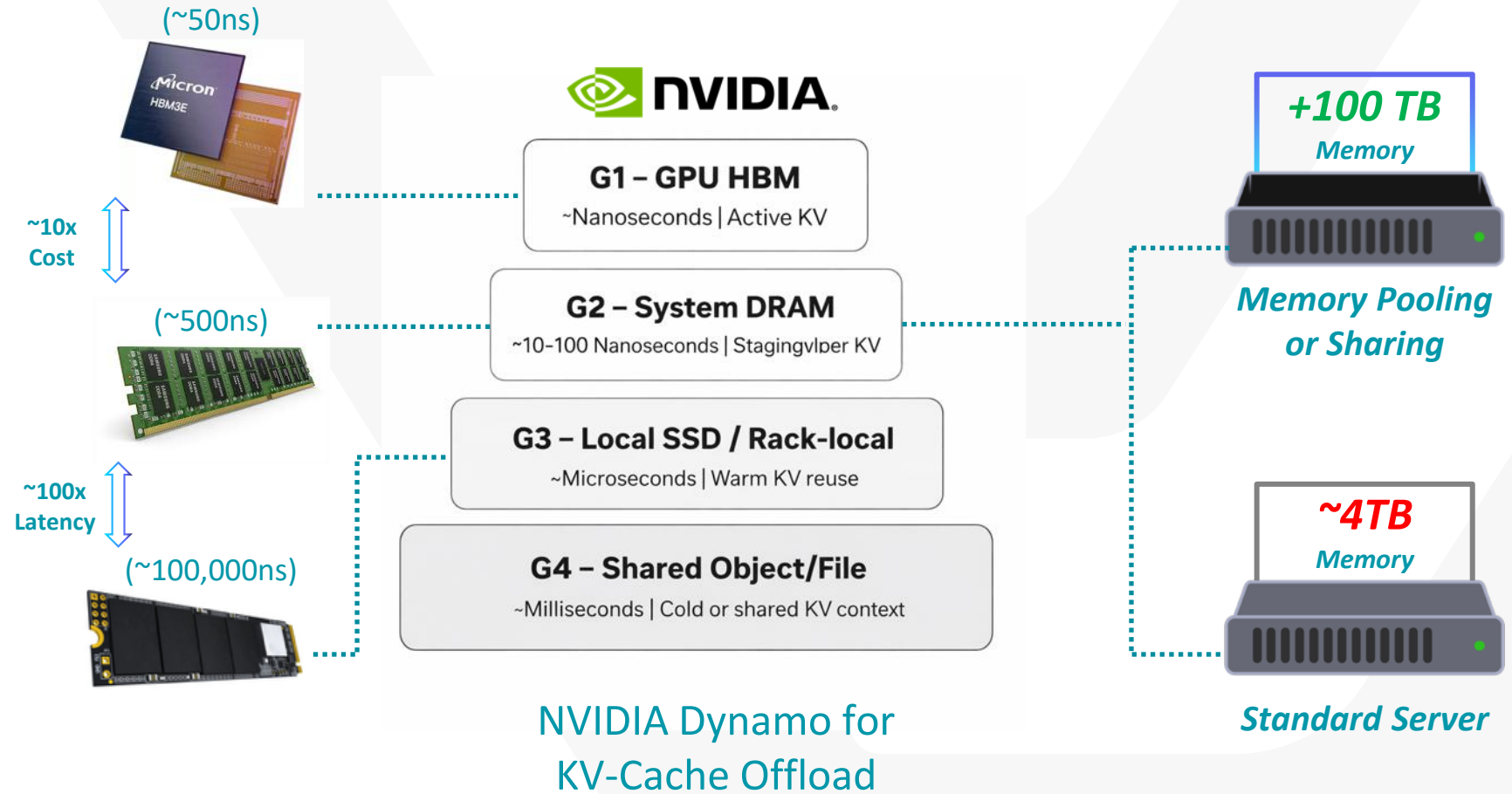
DRAM is Optimal Solution for Inference GPU Offload

DRAM pooling delivers optimal cost-performance with high capacity at lower cost

Cost per GiB: \$\$\$\$\$\$
Performance: High
Capacity: Low (GiB)

Cost per GiB: \$\$
Performance: High
Capacity: High (TiB)

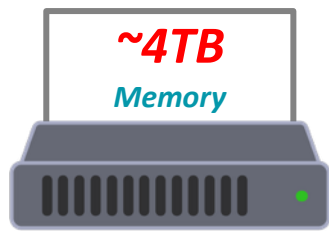
Cost per GiB: \$
Performance: Low
Capacity: High (TiB)



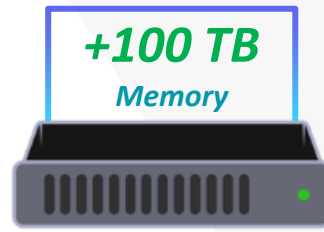
Larger Memory for Higher KV Cache Hit Rate

Memory pooling enables larger server memory for improved KV cache performance

Workload vs. KV Cache Size Requirement



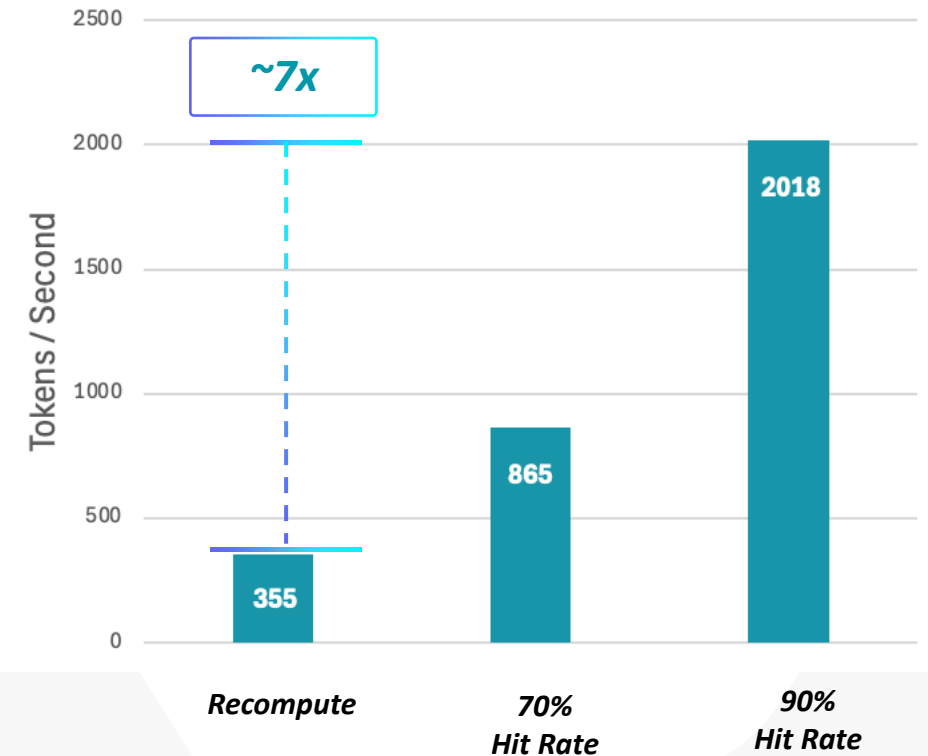
Standard Server



Memory Pooling

Scenario	Context	Blocks	Model: Llama 400B	
			KV / User	KV / 250 users
RAG / Chat	32K	2,048	15.8 GB	3.9 TB
Document Analysis	64K	4,096	31.5 GB	7.9 TB
Agentic	128K	8,192	63 GB	15.8 TB

Tokens/Second vs. KV Cache Hit Rate

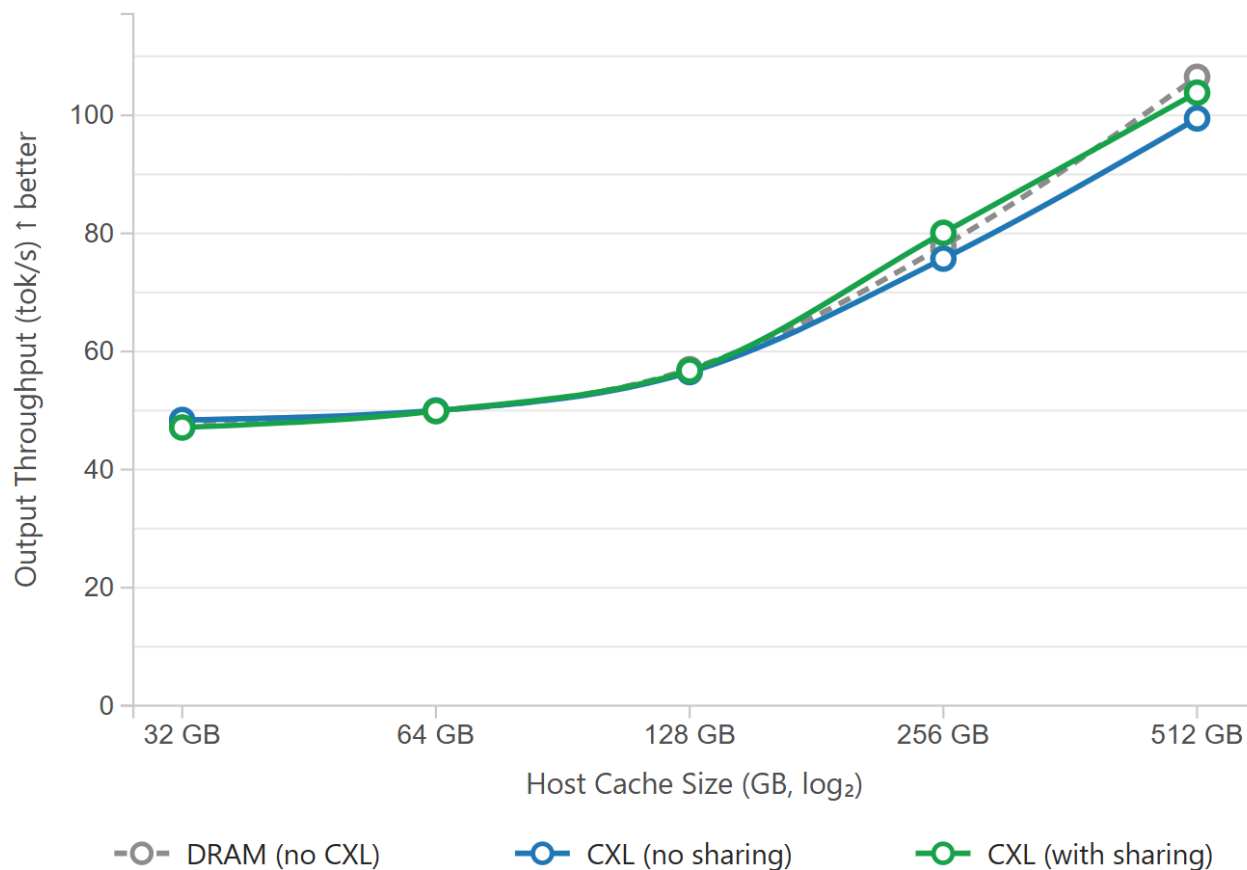


Memory Pooling Enables Higher Hit Rate for KV Cache

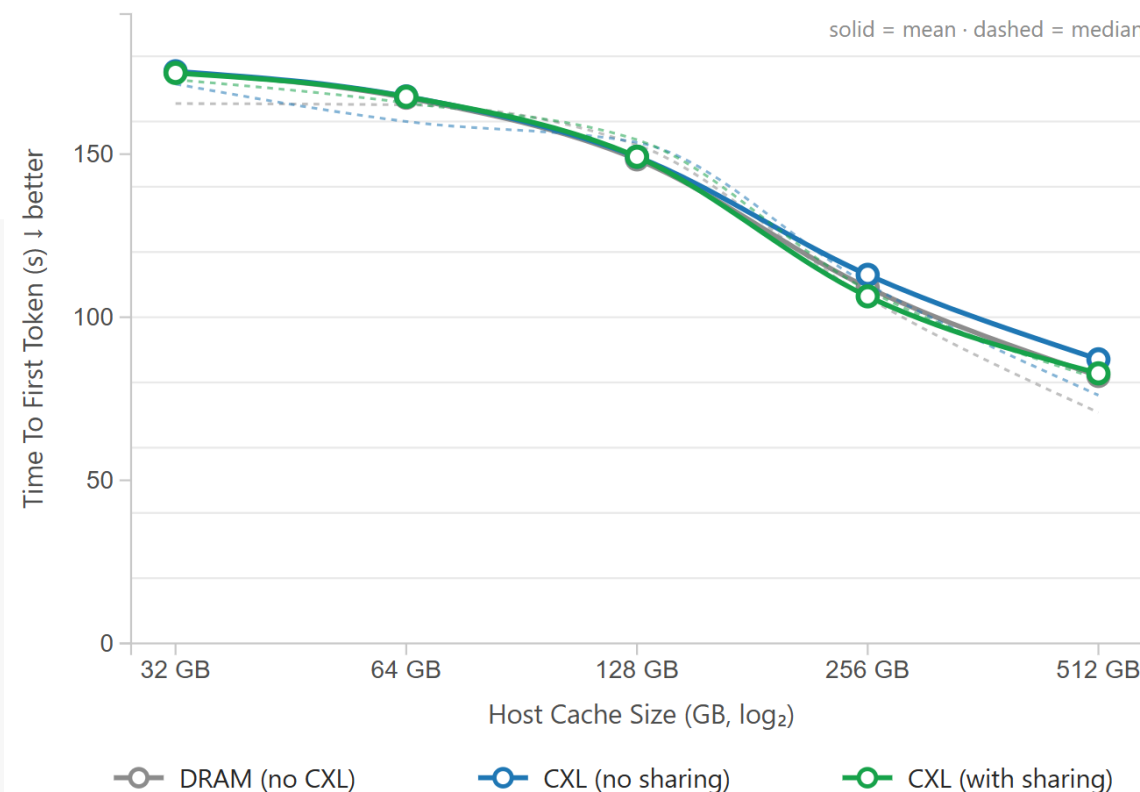
Memory Pooling Delivers DRAM-Class Performance

Dynamo with Memory Pooling Delivers 99% of Native DRAM Performance for Latency and Throughput

Output Throughput vs Host Cache Size



Time To First Token vs Host Cache Size

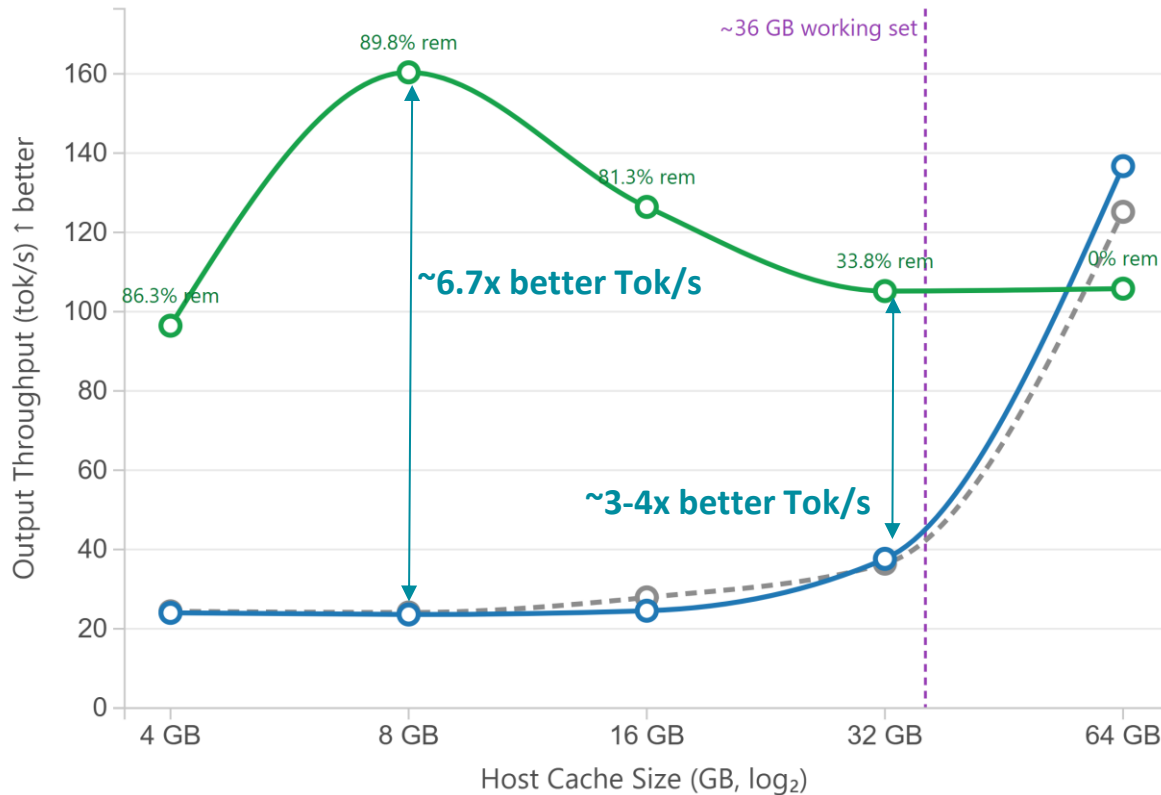


Multi-Server Memory Sharing for KV Cache

Shared, reused prefixes increase effective cache capacity and improve GPU utilization - Time to first token (TTFT) approaches full cache performance early

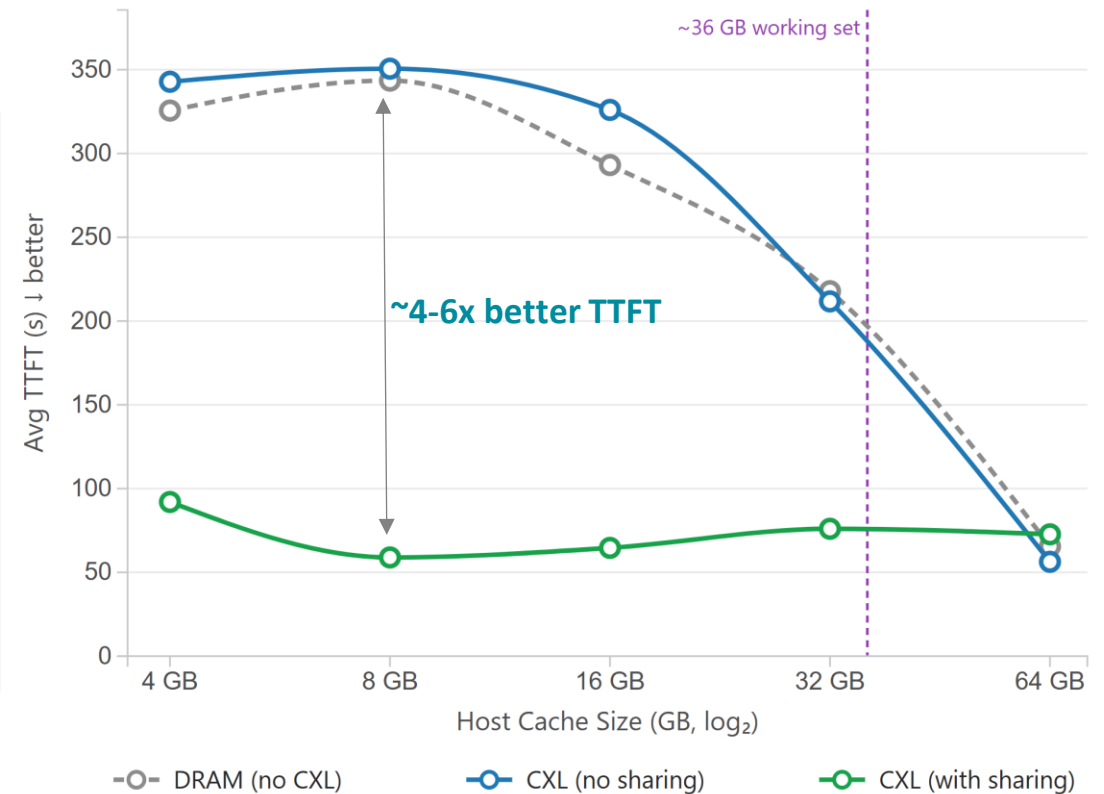
Output Throughput vs Host Cache Size

Higher is better. Sharing (green) peaks at 8 GB ($6.7\times$ DRAM), stays $3-4\times$ through 32 GB, converging above the ~ 36 GB threshold. Open markers = suspect/missing.



Average TTFT vs Host Cache Size

Lower is better. Sharing's cross-worker prefill reuse cuts TTFT $\sim 4-6\times$ at small cache (59 s vs 343 s DRAM at 8 GB). Absolute values are large (queue-bound at conc-64); read config-vs-config gaps.



AI Inference Acceleration with Memory Pooling

Industry results for memory pooling and scaling showing major AI performance gains by eliminating bottlenecks

	Metric	Performance Delta	Results
	TTFT	~10x Reduction	~7x More Tokens
	Throughput	~7x More Tokens	
	Utilization	~3x Higher GPU Utilization	

	Metric	Performance Delta	Results
	TTFT (Avg)	~10x Lower Avg TTFT	~10x Faster TTFT
	TTFT (P99)	~6x Lower P99 TTFT	
	Throughput	~1.6x Higher Throughput	

	Metric	Performance Delta	Results
	Throughput	~22x Higher	~22x Higher Throughput
	Efficiency	~60x Lower Energy	
	Cost Efficiency	~7x Better	

Key Workloads and Benefits

AI Inference

Highest Token/\$ and Token/W

~10x Higher Performance



In Memory Database

Reduced Processing Time

~30x Faster Queries

Graph Database
Processing Time

8Hrs → 15min



Virtualization

Increased VMs Per Server

~4x Higher VM Per Server

Legacy



2TB DRAM

LIQID



10TB DRAM / CXL

2TB DRAM + 576 Cores

@ 8 Cores Per 128GB = 16 VMs

(Insufficient Memory)

\$5,031 Per VM

10TB DRAM + 576 Cores

@ 8 Cores Per 128GB = 72 VMs

(Balanced Configuration)

\$2,229 Per VM

50% Reduced Cost Per VM

CXL Cuts NVMe Writes and Extends Drive Life

Multiple mechanisms work together to cut the total volume of writes that ever reach flash and reduce costs



Larger Cache, Less Write-Through

A DRAM-class capacity tier in front of NVMe means far more data stays resident before eviction, so fewer writes ever pass through to disk.



Write Filtering, Fewer Unnecessary Writes

Class-aware placement keeps ephemeral, low-reuse blocks in cache instead of writing them out to NVMe at all.

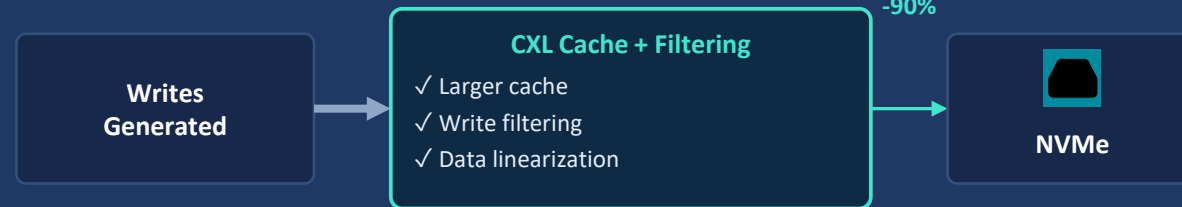
COMBINED EFFECT

Up to 90%

fewer writes reach NVMe

Bigger cache, smarter filtering, and data linearization

HOW IT WORKS



Lower Endurance Requirement

5 DWPD → 1 DWPD

Absorbing and filtering writes before they reach NVMe drops the sustained write pressure the media must tolerate — from 5 DWPD down to 1 DWPD.



Lower Media Cost

~65% lower media cost

1 DWPD NAND is dramatically cheaper than 5 DWPD NAND. The same workload can now run on lower-endurance, lower-cost media — cutting drive cost by up to ~65%.

Strategic Memory Vendor Partnerships



CXL Memory
CXL Pooling & Sharing

✓ *CXL Supported*



RDIMM



CXL Memory
CXL Pooling & Sharing

✓ *CXL Supported*



RDIMM, E3.S, AIC



CXL Memory
CXL Pooling & Sharing

✓ *CXL Supported*



RDIMM



CXL Memory
CXL Pooling & Sharing

✓ *CXL Supported*



RDIMM and E3.S

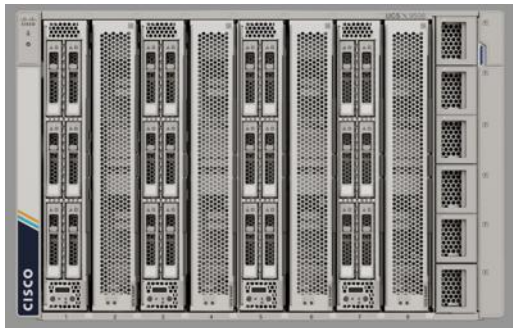
Strategic Server OEM Partnerships



UCS Servers

GPU and CXL Pooling

✓ *GPU Supported*
CXL Under Qualification



PowerEdge Servers

GPU and CXL Pooling

✓ *GPU Supported*
CXL Under Qualification



ProLiant Servers

GPU and CXL Pooling

GPU Supported
CXL Under Qualification



X14 Servers

GPU and CXL Pooling

✓ *GPU Supported*
✓ *CXL Supported*



Continue the Series: Endless Memory with CXL



IT Brand Pulse AI Leadership Forum. Each episode builds on today's foundations.

EPISODE 2

Wed, Oct 28

The Emerging Memory-Centric Data Center

Architectures beyond single-server memory for AI inference, KV cache, agentic AI, databases, analytics and HPC

EPISODE 3

Wed, Dec 2

Meta and Penguin Solutions on Reusing DDR4

The economics of redeploying installed-base DDR4 as CXL-attached capacity

EPISODE 4

Wed, Jan 13

PNNL Abaco: CXL Memory Pooling in the Real World

A complete production system with PNNL, Micron, Liquid and Scalemem

Want to see CXL memory pooling running? liquid.com | info@liquid.com

Thank You



Contact

11400 Westmoor Circle, Suite 225
Westminster, CO 80021

Phone: 303.500.1551

Website: liquid.com

General Info: info@liquid.com

Partner Info: partner@liquid.com