Boost your AI performance with CXL

September 16, 2026

Meet Our Speakers

Moderator

Siamak Tavallaei
Samsung

Panelist

Sandeep Dattaprasad
Astera Labs

Panelist

Hasan Maruf
Meta

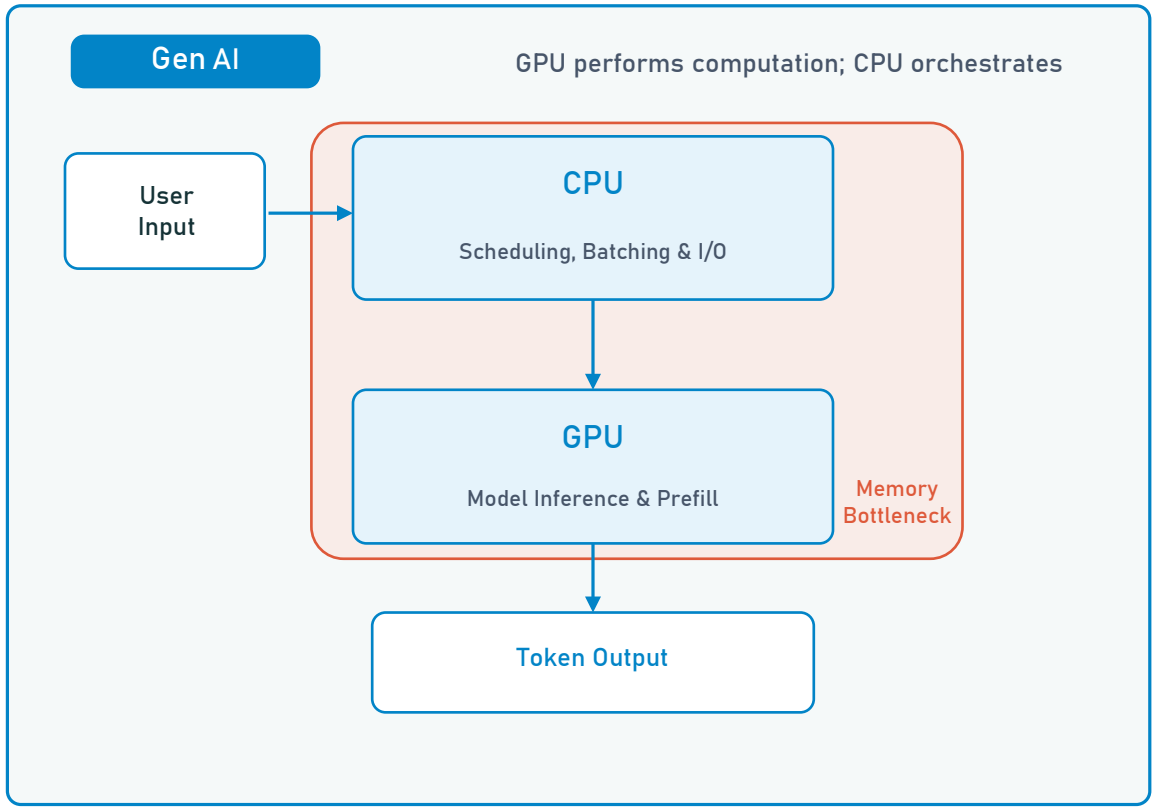
Panelist

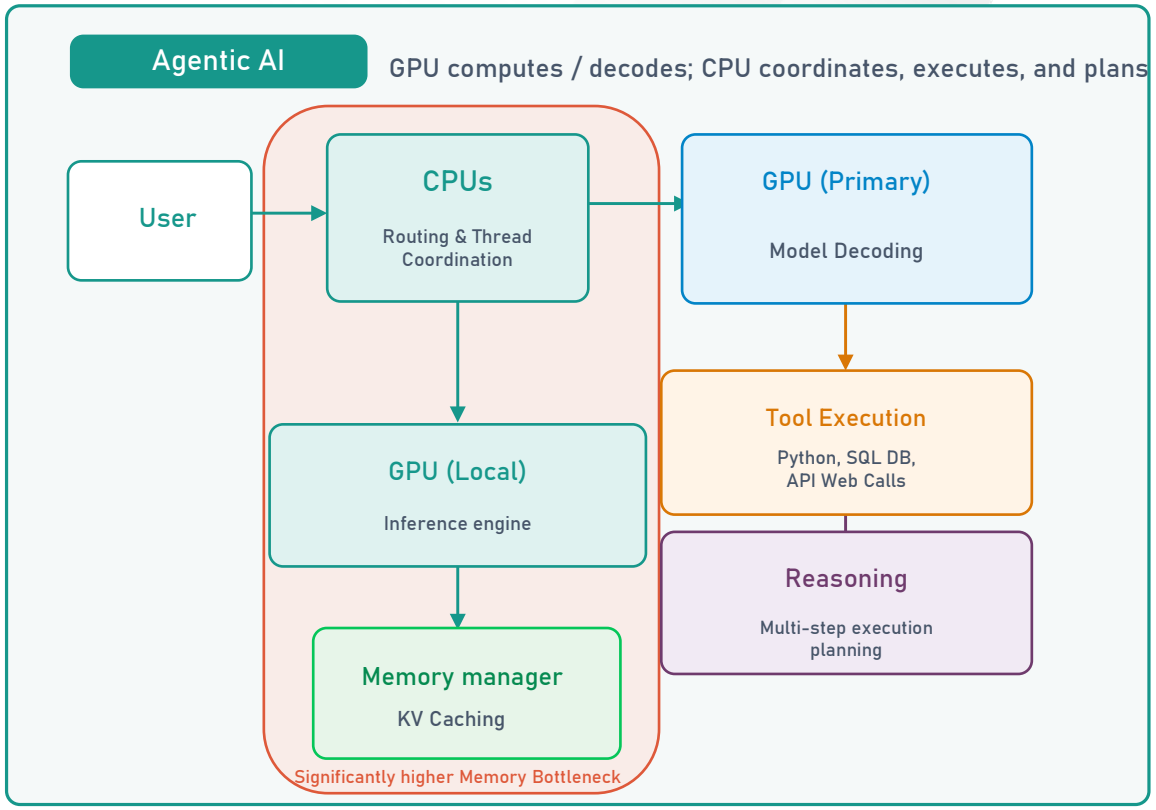
Joseph Baco
Primemas

Boost your AI performance with CXL

Sandeep Dattaprasad, Astera Labs

AI Inference Trends & Bottlenecks





POWER
5–7 year lead times; efficiency = capacity

MEMORY
HBM sold out; DDR5 $/GB costs elevated

CONNECTIVITY
Rack is the new server; pod is the new rack

UTILIZATION
Idle capacity is the invisible wall

Strategy A: Scale Capacity via Disaggregated KV-Cache using CXL Memory

Core Tokenomics Approach:

Scales long-context token generation while minimizing host and GPU infrastructure

Address the GPU Memory Wall

Offload KV cache from expensive HBM to cheaper CXL DRAM

→ CXL attached memory for capacity increase

→ Enable higher batch sizes and lower GPU node counts

→ Drastically drived down cost per million tokens ($/M tokens)

Token Economic Drivers:

Increase GPU Token Throughput:

- Boosts memory capacity per node
- Increases output tokens per GPU-hour

Reduce CapEx per Token:

Fewer compute resources (CPUs, motherboards, chassis, switch ports) → reduces the fixed infrastructure

Key Takeaway: Generates equivalent token volume across long-context AI workloads with fewer server/GPU hosts | Drastically drived down cost per million tokens ($/M tokens)

Strategy B: Rack-Level Memory BOM Optimization

Core Tokenomics Approach:

Optimize Server Hardware BOM with CXL-Tiered Memory

Replace high-cost, high-density DRAM (128GB DIMMs) with lower-cost DRAM + CXL memory

→ Maintain the required target RAM per GPU host at a lower BOM

- Lower-sized DIMMs on compute
- \+
- CXL memory with lower-sized DIMMs

→ Immediate CapEx savings per host node

- Lower $/GB

Token Economic Drivers:

Lower Amortized Hardware Cost per Token:

Eliminates the steep price premium of high-density DRAM modules

Preserves TTFT & TPOT SLAs:

Provides the memory BW and capacity needed by reducing TTFT without forcing node redesigns

Immediate Rack-Scale ROI:

Delivers direct unit-level CapEx savings per rack without changing inference framework

Key Takeaway: Immediate CapEx savings per host node | Reduced the baseline infrastructure cost component of token pricing

Flexible KV Cache Solutions for LLMs and Agentic AI with Leo

LLM Application Examples

Chatbots

reuse early chat content as context for later chat input

Bug Fixing

Frequently using entire code repository as context

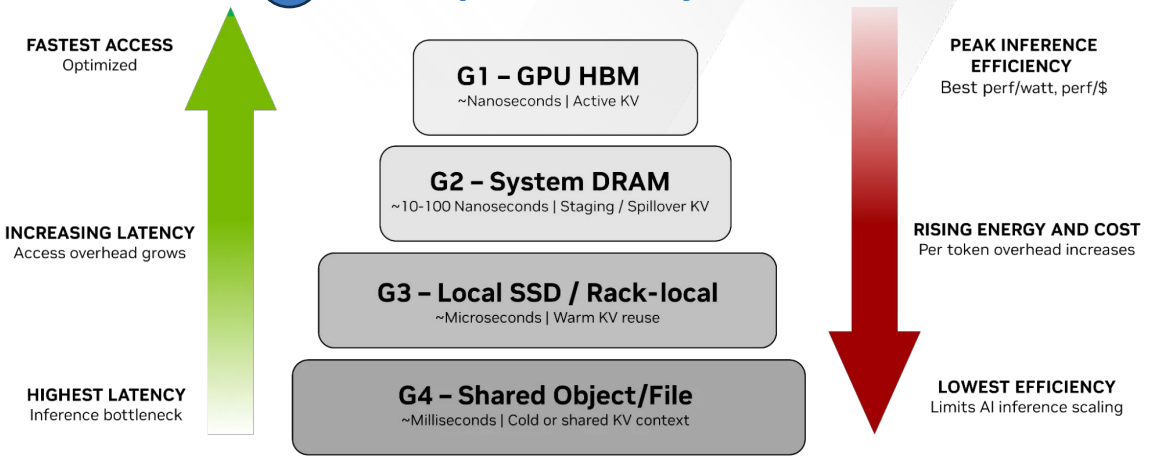
Agents

Assistant and rec sys query large document sets

1 Increasing KV Cache Capacity

- Larger models
- Longer context
- Multi-turn conversations
- Reuse across agents
- Concurrent users/sessions
- Long term memory

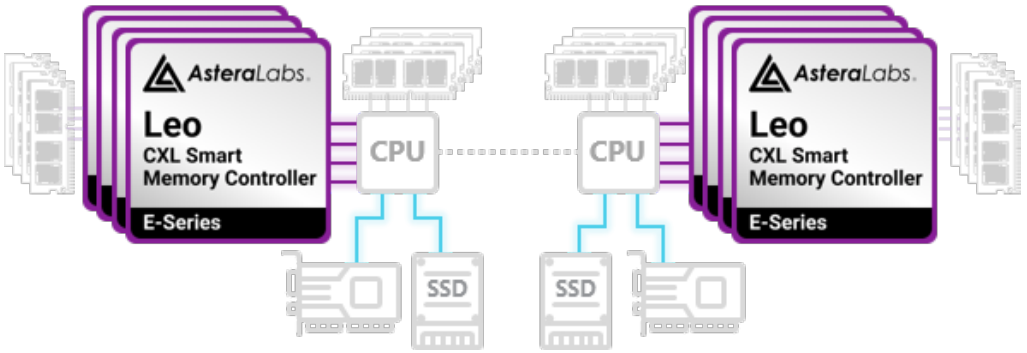
2 Multiple Memory Tiers



3 Purpose-built for AI Inference

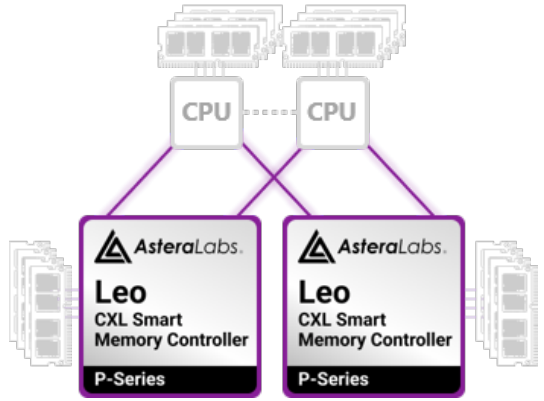
Agentic AI, KVCache, General Purpose Servers

Reused DDR4 or DDR5 DIMMs



4 High Memory Utilization with Sharing & Pooling

Multi-Socket Servers



Ecosystem Momentum

Astera Labs Bolsters Leo Smart Memory Controller Family for Agentic AI and General-Purpose Cloud Workloads

See the Demo in Booth 904

Astera Labs' Leo CXL Smart Memory Controllers on Microsoft Azure M-series Virtual Machines and Intel Xeon Overcome the Memory Wall

Scaling Agentic AI With CXL Memory Expansion and Extended PCIe 6 Signal Reach

AI Your Way at AMD Advancing AI 2026: Why Open Standards Power AI Everywhere

Scaling CXL® to Millions of Servers Vistara for Hyperscale Efficiency

Hasan Al Maruf, Meta, Inc.

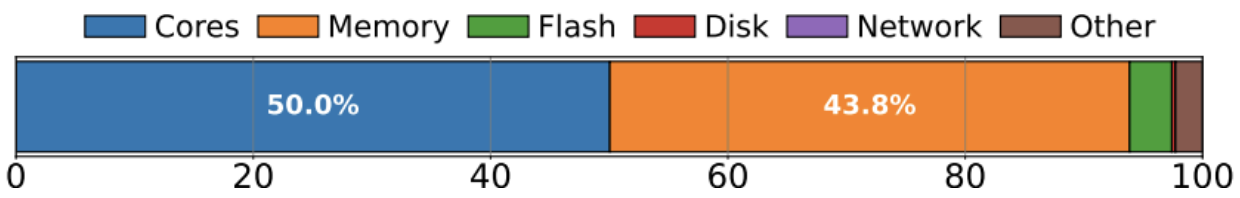
Memory is the Hyperscale Bottleneck

43%
of servers are memory-bound
not compute-bound

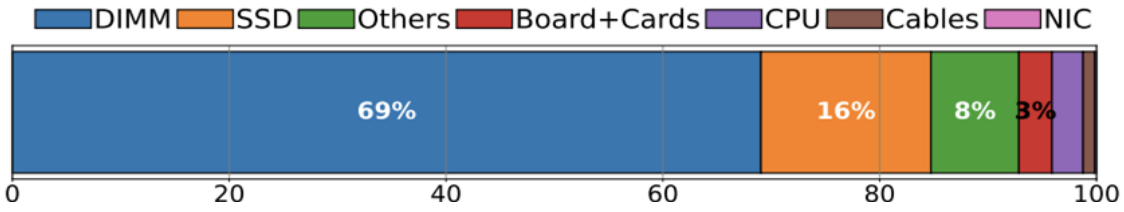
69%
of a server's embodied
carbon footprint is DRAM

10–14 yr
DRAM lifetime
servers retire at 5–7 yr

Fraction of Servers bottlenecked by resource (%)



Emissions by Component (%)



CXL promised to address all three at once

What Fleet-Wide CXL Demands?

Cost-effective Expansion

Lower Carbon

Easy Adoption

Generic & Broad

Operational Efficiency & Low Maintenance Overheads

Vistara: one co-designed stack that meets all the demands

Vistara: End-to-End CXL Stack

Software Linux Memory Tiering

The complexity of CXL is invisible to applications. Keeping tooling and maintenance simple

Platform MemServer

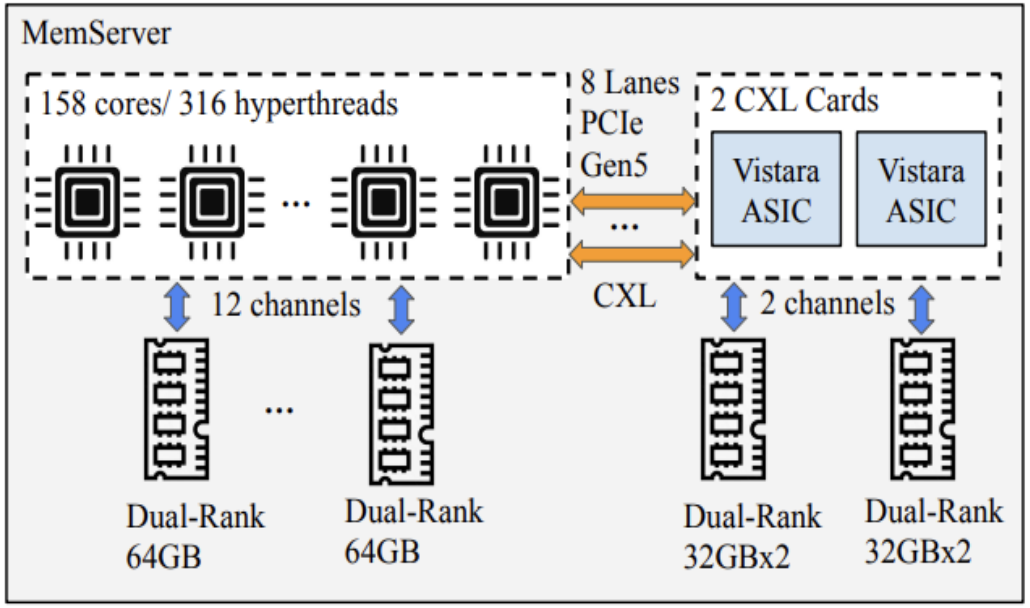
High memory-to-compute ratio for memory-bound services

ASIC Vistara

Power & cost efficient, performant expansion with DDR4 reuse

One stack, co-designed end-to-end, running in production today

MemServer: 1TB Server



Vistara ASIC (2x ASICs)

Host Connectivity x16 PCIe Gen5 (x8 used in prod)

Memory Interface 2x DDR4 2DPC channels

Configuration 4x 32GB DDR4 DIMMs 2400MT/s

Production Operation Conditions	Local memory [60% BW util]	CXL Memory [20% BW util]
	234 ns	280 ns (+50ns)

1 TB memory / server

3 : 1 local : CXL

250 ns CXL Idle Latency

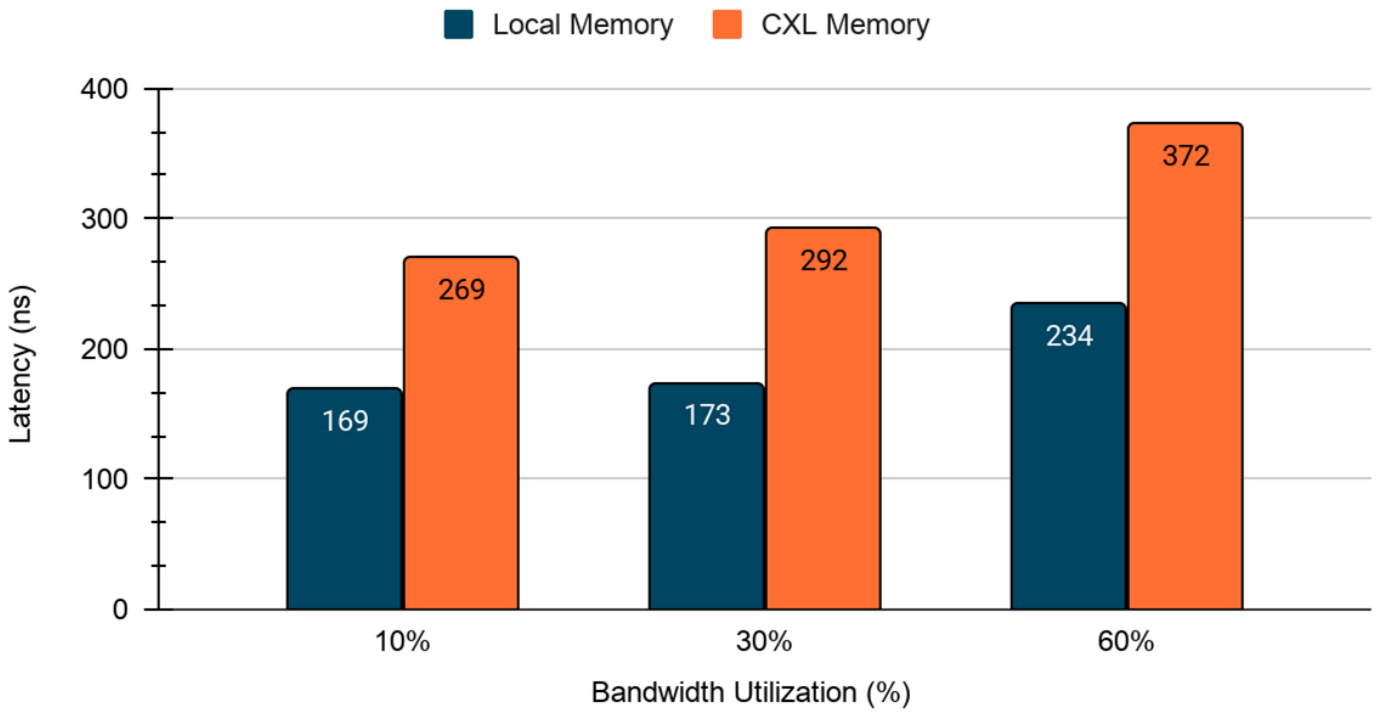
48 GB/s Peak Effective CXL BW

+50 W for both ASICs+DIMMs

MemServer platform designed for memory constrained services with 3:1 local:CXL ratio

Local vs CXL Memory in Production

	Local	CXL
Capacity	768 GB	256 GB
BW	614 GB/s	58 GB/s
Power	132 W	30 W
Power/GB	1x	0.7x
Cost/GB	1x	0.13x



Practical trade-offs of deploying CXL memory shows its potential for cost-effective capacity expansion

Already in Production

ML Parameter Server

Fewer servers, less fan-out, fresher models

−25%
fleet size

Cache

Bigger cache partitions cut backend traffic

−29%
avg. query processing time

+25%
throughput

−25%
fleet size

Development Infrastructure

More CI/CD jobs & VMs per server

−33%
fleet size

Data Warehouse

+33% executor packing, fewer OOMs

−10%
execution time

−30%
fleet size

Server count reduction & DRAM reuse cut cost and carbon footprint

Vistara: Making CXL Real!

First at scale

The first reported CXL memory expansion deployed at scale.

Co-design wins

Custom ASIC + transparent OS tiering, built production-grade solution.

Broad impact

Delivers performance gains across a wide range of real-world workloads.

CXL Memory expansion with DRAM Reuse pays

High memory-to-compute ratio achieved with DDR4 reuse dramatically cuts cost and carbon footprint landing huge TCO wins!

Memory Disaggregation: Practical Solutions

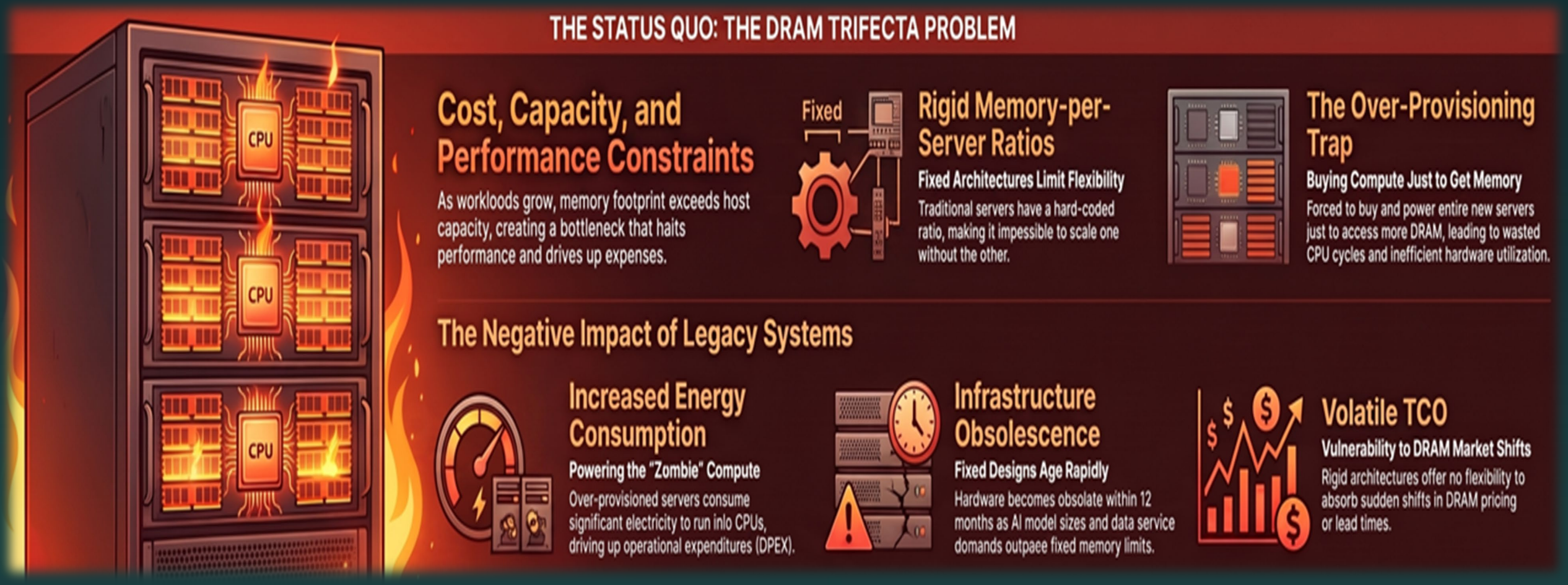
Joseph Baco, Primemas Inc.

DRAM Memory Requirements

LLMs & Agentic AI

The memory wall ***was*** real and growing since ~2020

Life Inside the Wall



Modern Datacenter demands have outgrown traditional host-attached memory

CXL: Emerging Solutions

Real solutions. Real products. Emerging now.

CXL promotors include Alibaba. Cisco. Dell EMC. Google. HPE. Huawei. Intel. Meta. Microsoft.
Source: HPCwire

CXL 4.0 & Memory Pooling win Best of Show — FMS 2026
Source: Forbes

"Unprecedented momentum … across the industry." — Jim Pappas, CXL Consortium
Source: Forbes

1.11x faster LLM training
2.75x cheaper interconnect
Source: ByteDance (CXL-CCL)

25% of Azure DRAM sits stranded at peak allocation
Source: Microsoft Research

89.6% lower time-to-first-token
7.35x higher LLM throughput
Source: Alibaba Cloud (Beluga)

75% higher GPU utilization, >2x inference throughput
Source: Astera Labs

8.3x cheaper host interconnect hardware
Source: Alibaba Cloud (Beluga)

CXL Memory Solutions: $1.39B → $5.42B by 2031 — 31% CAGR
Source: Mordor Intelligence

Samsung. SK hynix. Micron. All racing to CXL 3.2 in 2026.
Source: TradingKey

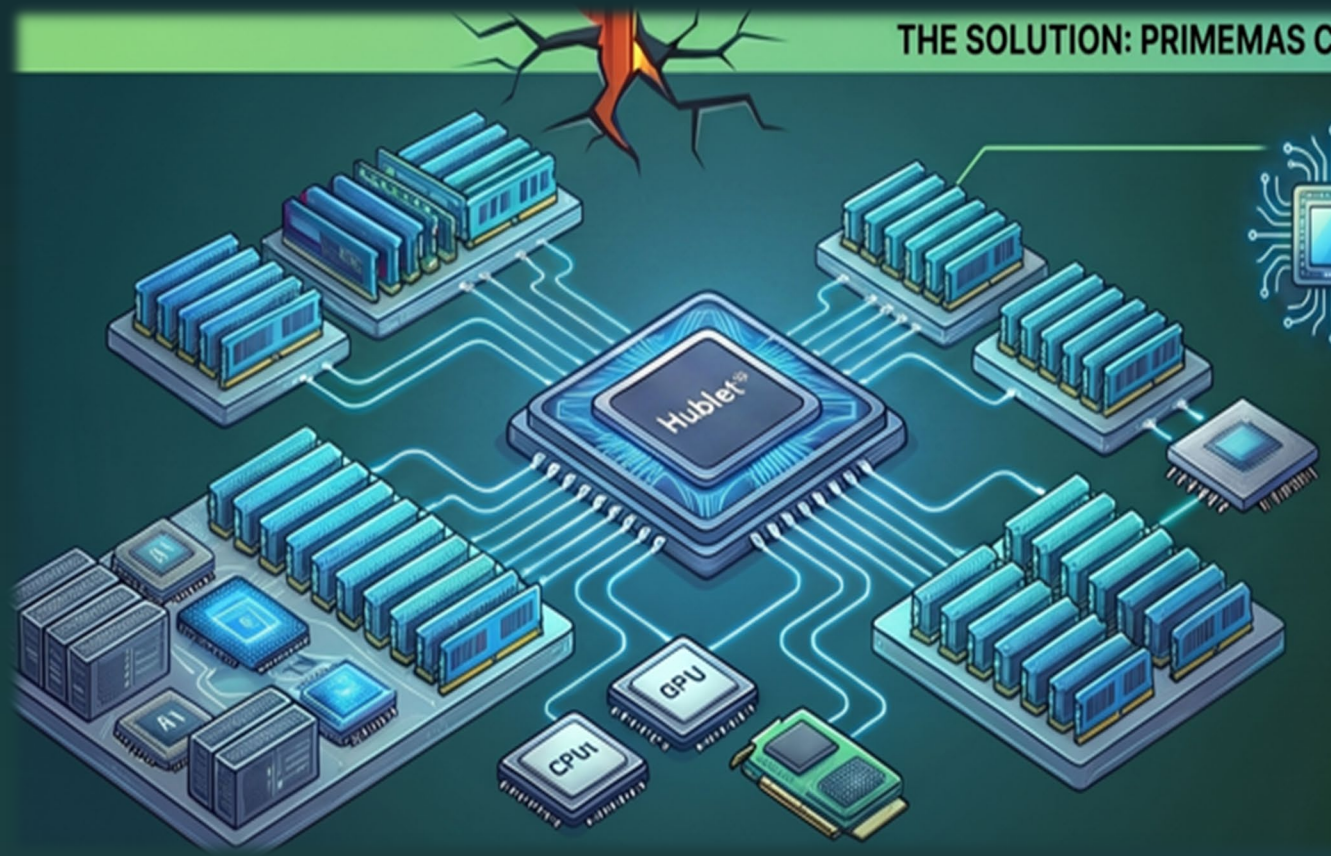
SK hynix ships 96GB CXL 2.0 modules — 2022
Source: ServeTheHome

Samsung Pangea v2: 10.2x traditional RDMA transfer
Source: TradingKey

• Ecosystem • Hyperscaler evidence • Economics • Memory vendors

CXL Real-World Deployment
Life Outside the Wall

THE SOLUTION: PRIMEMAS CXL DISAGGREGATION



THE HUBLET® ADVANTAGE

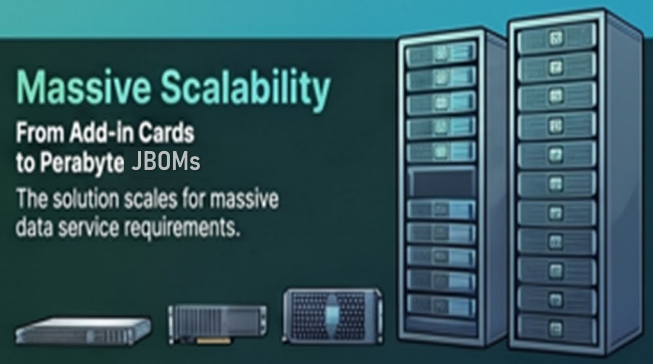
Switchless Low-Latency Architecture

Supports 14, 16, or 32+ DIMMs without the added latency, cost, and power overhead of a traditional switching layer.

Massive Scalability

From Add-in Cards to Perabyte JBOMs

The solution scales for massive data service requirements.



THE BENEFITS: A HIGH-DENSITY FUTURE

Massive Performance Gains

>10x proven performance gains on common DDR/HBM bound workloads.
> 6x leap in tokens per second

Maximized De-coupling Memory from Compute

CXL allows memory to be treated as an independent layer.

Meaningful TCO Reduction

Direct Impact on the Bottom Line

By eliminating over-provisioning and reducing energy waste, disaggregated memory delivers immediate improvements in performance-per-dollar.

Future-Proofing

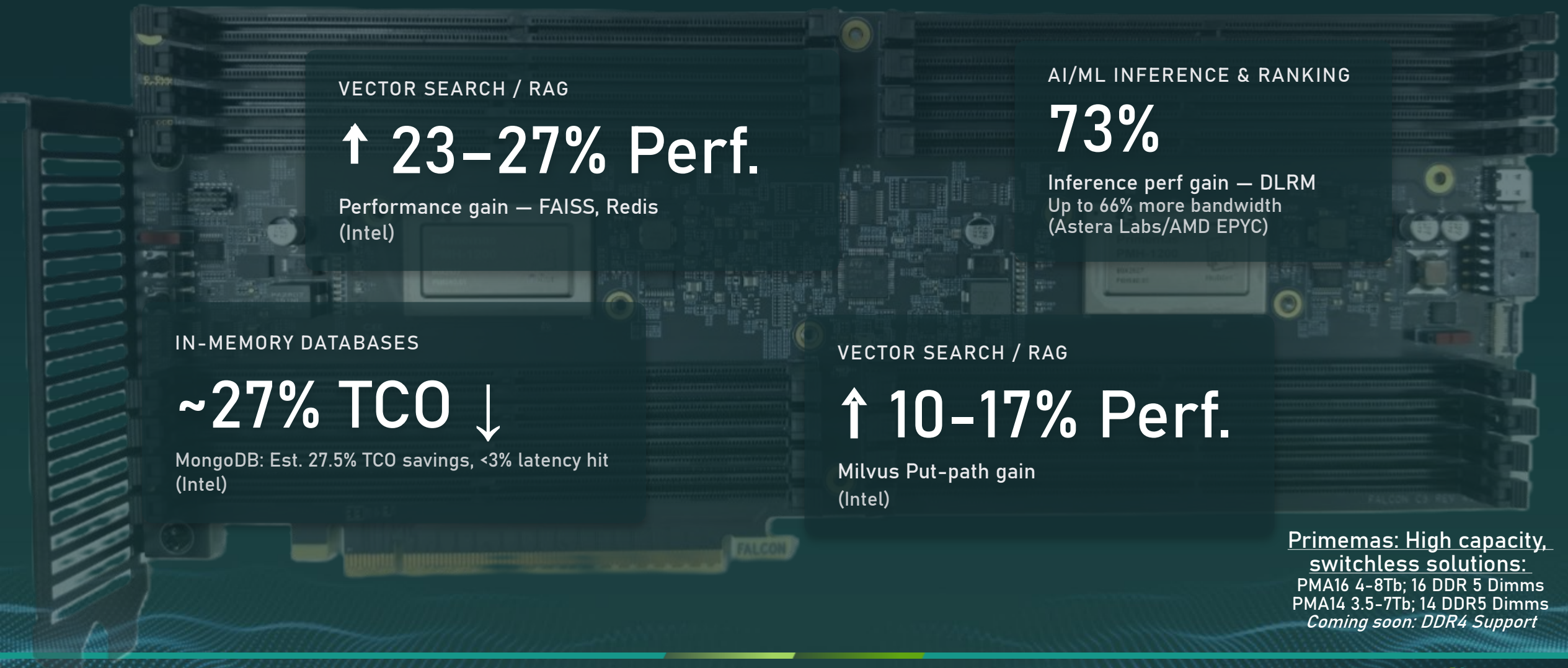
Agility for Next-Gen Workloads

Disaggregation makes memory investments portable and adoptable to AI models that haven't even been designed yet

... Rethink DRAM as an independently scalable, upgradeable, disaggregated resource

Practical Deployment 1: Add-In Cards

Maximize capacity/bandwidth for one host's own working set



VECTOR SEARCH / RAG

↑ 23–27% Perf.

Performance gain — FAISS, Redis
(Intel)

AI/ML INFERENCE & RANKING

73%

Inference perf gain — DLRM
Up to 66% more bandwidth
(Astera Labs/AMD EPYC)

IN-MEMORY DATABASES

~27% TCO ↓

MongoDB: Est. 27.5% TCO savings, <3% latency hit
(Intel)

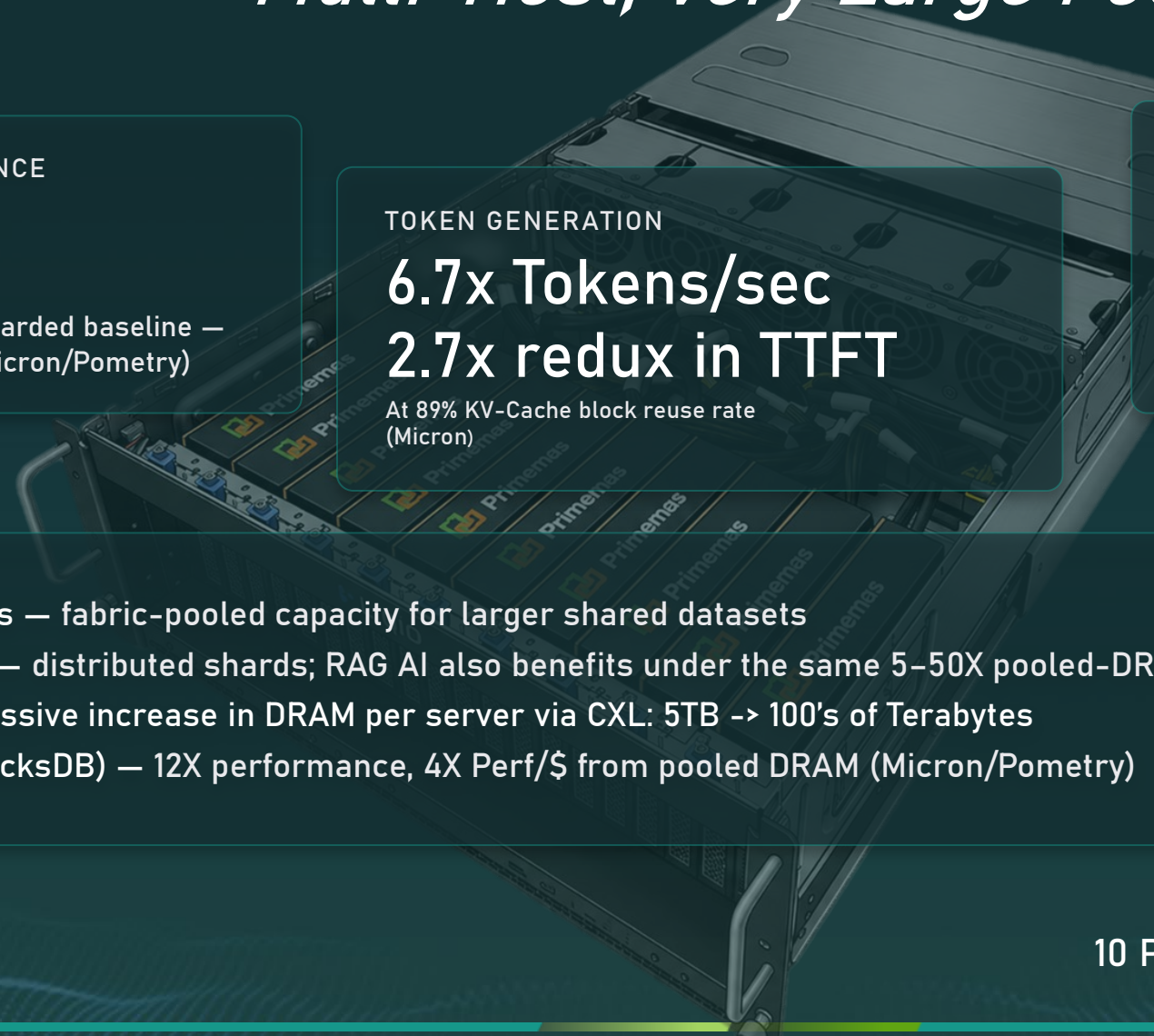
VECTOR SEARCH / RAG

↑ 10–17% Perf.

Milvus Put-path gain
(Intel)

Primemas: High capacity, switchless solutions:
PMA16 4-8Tb; 16 DDR 5 Dimms
PMA14 3.5-7Tb; 14 DDR5 Dimms
Coming soon: DDR4 Support

Practical Deployment 2 – JBOM
Multi-Host, Very Large Pool



GRAPH DB PERFORMANCE

24–40X

higher throughput vs. sharded baseline — no sharding required (Micron/Pometry)

TOKEN GENERATION

6.7x Tokens/sec
2.7x redux in TTFT

At 89% KV-Cache block reuse rate (Micron)

PERFORMANCE/COST

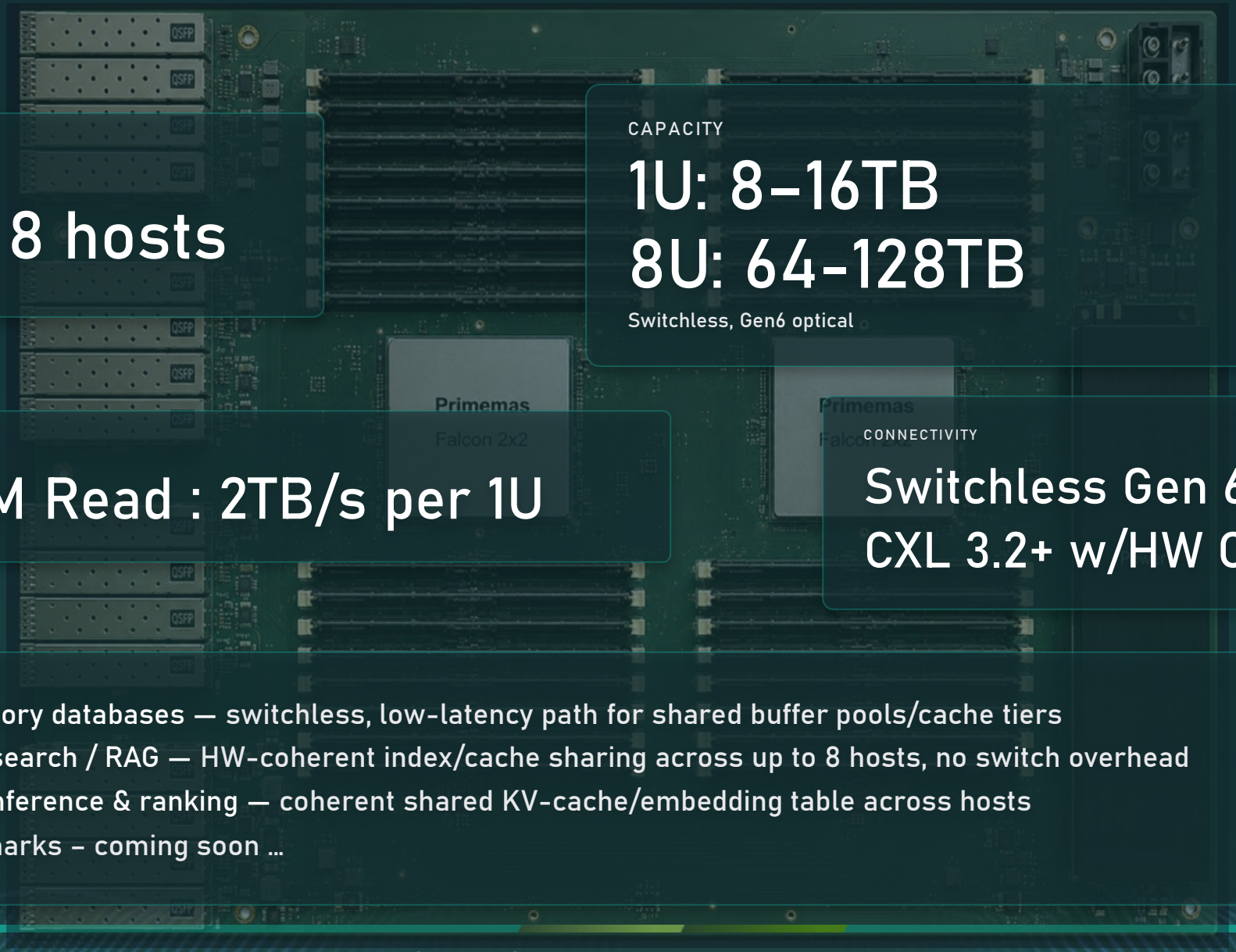
8.0x Perf/$

Reduction of TCO by 88% for equivalent perf. (Micron)

In-memory databases — fabric-pooled capacity for larger shared datasets
Vector search / RAG — distributed shards; RAG AI also benefits under the same 5–50X pooled-DRAM uplift band (Micron/Pometry)
Indexing Cache — Massive increase in DRAM per server via CXL: 5TB -> 100's of Terabytes
Key-value stores (RocksDB) — 12X performance, 4X Perf/$ from pooled DRAM (Micron/Pometry)

Primemas Solution: 40-80TB JBOM
10 PMA 16's w/Marvell switch & Liqid Chassis

Practical Deployment 3 – Multi Host, Low Latency



HOST SCALE

Up to 8 hosts

CAPACITY

1U: 8–16TB
8U: 64–128TB

Switchless, Gen6 optical

BANDWIDTH

DRAM Read : 2TB/s per 1U

CONNECTIVITY

Switchless Gen 6 Optical
CXL 3.2+ w/HW Coherency

In-memory databases — switchless, low-latency path for shared buffer pools/cache tiers
Vector search / RAG — HW-coherent index/cache sharing across up to 8 hosts, no switch overhead
AI/ML inference & ranking — coherent shared KV-cache/embedding table across hosts
Benchmarks – coming soon …

Availability:
Come and talk to Primemas!

Practical Deployment #4: Abaco 3.0

Largest Capacity Shared Memory System for AI & HPC, Powered by Primemas CXL AIC, Micron, and Liqid

Zero-Friction Software Stack

- Most applications run without modification
- RHEL 9.1 Linux kernel
- Liqid Matrix Software as Fabric Manager
- Micron famfs kernel – Open-source submittal to Linux.org

Schedule : Production Q4 2026

- KV Cache for LLM inferencing using Dynamo
- Vector RAG demonstration by Scalemem/Micron
- Graph Analytics/RAG, large DB by software partners

Rack Specifications

1× Fabric Manager host

5× AI Compute Servers
- Including CXL host adapter cards

4× Memory Chassis (up to 160TB total)
- 40TB per 4U chassis
- 10 Primemas CXL Add-in Cards (AICs) per chassis
- 4TB per AIC (16x 256GB DDR5 RDIMM)

Production Ready Reference Design

- Supermicro Hyper X14 Servers

- LIQID EX-5410C Memory Expansion Chassis
- Primemas 4TB Mem Expansion Add-in Card
- Micron 256GB DDR5 RDIMMs

Q&A

Please share your CXL questions!

Thank You

Visit the CXL Pavilion at Booth no. 213 or www.ComputeExpressLink.org to learn more!