

The Imperative for Memory Innovation in the AI Era

AI Leadership Forum, September 30

Arthur Sainio, Director, DRAM

David Selle, Manager, Business Development



Why Penguin Solutions

The AI Factory platform company bringing unmatched expertise to the intersection of AI compute and memory.



25+

AI and HPC
Years Experience



~100,000

GPUs Deployed &
Managed



4+ Billion

Hours of
GPU Runtime



30+

Advanced Memory
Years Experience

AI Hypergrowth is Driving Multi-Dimensional Memory Pressure

Design teams face the squeeze of capacity, bandwidth requirements, supply, and cost

1

Compute Has Outpaced Memory

Peak compute has scaled roughly 1,000x faster than the bandwidth feeding it. Accelerators now spend their time waiting on memory, not on math.

2

Memory Demand is Stacking

Model weights, inference KV cache and agentic state accumulate rather than replace one another – requiring roughly 80x more memory per system by 2030.

3

AI Inputs Compound Supply Constraint

Tokens, context windows and model size have each grown 100x to 1,000x in a few years, while HBM supply has merely doubled per year.

4

Memory Has Repriced: High Cost is the New Norm

Per-bit pricing bottomed in 2023 and has since dramatically repriced higher multiple times over; analysts see no return to the old floor through 2030.

Compute Has Outpaced Memory

Compute

- Roughly 1,000,000x since 1980, from single-core CPUs to today's accelerators

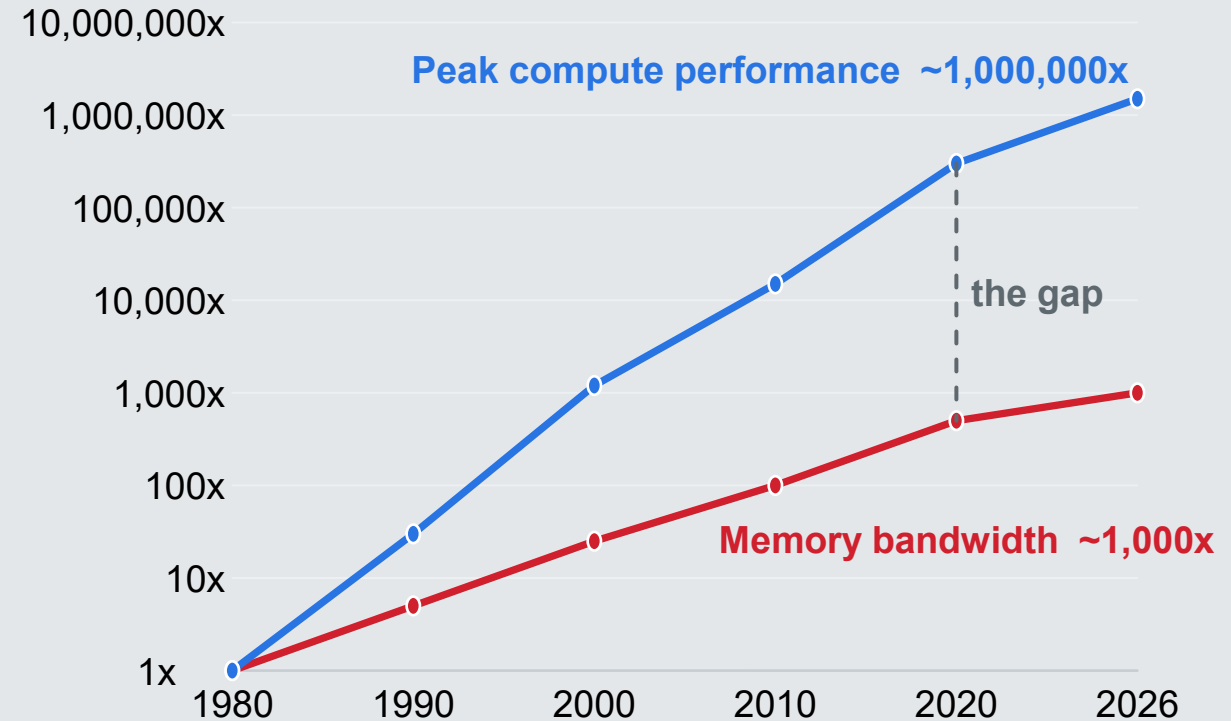
Memory Bandwidth

- Roughly 1,000x over the same period, limited by pins, packaging and power

Resulting Gap

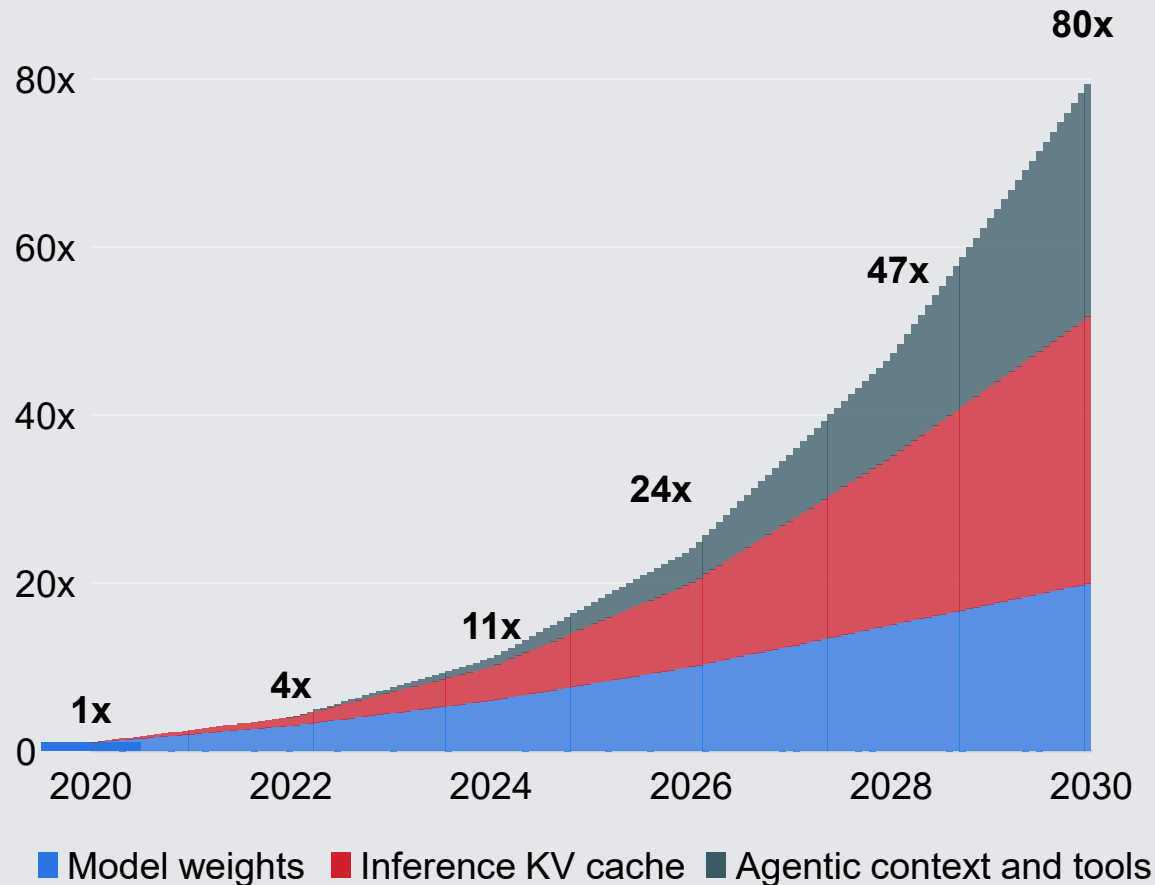
- A 1,000x divergence. Processors now spend their time waiting on memory aka the memory wall

Four Decades of Compute and Memory Growth



Source: Hennessy & Patterson, Computer Architecture: A Quantitative Approach (6th ed.), processor-memory performance gap; Gholami et al., "AI and Memory Wall," IEEE Micro, 2024. Curves are order-of-magnitude illustrations indexed to 1980, plotted on a log scale, not vendor-specific measurements.

Relative Memory Demand Per AI System



Illustrative projection by Penguin Solutions, indexed to 2020. Component trends drawn from published analysis of model-parameter growth, KV-cache scaling in inference (AceCloud, "CXL and the AI Memory Wall," August 2026) and agentic workload memory footprints (Samsung Semiconductor, July 2026). Directional, not a forecast of specific deployments.

Memory Demand is Stacking

Model scale, inference context, and agentic state each drive cumulative demand for memory

LLM Model Scale

- Parameter counts keep rising; weights must be resident before a single token is served

Inference

- KV cache scales with context length and concurrency, on top of the weights

Agentic AI

- Persistent context, tools and multi-step state add a third layer - the fastest growing of the three.

AI Inputs Compound Supply Constraints

METRIC	VALUE PROGRESSION			CHANGE
AI tokens / month Google, all products	9.7T MAY '24	480T MAY '25	3.2Q MAY '26	~330x in two years
Frontier context window max tokens	2K '20	128K '23	2M '24	~1,000x in four years
Largest frontier model parameters	1.5B '19	175B '20	~1.8T* '23	~1,000x in four years
HBM per flagship GPU NVIDIA spec	16GB P100 '16	80GB H100 '22	288GB B300 '25	18x in nine years
HBM bit supply per year	1.5 EB '23	2.8 EB '24		~2x in one year

Confirmed specs in blue. *GPT-4-class parameter counts are reported, not disclosed, and sparse mixture-of-experts makes raw counts approximate. Source: Memory Analyst, "The AI Memory Thesis," track record; underlying data from Google I/O 2026, OpenAI, Anthropic and Google model cards, NVIDIA datasheets, Yole Group.

Source: "Memory is cyclical. But Demand is inflecting faster than supply," Memory Analyst author

High Memory Price is the New Norm

2023: The Market Floor

- DRAM revenue hits lowest level since 2016; ASPs fell 20% in 1Q23 alone (Trend Force, Yoke Group)

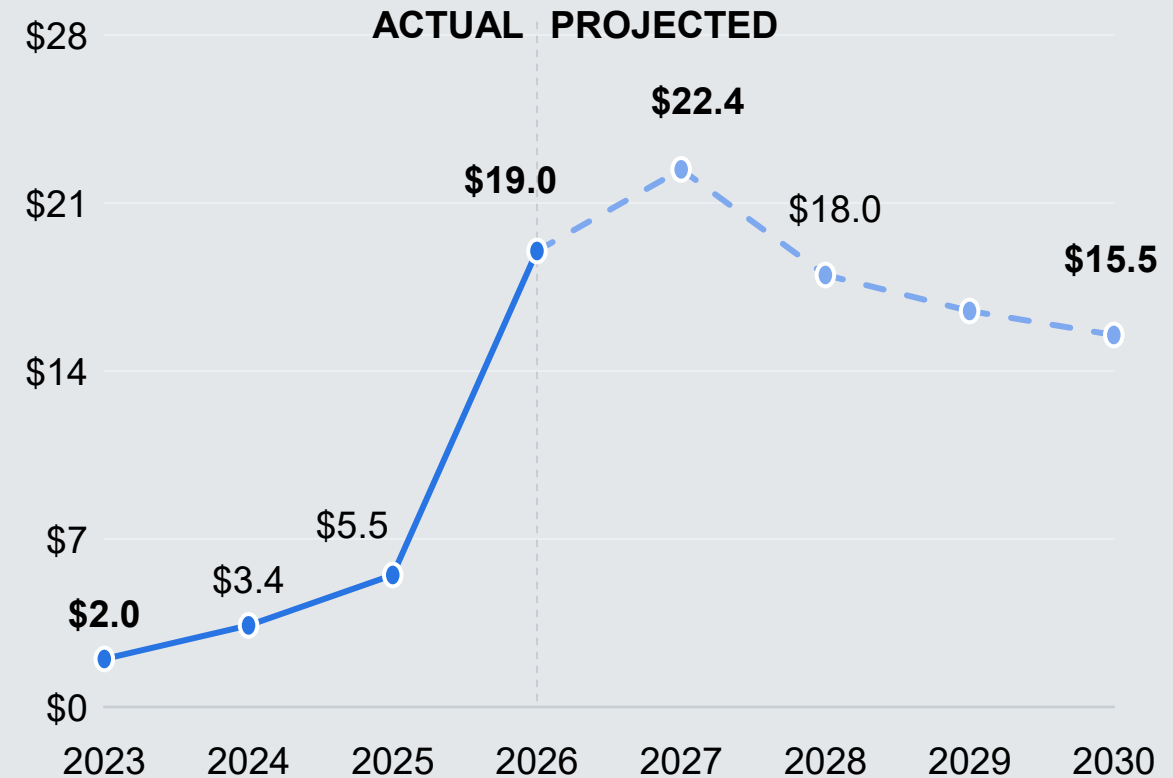
2026: The Spike

- 64GB DDR5 RDIMM contract pricing runs from \$873 in 1Q26 to a forecast \$1,586 by 4Q26, an 82% rise in one year

2027-2030: The New Floor

- Tight supply into 2027, a modest correction in 2028, and sustained vendor profitability through 2030 - no return to 2023 pricing (Gartner)

Server DDR5 Contract Price (USD per gigabyte, annual average)



Actuals 2023-2026 from TrendForce DRAM contract-price reporting and tracked DDR5 RDIMM contract pricing; 2027-2030 directional projection built from Gartner (Forecast: NAND Flash Market Statistics, Supply and Demand, Worldwide, 2023-2030, June 2026), TrendForce (memory outlook, July 2026), Morgan Stanley (Asia-Pacific technology research, July 2026) and Yole Group (NAND and DRAM markets, September 2024). Annual averages, directional - not a price quote.

Organizations Must Take a Multi-Faceted Approach

No single memory solution exists to solve all four challenges. It is essential to pursue the right approach for the right problem.

Design-In of New Memory Components

- Breaking down the memory wall
- Use of SOCAMM, MRDIMM and CXL add-in cards

Today: The Memory Advantage in the AI Era

Creation of Novel Memory Tiers

- KV cache servers and memory pooling
- Create memory tiering

Oct. 28: Scaling AI with MemoryAI KV Cache Servers

Finding Economic Efficiencies from Prior Investment

- Memory re-use: finding capacity you already own

Nov. 23: Unlocking New Value from DDR4 with CXL

Planning for Photonic Breakthroughs on the Horizon

- Breaking bandwidth barriers with photonics

Jan. 13: The Future of Memory: Photonics and Beyond

Today's Topic:

The Memory Advantage in the AI Era

September 30, 2026



Agenda

AI Is Reshaping Infrastructure and Memory Demand

The Shift to Workload-Optimized Memory

DDR5 Solutions for Performance, Density, and Reliability

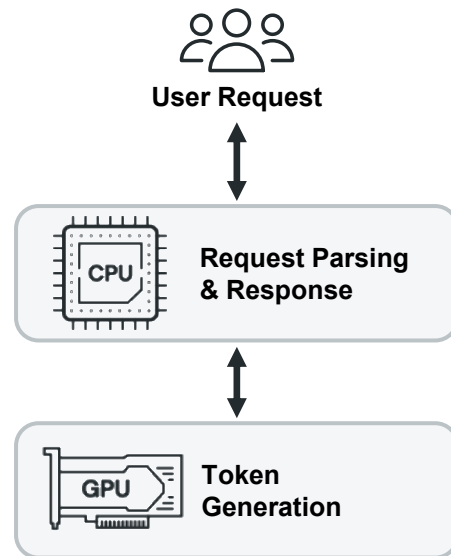
Emerging Architectures: MRDIMM, SOCAMM2, and CXL

Penguin's Memory Portfolio for the AI Era

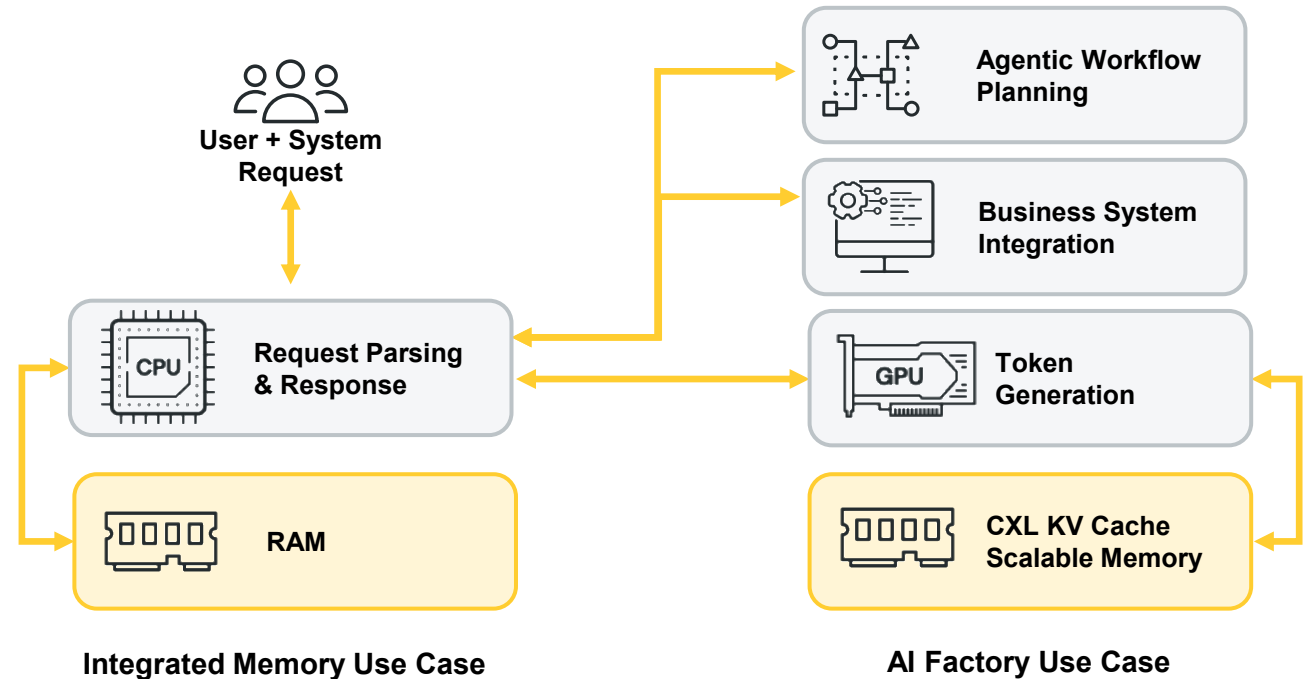
Integrating AI Within Business Flows is Changing the Landscape

Inference and Agentic AI workflows move into production at scale, driving increased demand for CPUs and memory

Pre-Agentic AI Inference

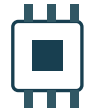


Inference at scale with Agentic AI



Agentic AI Increases Demand for Memory

More CPUs • More System Memory • Memory-centric AI Infrastructure



CPU : GPU Ratio

More host CPU per accelerator (fleet-level)

TODAY



AGENTIC AI



C = host CPU G = GPU

1 : 4



Toward 1 : 1

Agent workflows are orchestration-heavy, so the fleet-level CPU:GPU ratio moves toward 1:1 — via added sockets and dedicated CPU racks alongside GPU nodes. Both pull DDR5 RDIMM content.



System Memory per GPU

2–4x more capacity per accelerator

TODAY

~512 GB

AGENTIC AI

1–2 TB per GPU

Capacity per accelerator

~512 GB



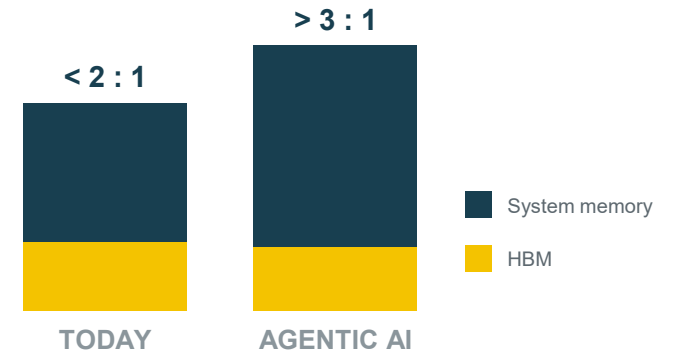
1–2 TB

Larger working sets and persistent agent context push capacity from ~512 GB toward 1–2 TB per GPU, driving DDR5 and LPDDR5 attach.



System Memory : HBM

Capacity grows faster than HBM



TODAY

AGENTIC AI

< 2 : 1



> 3 : 1

A capacity ratio, not bandwidth: value moves beyond the accelerator toward capacity and tiering — the MRDIMM and CXL opportunity.

Sources: TechInsights, DRAM Market Update Q2 2026 and "System Memory is No Longer an Afterthought in AI Server Design," Apr 2026; AMD, Arm and Intel commentary, May–Aug 2026; TrendForce, May 2026.

AI Is Reshaping the Server Memory Hierarchy

From one-size-fits-all DRAM to workload-optimized memory architectures

Four AI-driven Market Transitions

1

Mainstream
DDR5 RDIMM

Core CPU-attached capacity for compute, storage and enterprise AI.

2

Performance
DDR5-12800 MRDIMM

Extreme host-memory bandwidth for AI, HPC and analytics.

3

Efficiency
LPDDR5x SOCAMM2

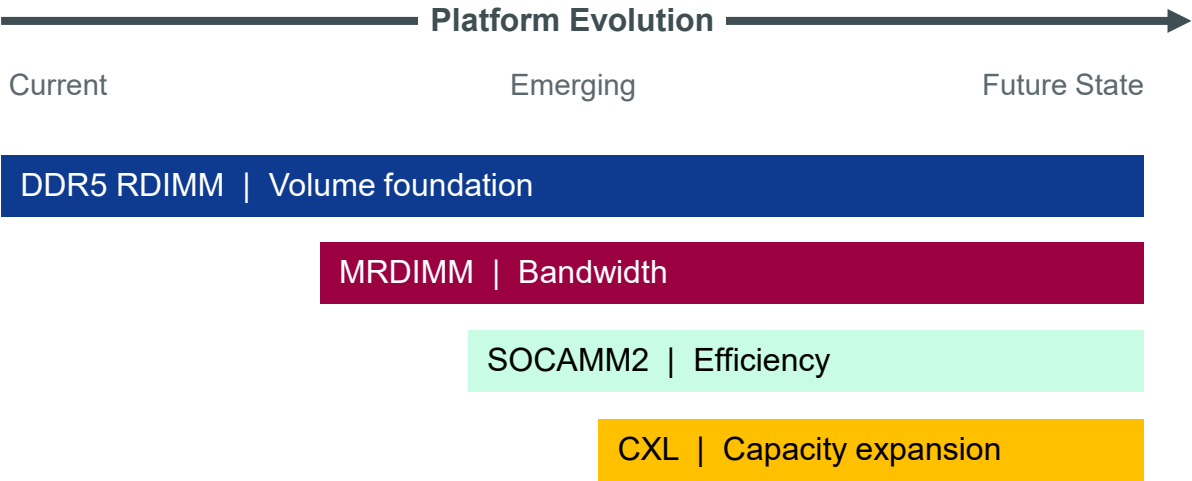
High bandwidth-per-watt in compact, modular AI platforms.

4

Expansion
CXL Memory

Tiered capacity, pooling and composable memory infrastructure.

Server Memory Evolution



AI infrastructure is creating specialized memory tiers optimized for performance, power efficiency and scalable capacity.

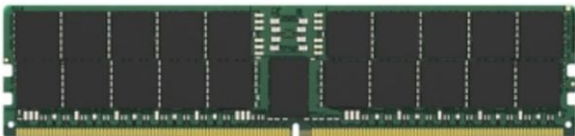
AI Infrastructure Is Driving Memory Specialization

Match bandwidth, capacity, power and form factor to the workload

Mainstream	Performance	Efficiency	Expansion
DDR5 RDIMM	MRDIMM GEN2	LPDDR5x SOCAMM2	CXL MEMORY
Capacity	Bandwidth	Power	Scale
Best-fit Workloads General compute Storage + enterprise Mainstream AI servers	Best-fit Workloads AI inference + training HPC + simulation Real-time analytics	Best-fit Workloads AI inference platforms Power-constrained racks High-density systems	Best-fit Workloads Tiered capacity Memory pooling Composable infrastructure
Architecture 5,600–8,000 MT/s Up to 256GB (3DS)	Architecture Up to 12,800 MT/s Up to 256GB+	Architecture Up to 9,600 MT/s 55-70% lower-power	Architecture PCIe / EDSFF attached Multi-TB potential
Best general-purpose TCO	Maximum host throughput	Bandwidth per watt	Flexible capacity expansion

DDR5 RDIMMs built for AI and data center uptime

Scale CPU-attached memory performance with a broad DDR5 portfolio and optional Zefr™ high-reliability screening.



16GB–256GB

DDR5-5600 | 6400 | 7200 | 8800

Multiple rank and device configurations

Accelerate AI infrastructure

CPU-attached memory for AI inference, HPC, virtualization, compute and storage servers.

Scale platform choice

Capacity and speed options to align memory cost, bandwidth and density to system requirements.

Protect mission-critical uptime

Zefr™ screening is designed to eliminate over 90% of memory reliability failures.

SMART Zefr™ High Reliability

Optional proprietary screening for uptime-focused deployments

Match the workload, then differentiate with SMART reliability and configuration flexibility.

AI / ML

HPC

DATA CENTER

COMPUTE + STORAGE

On-Die ECC Can Hide Failing DDR5 Modules

GFLOPS degradation exposes correction activity before early-life failures becomes an outage

Hidden correction activity

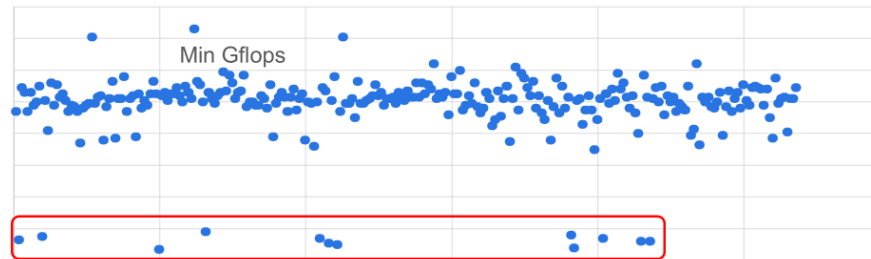


- On-die ECC corrects single-bit errors internally
- Corrections may not appear in standard system logs
- A degrading module can still look “healthy”

Risk: Performance loss or sudden ELF

Performance reveals the hidden signal

Minimum GFLOPS across tested DDR5 RDIMMs



Repeated ECC correction activity introduces latency; lower GFLOPS identifies modules that warrant deeper screening.

Outlier GFLOPS = early warning flag

SMART Zefr™ mitigates the risk

1 Stress

Elevated temperature, voltage and extended test time

2 Detect

ECC events and GFLOPS outliers surface weak modules

3 Screen

Reject suspect modules before customer deployment

Outcome: Higher uptime and fewer field failures

Treat sustained GFLOPS outliers as an actionable precursor signal, then use Zefr screening to remove weak modules before deployment.

[Technical brief ↗](#)

DDR5 RDIMMs Built for Immersion-Cooled AI Infrastructure

Protect memory reliability as compute density and thermal demands rise

Protect critical memory

Conformal coating creates a barrier against oxidation, moisture ingress and corrosive fluid exposure.

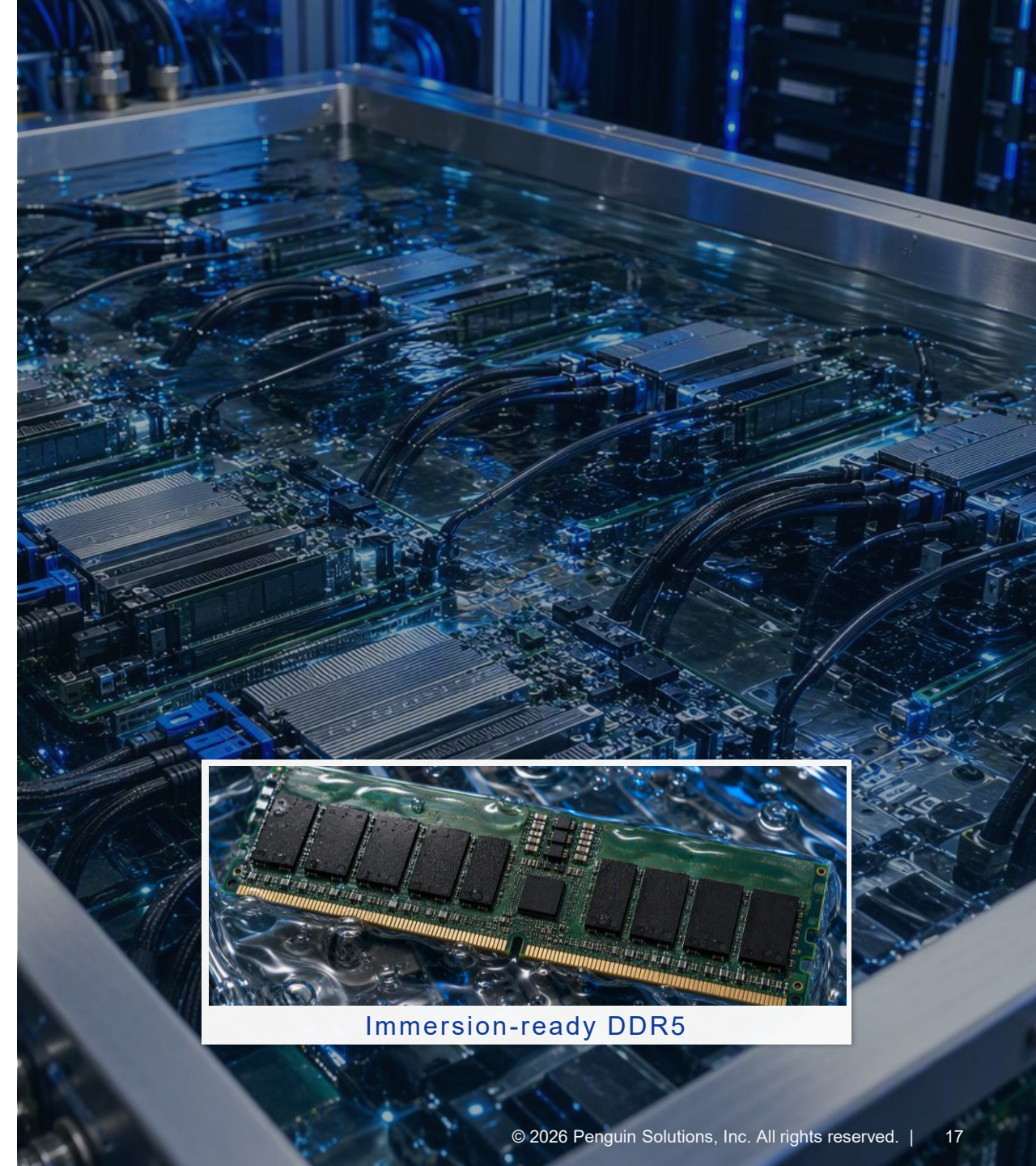
Sustain reliability at scale

Zefr-tested DDR5 RDIMMs support uptime-focused deployments in high-density immersion-cooled servers.

Match demanding workloads

32GB to 256GB capacity options for AI/ML, HPC and dense database analytics. SRL part-number prefix.

High-density DDR5 performance with immersion-ready protection for AI data centers.



Immersion-ready DDR5

DDR5 VLP RDIMMs for AI and Data Center Blades

Maximize memory density and airflow in space-constrained blade and 1U platforms

Very Low Profile. High-density Compute



Vertical placement enables compact memory layouts in blade and 1U systems while preserving airflow around densely packed components.

18.75

mm module height

Designed for vertical installation in dense blade architectures

DDR5 VLP RDIMM

32GB	2Rx8	16Gb	5600 / 6400
48GB	2Rx8	24Gb	5600 / 6400
64GB	2Rx8	32Gb	5600 / 6400

Capacity and speed availability should be validated for the target platform.

Purpose-built For Dense Infrastructure

Ai Compute Blades

Compact CPU-attached memory for dense AI infrastructure and accelerator host systems.

Compute + Storage

Increase memory capacity in blade servers where board height and airflow are constrained.

Network + Telco

Fit telecom and networking blades built for high availability and compact chassis.



Blade-optimized form factor

Maximize system density, maintain airflow and scale DDR5 capacity in space-constrained AI and data center blades.

DDR5-12800 MRDIMMs for AI and Data Centers

Double host-memory bandwidth to keep data-intensive compute fed

Extreme Bandwidth, Standard DDR Economics



Multiplexed rank architecture accesses memory ranks in parallel to increase channel throughput.

2X PEAK BANDWIDTH

DDR5-6400 RDIMM

6,400 MT/s

SMART MRDIMM

12,800 MT/s

DDR5-12800 PRODUCT

32GB 2Rx8 SR4G8RM528JNC/SQ/MH

64GB 4Rx8 SR8G8RM528JNC/SQ/MH

128GB 4Rx8 SRAG8RM548JHM/ME

*Platform support is CPU- and system-dependent;
validate against the target server roadmap.*

Built For Bandwidth-intensive Workloads

AI Inference

Feed CPUs and accelerators for LLM inference and memory-bound model execution.

HPC + Simulation

Keep high-core-count processors supplied for scientific and engineering workloads.

Real-time Analytics

Increase throughput for large databases, in-memory analytics and complex queries.

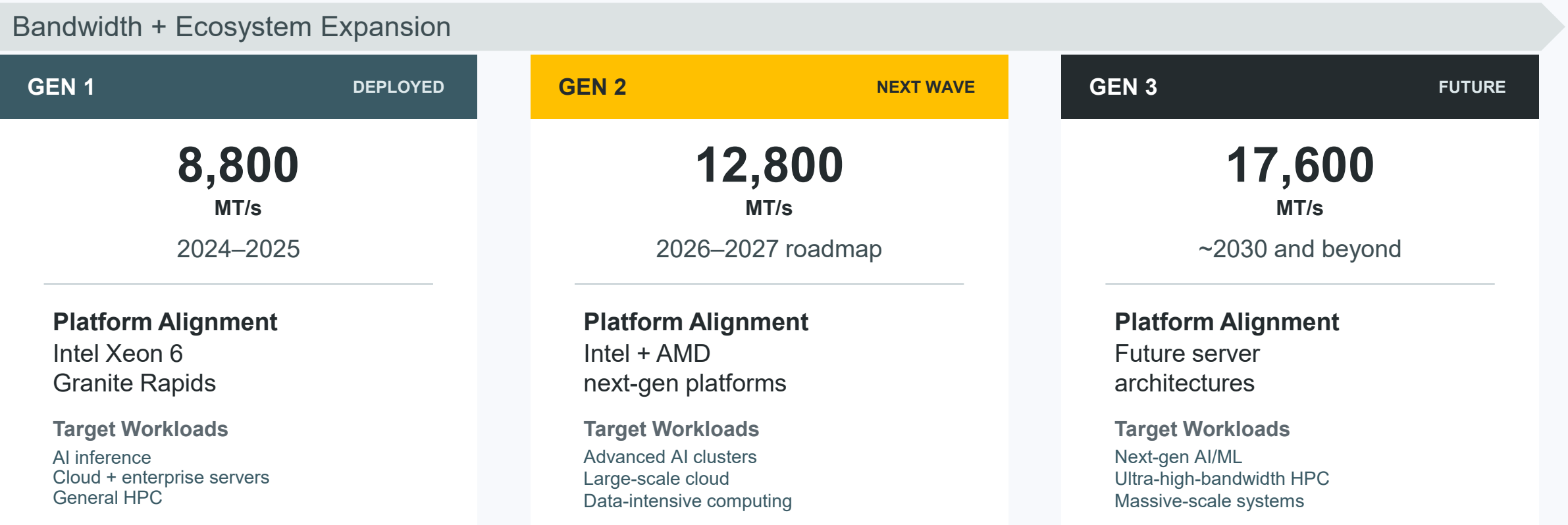
Cloud Infrastructure

Scale host-memory bandwidth for data-intensive enterprise and cloud platforms.

More bandwidth per memory channel for AI and data center performance, without moving all host memory to HBM.

DDR5 MRDIMM: Scaling Bandwidth for the AI Era

A three-generation roadmap from 8,800 to 17,600 MT/s



MRDIMM creates a scalable host-memory bandwidth path for increasingly data-intensive AI, cloud and HPC platforms.

DDR5-12800 MRDIMM: The Practical Host-Memory Choice

Use MRDIMM for scalable capacity and host bandwidth; reserve HBM for the accelerator's hottest data.

Cost / TCO

MRDIMM Advantage

- ✓ DDR-based module economics
- ✓ Socketed and field-replaceable
- ✓ Avoids advanced HBM packaging

HBM Trade-off

- Premium memory + packaging
- Fixed to accelerator package
- Higher integration complexity

Performance

MRDIMM Advantage

- ✓ 12,800 MT/s
- ✓ Up to 2× DDR5-6400 peak bandwidth
- ✓ Strong fit for memory-bound CPU workloads

HBM Advantage

- Highest bandwidth near the GPU
- Best for bandwidth-critical working sets
- Not the cost-effective answer for all data

Density + Flexibility

MRDIMM Advantage

- ✓ 32GB–128GB per module
- ✓ Scales across server DIMM slots
- ✓ Capacity can be configured and upgraded

HBM Trade-off

- Capacity fixed at manufacture
- Not socketed or independently serviceable
- Scaling capacity requires more accelerators

MRDIMM delivers standard DDR economics + serviceable density + up to 2× DDR5-6400 bandwidth

LPDDR5x SOCAMM2

A modular, low-power memory architecture moving from early adoption to AI data center scale

What It Is

**JEDEC-standard, detachable
LPDDR5/5X server memory**

SOCAMM2 combines low-power DRAM with a compact compression-attached module, bringing upgradeability and serviceability to high-bandwidth AI system memory.

64–256GB

module capacities

Up to 9600 MT/s

published speed

Up to 154 GB/s

bandwidth per
module

Why It Matters To AI Infrastructure

High Bandwidth

Feeds CPU and accelerator-centric inference workloads

Lower Power

Reduces memory energy and cooling burden

Compact + Modular

Improves density, airflow and field serviceability

Adoption Trend: From Qualification To Platform Scale

1 Early supplier validation

Customer samples and initial SOCAMM2 products establish the form factor.

2 AI platform adoption

JEDEC standardization and next-generation accelerator platforms drive qualification.

3 Ecosystem expansion

Multiple memory suppliers, OEMs and hyperscalers broaden deployment options.



AI Ecosystem Momentum Builds Around LPDDR5x SOCAMM2

Modular, high-bandwidth, low-power memory for next-generation AI infrastructure

High Bandwidth

Up to 154 GB/s per module

Lower Power

Reduces memory energy + cooling

Modular

Upgradeable + serviceable

Compact

Improves density + airflow

Ecosystem Adoption Signals

NVIDIA

Vera / Rubin

AI Role

Flagship Arm-based AI platform for large-model training and inference.

SOCAMM2 Signal

Internal roadmap: up to 1.5TB via 16 modules and up to 1.2TB/s.

AMD

EPYC “erano”

AI Role

Power-optimized host CPU aligned to next-generation Instinct GPU systems.

SOCAMM2 Signal

Native LPDDR5x SOCAMM2 support.

QUALCOMM

Next-Gen AI

AI Role

Enterprise inference architecture evaluating standardized modular memory.

SOCAMM2 Signal

Exploring SOCAMM2 as an alternative to soldered LPDDR layouts.

INTEL

TBDs

AI Role

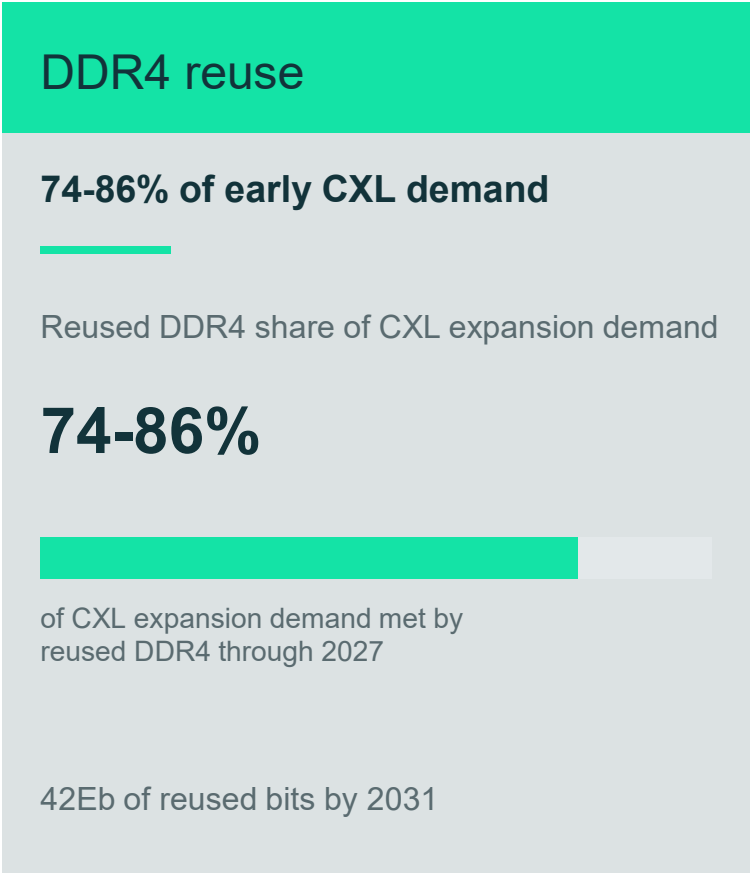
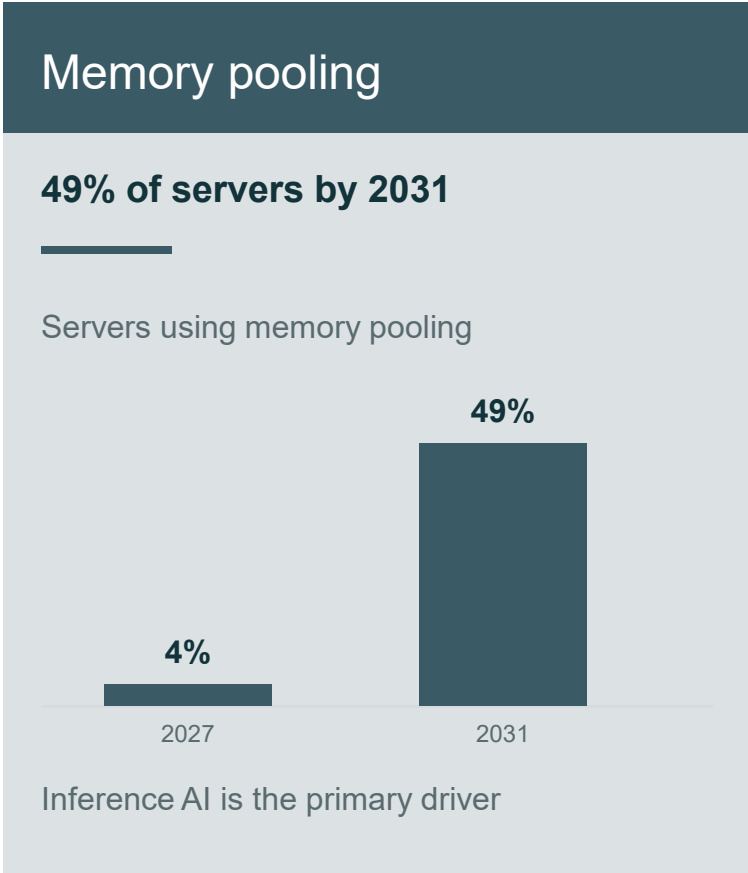
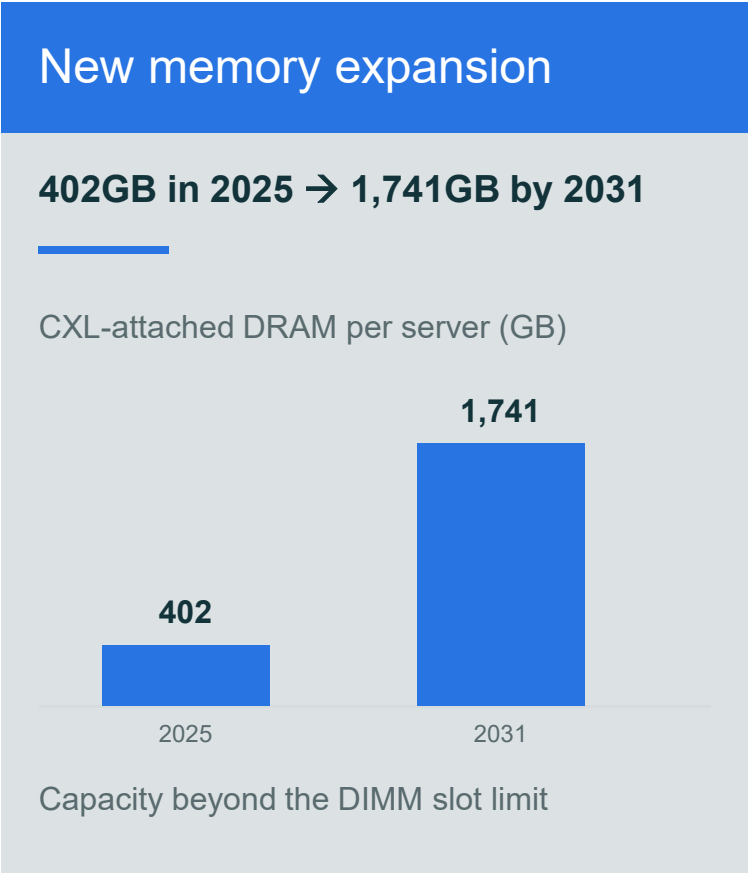
AI host CPU roadmap for inference and memory-intensive workloads.

SOCAMM2 Signal

SOCAMM2 configuration TBD .

CXL Expands Memory Capacity for AI Infrastructure

Capacity expansion, pooling, and DDR4 reuse increase memory content per server



Source: TechInsights, DRAM Market Report Q2 2026 (pp. 47-52)

CXL Memory Solutions for AI

A form-factor and media portfolio aligned to server constraints, workload latency and serviceability requirements



MAX CAPACITY

**DDR5 CXL 8-DIMM
AIC**

Up to 1TB

DDR5-4800 | CXL 2.0 x16
(2 x x8)

FHHL, dual width

High-density expansion
for AI, HPC and databases



BALANCED

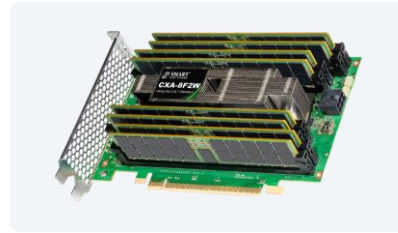
**DDR5 CXL 4-DIMM
AIC**

Up to 512 GB

DDR5-4400 | CXL 2.0 x16

FHHL, single width

Capacity expansion where
slot width and power are
constrained



REUSE

**DDR4 CXL 8-DIMM
AIC**

512 GB with 64 GB RDIMMs

DDR4-3200 | CXL 2.0 x16

FHHL, dual width

Re-deploy qualified DDR4
in new DDR5-era servers



SERVICEABLE

CXL CMM

64 / 96 / 128 GB

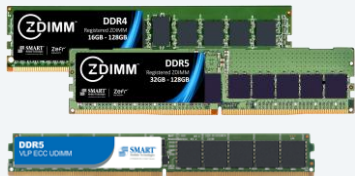
DDR5-4800 | CXL 2.0 x8

EDSFF E3.S 2T

Front-of-box expansion
with easier service access

Memory Solutions for AI

A form-factor and media portfolio aligned to server constraints, workload latency and serviceability requirements.



CORE CAPACITY
DDR5 RDIMMs & VLP

16GB–256GB | 5600–7200 MT/s

Mainstream compute, storage and AI/ML servers.



SUSTAINABLE COOLING
Liquid Immersion RDIMMs

Up to 256GB | Conformal coated

Fluid-exposed, high-density immersion-cooled servers.



EXTREME BANDWIDTH
DDR5-12800 MRDIMMs

32GB–128GB | 12,800 MT/s

AI inference, HPC and real-time analytics.



EMERGING AI MEMORY
LPDDR5x SOCAMM2

Modular LPDRAM
55-70% lower-power

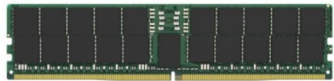
Next-generation AI inference platforms.

The Memory Advantage: Match Memory to the Workload

AI infrastructure requires a portfolio optimized for capacity, bandwidth, efficiency and expansion

Core Capacity

DDR5 RDIMMs



16GB–256GB

Volume foundation for compute, storage and mainstream AI. Broad speed, density and reliability options.

Host Bandwidth

DDR5 MRDIMMs



Up to 12,800 MT/s

Up to 2× DDR5-6400 peak bandwidth for AI inference, HPC and real-time analytics.

Power Efficiency

LPDDR5x SOCAMM2



Up to 154 GB/s

Compact, modular memory that improves bandwidth per watt, density, airflow and serviceability.

Capacity Expansion

CXL Memory



Up to 1 TB per AIC

Adds tiered capacity beyond DIMM slots and enables pooling, composability and DDR4 reuse.

Lead with the workload, then select the memory architecture that delivers the right balance of performance, capacity, power, and TCO.

Where This Memory Innovation Series Goes Next

October 28



Scaling AI with MemoryAI KV Cache Servers

How KV cache servers turn memory capacity into inference throughput, and the advantages for cluster design

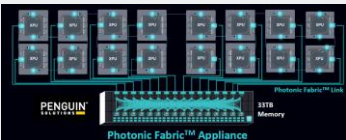
November 23



Unlocking New Value from DDR4 with CXL

How to add capacity from DDR4 you already own, and the math of cost, supply and sustainability

January 13



The Future of Memory: Photonics and Beyond

Where interconnects are heading, and what photonics means for pooling memory outside the rack

Talk to Penguin

Evaluate our Integrated Memory portfolio

Discuss your product qualification process

Explore design-in and integration with our team

Today's Presenters



Arthur Sainio
Director, DRAM
arthur.sainio@penguinsolutions.com



David Selle
Manager, BD
david.selle@penguinsolutions.com

