PENGUIN SOLUTIONS

Scaling AI with MemoryAI™ KV Cache Servers

AI Leadership Forum

Kumar Natarajan, Principal Product Manager



Why Penguin Solutions

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



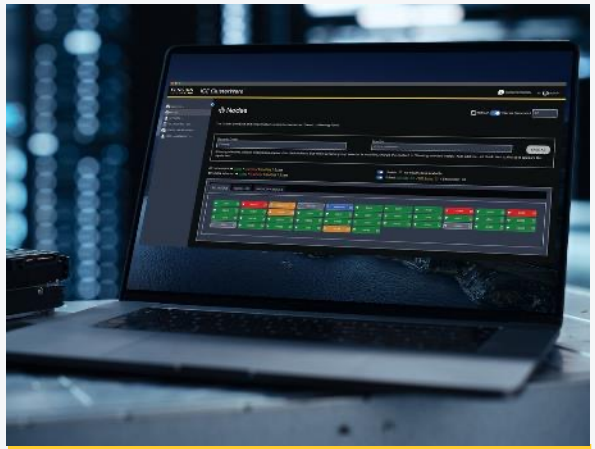
25+

AI and HPC
Years Experience



~100,000

GPUs Deployed &
Managed



4+ Billion

Hours of
GPU Runtime



30+

Advanced Memory
Years Experience

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

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.

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved.

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.

Creation of Novel Memory Tiers

- KV cache servers and memory pooling
- Create memory tiering

Today: 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

Design-In of New Memory Components

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

On-Demand: The Memory Advantage in the AI Era

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

Agenda

Inference and KV Cache

Offloading KV Cache to External Tier

Penguin MemoryAI™ KV Cache Server

Performance Testing with Penguin MemoryAI™

PENGUIN SOLUTIONS

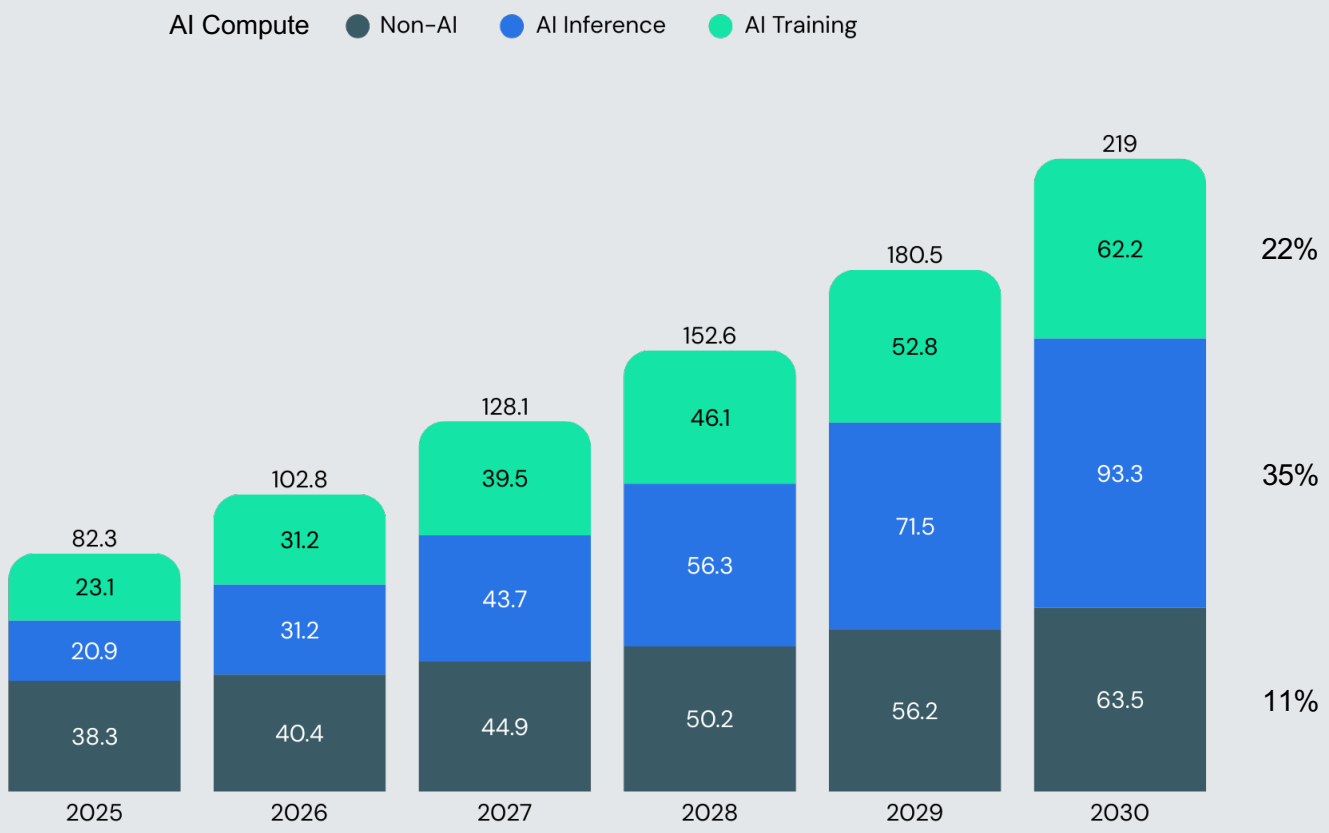
© 2026 Penguin Solutions, Inc. All rights reserved. |

AI Training vs. Inferencing

> Inference workloads could make up more than 40% of data center demand in 2030, growing 35% CAGR until 2030.

McKinsey & Company
December 17, 2025

Global data center demand by workload 2025-2030, gigawatts



Note: includes all provider types
Source: McKinsey Data Center Demand Model [The next big shifts in AI workloads and hyperscaler strategies]

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

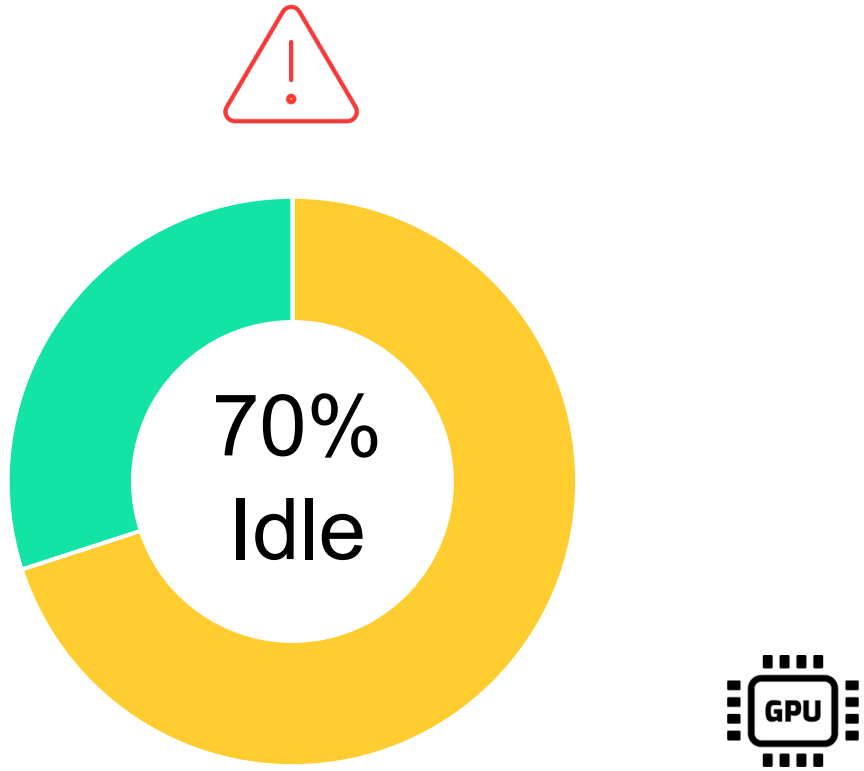
The “Memory Wall” Challenge in AI Inference

Memory is a fundamental limitation of enterprise-scale AI inference

- Memory usage in inference
  - Model loaded to HBM
  - Sequential prefill/decode process
    - Memory size grows linearly with context size
    - Memory consumption grows exponentially with concurrency

GPU is memory-bound during inference

If GPU becomes “memory-starved,” token generation slows and compute goes idle



PENGUIN SOLUTIONS

Source: The Impact of Storage on the AI Lifecycle - WEKA

© 2026 Penguin Solutions, Inc. All rights reserved. |

How KV Caching Addresses the Memory Wall

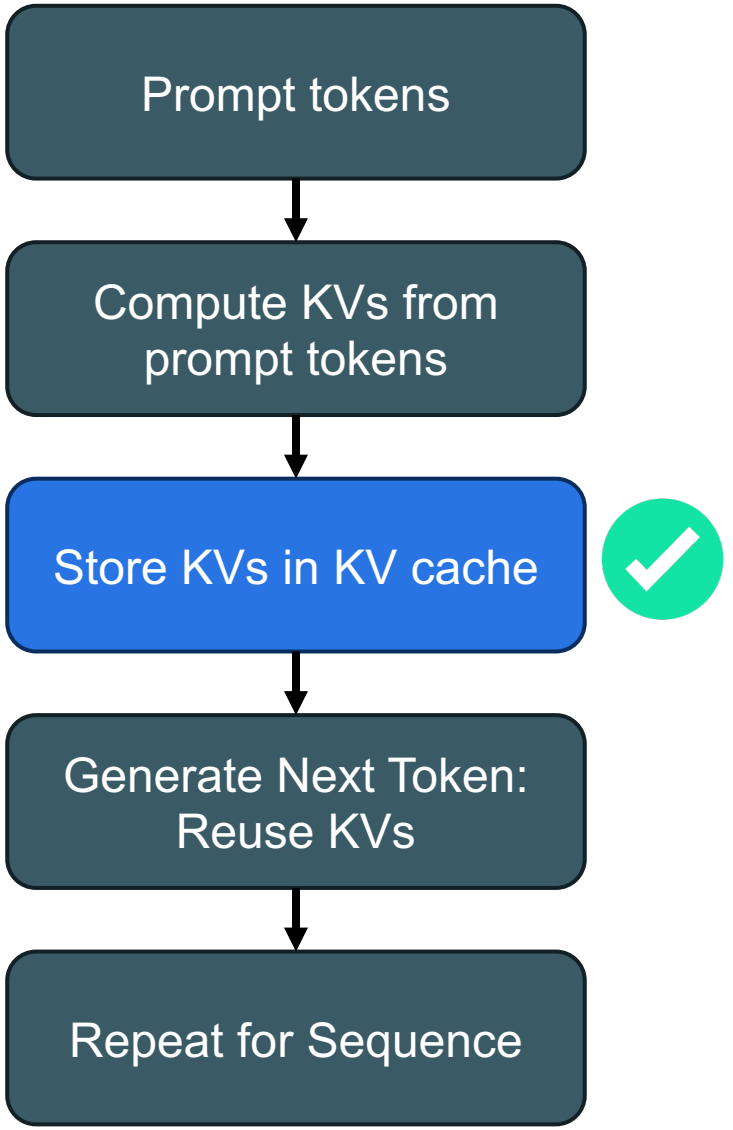
KV caching is an optimization technique in AI models to store computed "keys" (K) and "values" (V) in cache memory for re-use.

How it works

- Use of KV cache avoids re-calculation of keys and values.
- During inference, GPU reuses KV from past tokens.
- Reuse speeds up generation of new response, driving lower latency.

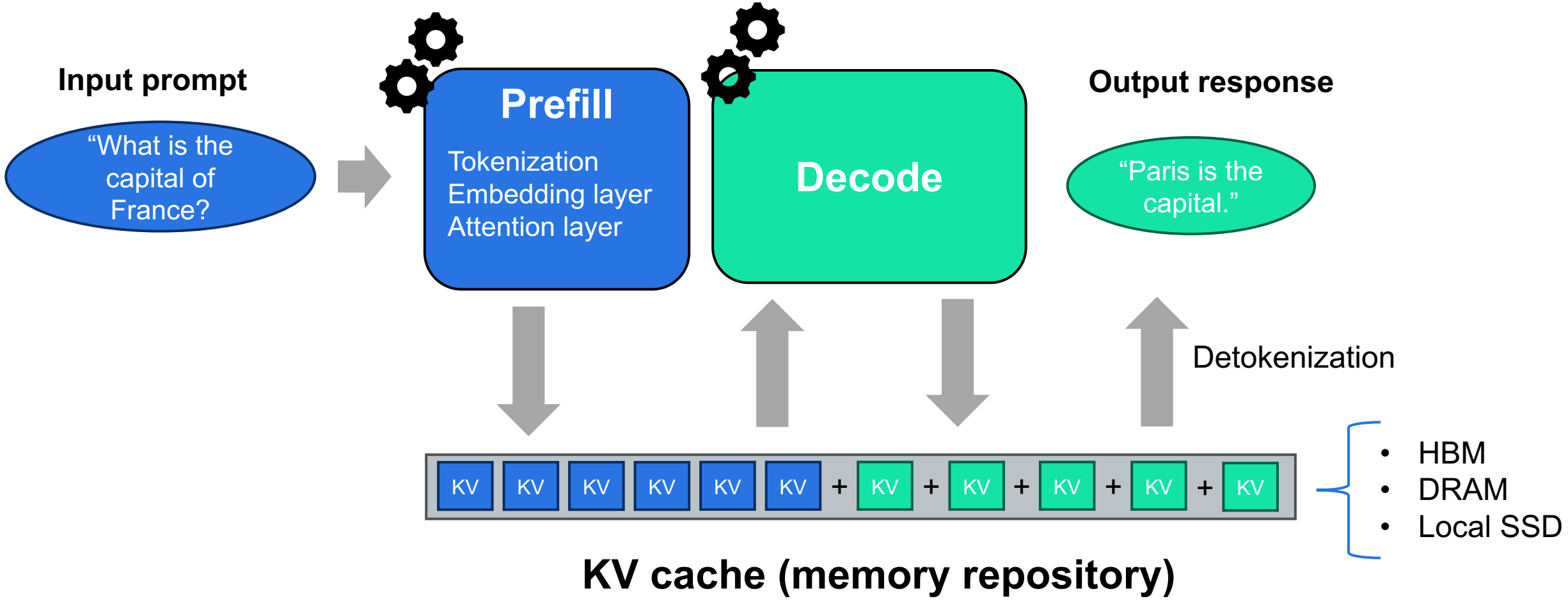
Why it matters

- Larger context requires more memory; KV reuse speeds TTFT
- Latency is essential for fast responses in chatbots or real-time AI.
- KV efficiency reduces GPU idle time of waiting for memory.



© 2026 Penguin Solutions, Inc. All rights reserved. |

Role of KV Cache in Inference

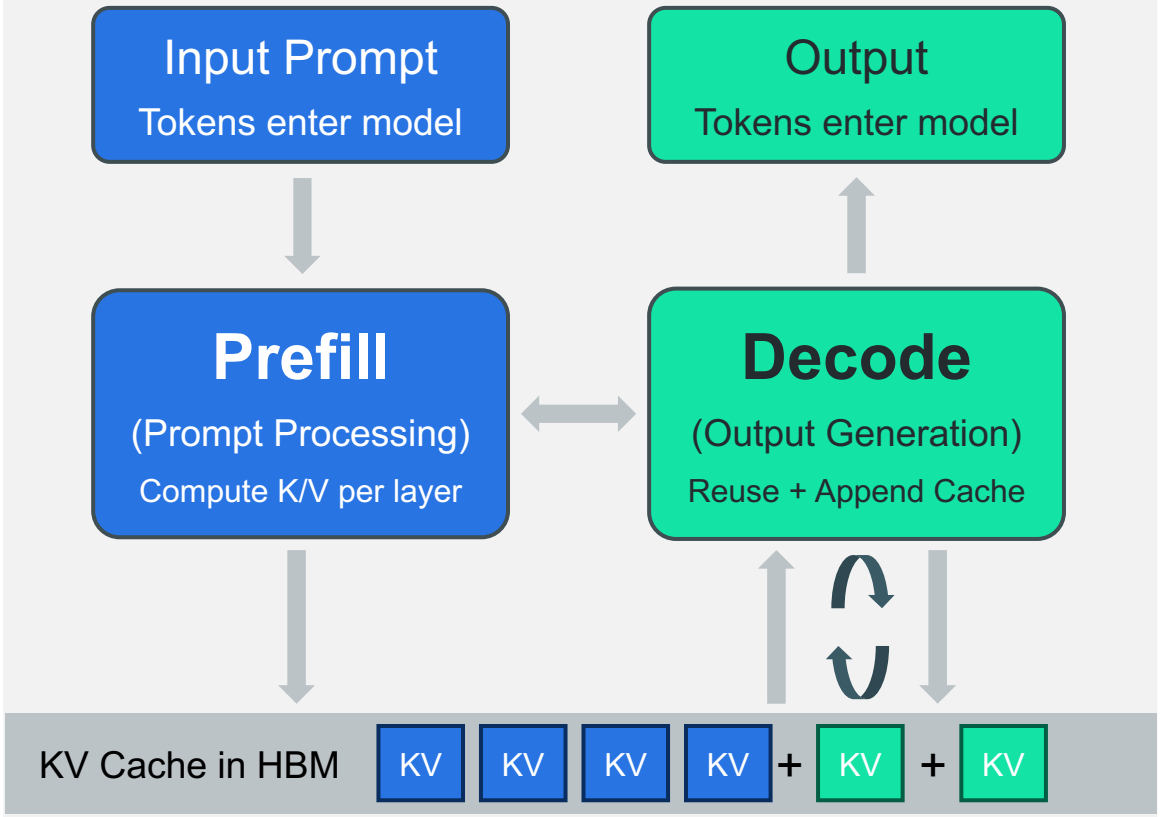


© 2026 Penguin Solutions, Inc. All rights reserved.

KV Cache: Fast Inference. Growing HBM Pressure

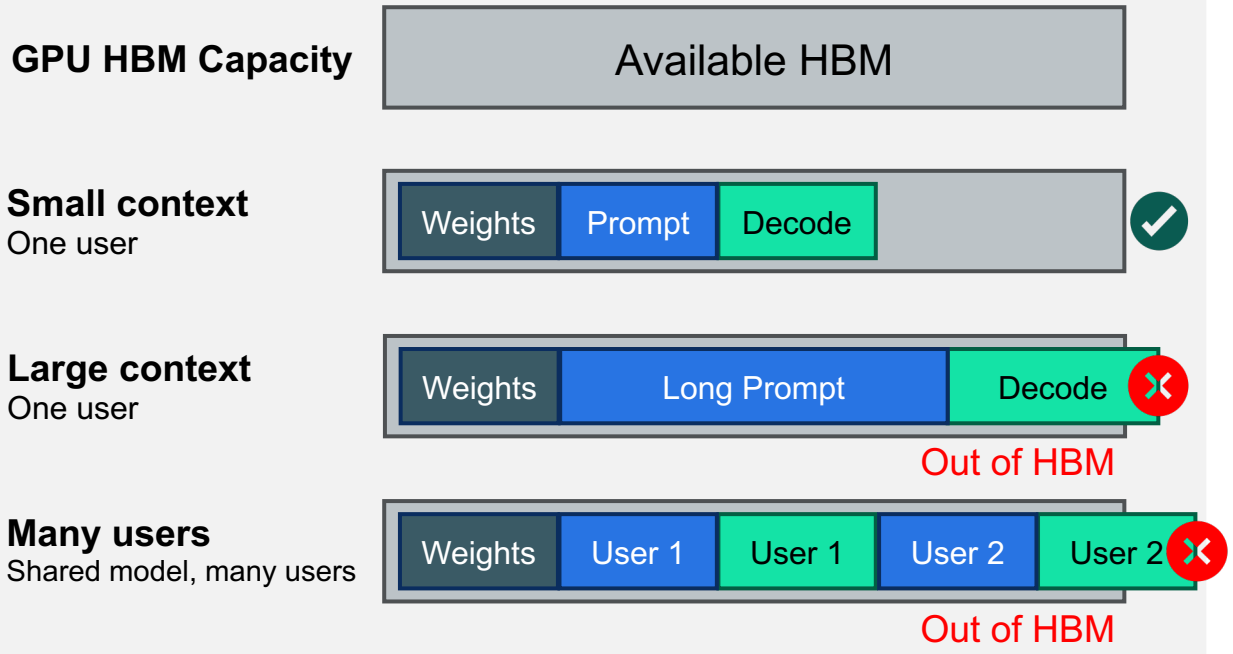
During LLM inference, each token leaves behind keys/values that speed the next token but consume GPU HBM.

How KV Cache Works



Why HBM Overflows

HBM used = model weights + KV Cache + activations
KV cache scales with tokens x layers x concurrent users

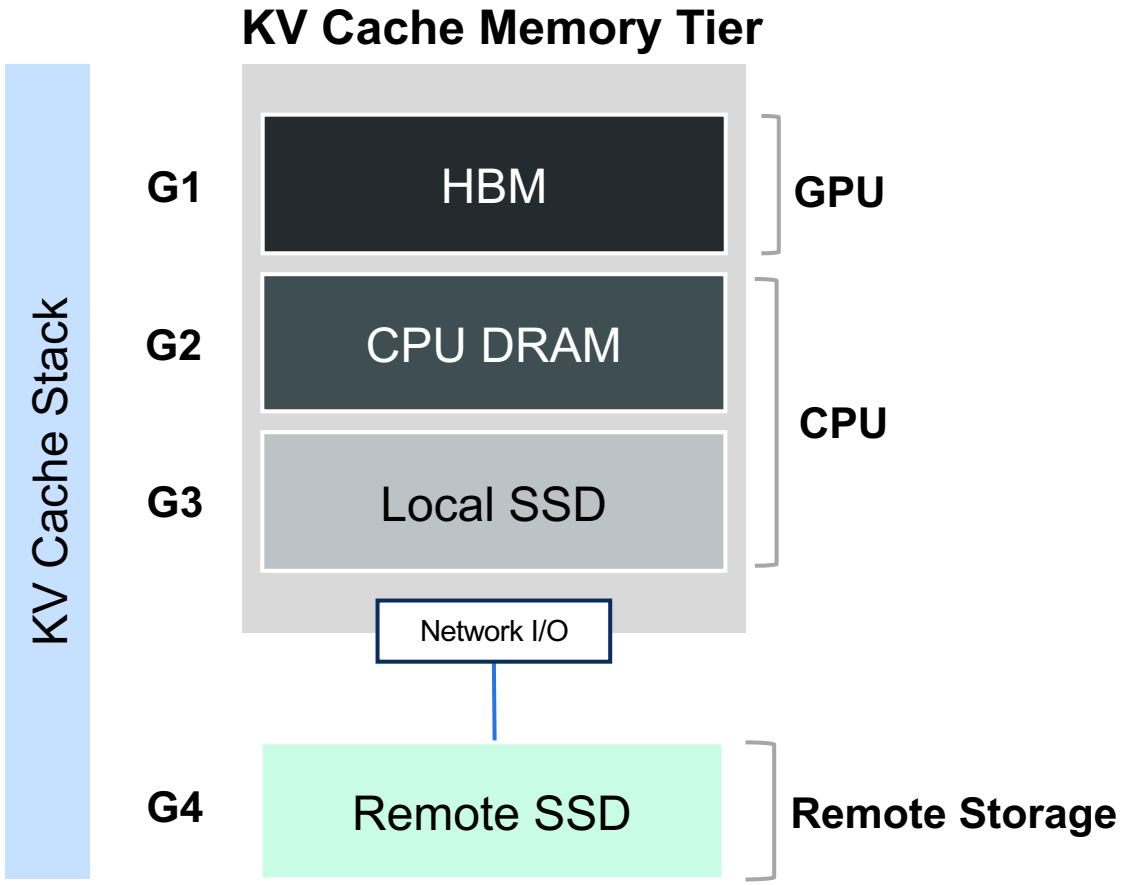


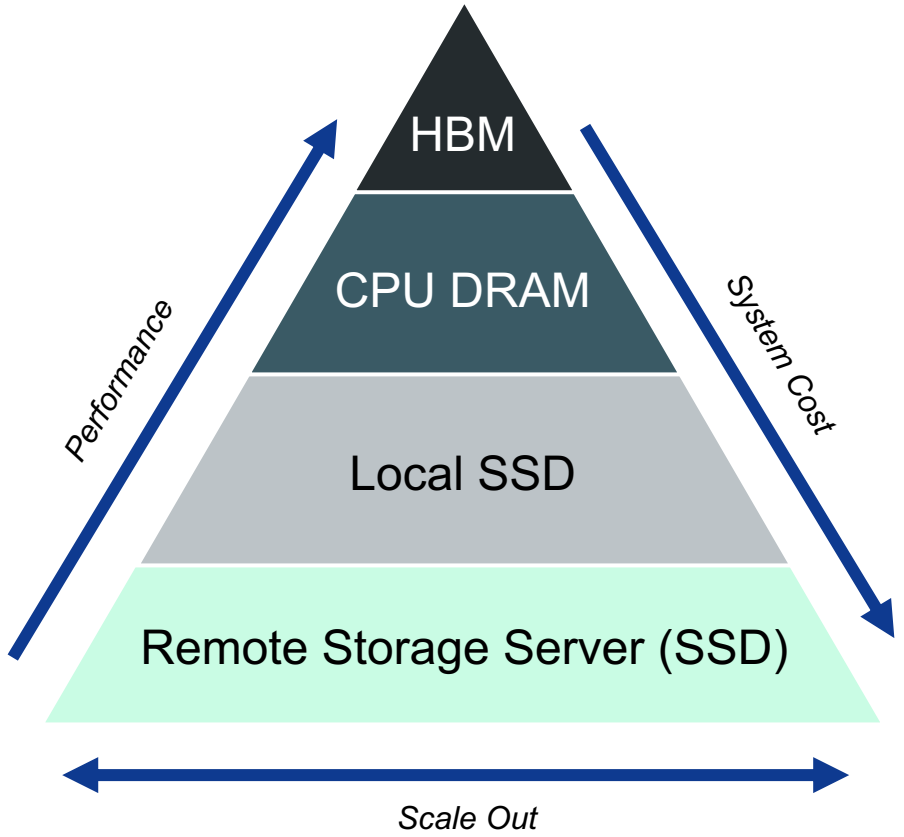
Takeaway: Long contexts + concurrency multiply KV Cache footprint, leaves less HBM, and triggers capacity spill or OOM.

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

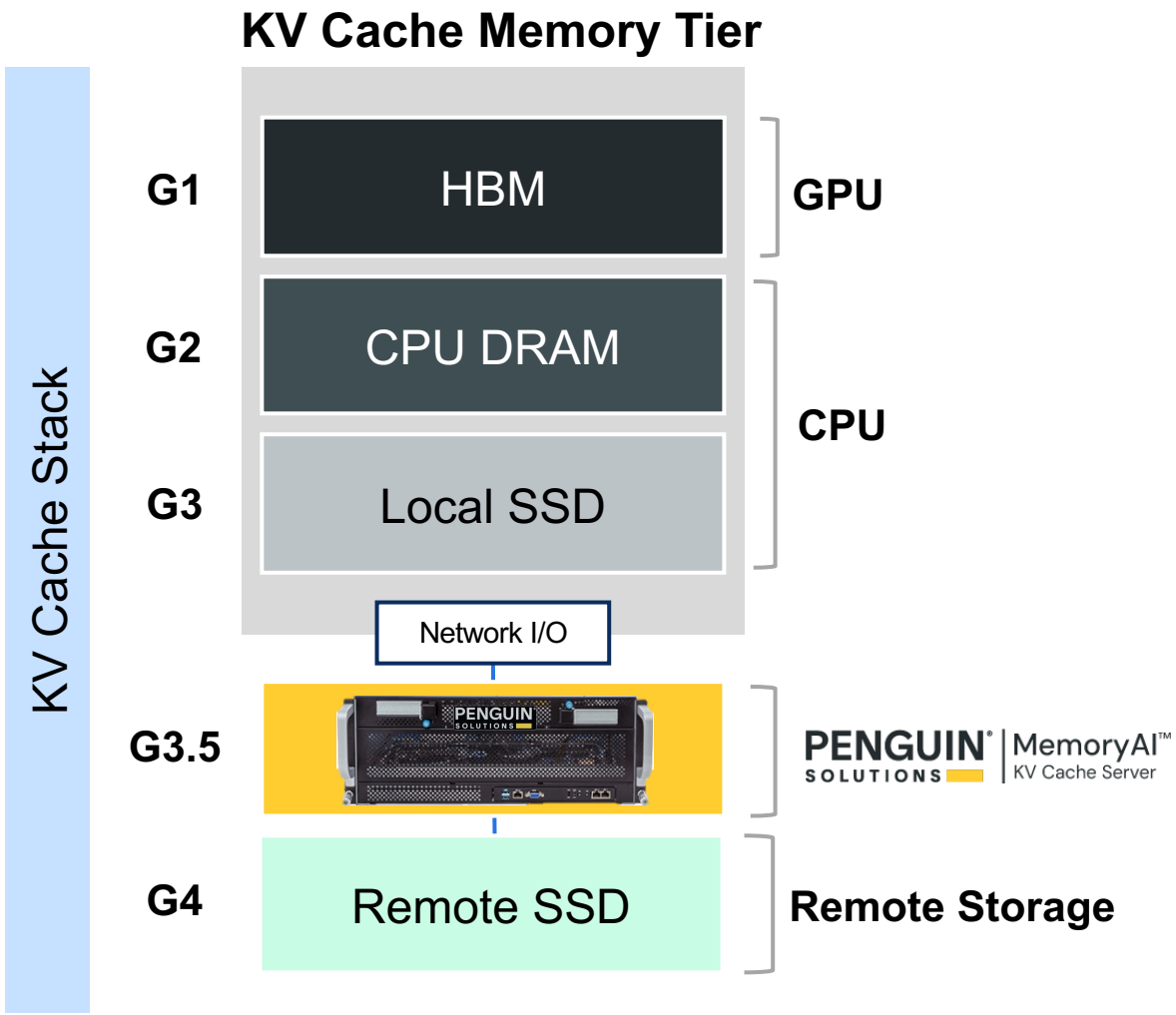
Offloading KV Cache to Other Memory Tiers

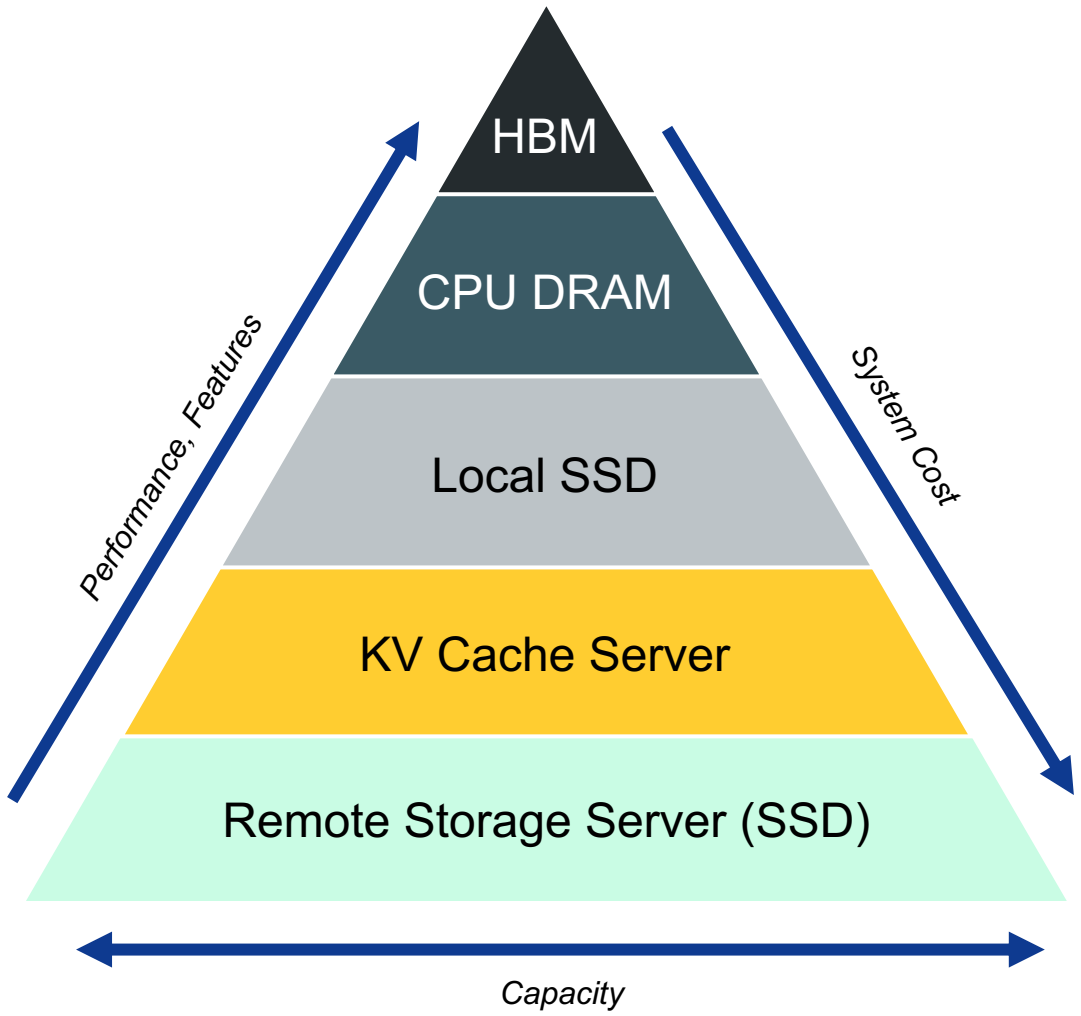




© 2026 Penguin Solutions, Inc. All rights reserved. |

Offloading KV Cache to Other Memory Tiers





© 2026 Penguin Solutions, Inc. All rights reserved. |

Penguin Solutions MemoryAI KV Cache Server

Patent-pending memory appliance for enterprise-scale production inference

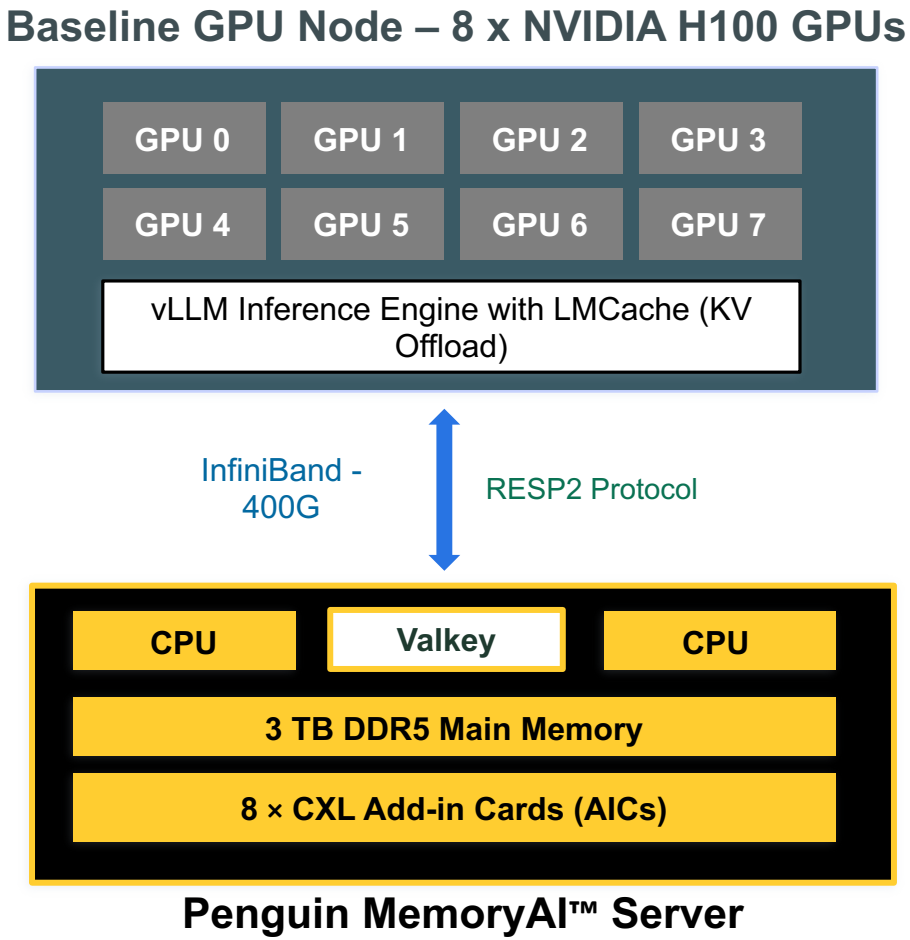


- **Total Memory - 11 TB**
  - Up to 8TB DDR5 CXL memory
  - 3TB (24 x 128GB RDIMM)
- **Processor**
  - Dual AMD EPYC™ 9004 Series (Turin)
- **Networking**
  - 2 x 400 Gb/s Ethernet cards
- **Rack Unit Size - 4U 19”**

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

MemoryAI KV Cache Server – Performance Testing



- **Workload:** Realistic Multi-Document RAG (Retrieval-Augmented Generation)
- **Scale:** 65K average input tokens per request, 128 concurrent users, simulating heavy enterprise congestion.
- **Storage Footprint:** Successfully stored and actively served over 650 GiB of KV cache data directly from the external CXL tier.
- **LLM used**: NVIDIA Llama-3.1-Nemotron-70B-Instruct-HF

PENGUIN SOLUTIONS

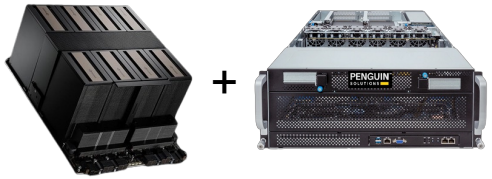
© 2026 Penguin Solutions, Inc. All rights reserved. |

MemoryAI KV Cache Server – Performance Results

GPU Only



GPU + MemoryAI



Performance Metric	vLLM Baseline	Penguin MemoryAI	Improvement
Short Output Throughput (125 tokens)	63.2 tokens/sec	339.3 tokens/sec	5x Higher Throughput
Short Output TTFT	91.5 seconds	1.97 seconds	46x Lower Latency
Long Output Throughput (1024 tokens)	409.9 tokens/sec	1042.8 tokens/sec	2.5x Higher Throughput
Long Output TTFT	8.63 seconds	3.4 seconds	2.5x Lower Latency
Overall Capacity	Limited to GPU HBM	11TB Total Memory (3TB Host + 8TB CXL)	Unprecedented Prefix Reuse Capabilities

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

Scale to 8 GPU Servers: Maintain 2× Throughput

Linear scale-out projection

One 11 TB MemoryAI KV Cache Server extends the measured node-level benefit across 64 GPUs.

MEASURED NODE-LEVEL PROOF

1 GPU server • 8 GPUs



700 GB

KV cache demand
70B-class workload



1 MemoryAI KV Cache Server

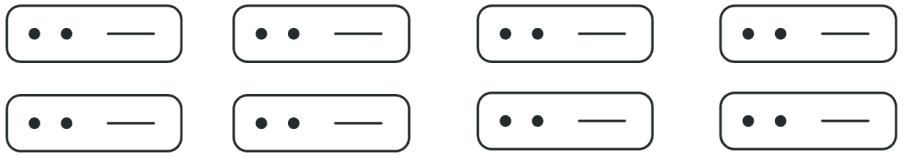
≈2× throughput measured

vs. GPU-only baseline

SCALE

LINEAR CLUSTER SCALE-OUT

8 GPU servers • 64 GPUs



shared KV path



5.6 TB

KV cache demand

11 TB usable

5.4 TB headroom

≈2× throughput maintained

with 8× more GPU servers

1 → 8 GPU servers — one shared KV cache tier preserves the 2× throughput advantage

Measured at one 8-GPU node; 8-node result is a linear capacity projection (8 × 700 GB = 5.6 TB).

PENGUIN SOLUTIONS

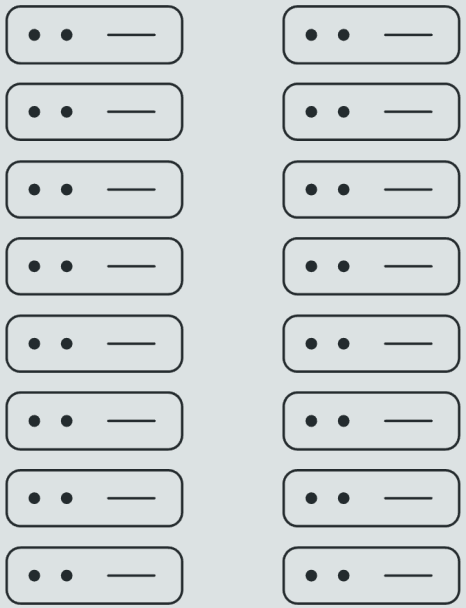
© 2026 Penguin Solutions, Inc. All rights reserved. |

MemoryAI KV Cache Server Cuts GPU Requirements in Half

Shared KV cache removes the memory bottleneck—unlocking equivalent throughput with half the GPU servers.

Without MemoryAI

Tightly coupled compute and memory

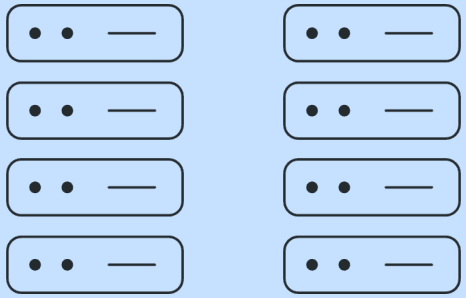


16 x Standalone GPU Servers

≈

With Penguin MemoryAI

Disaggregated architecture



+



8 x Standalone GPU Servers + MemoryAI

50% fewer GPU servers for the same performance target

For qualified, memory-constrained inference workloads.

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

MemoryAI Cuts CapEx by 35%

	Baseline 16 GPU servers	**MemoryAI** 8 GPU servers + KV Cache server
Estimated DGX H100 server-based unit cost	$300,000	$300,000
# of H100 servers **8 x H100 in each node**	16	8
KV Cache server cost	0	$700,000
# of KV Cache servers	0	1
Total CAPEX Spend	$4,800,000	$3,100,000
CAPEX Savings		**-35%**

$1.7M lower CapEx — 8 GPU servers + 1 KV Cache server matches the 16-server baseline
Other benefits include power, cooling and rack space savings

PENGUIN SOLUTIONS

© 2026 Penguin Solutions, Inc. All rights reserved. |

Where This Memory Innovation Series Goes Next

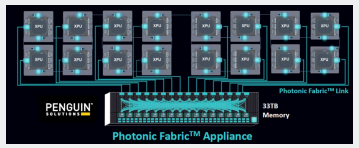
November 23



Unlocking New Value from DDR4 with CXL

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

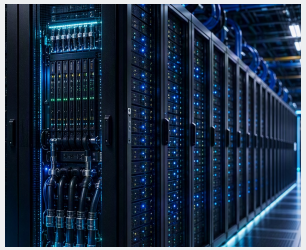
January 13



The Future of Memory: Photonics and Beyond

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

On-Demand



The Memory Advantage in the AI Era

How AI hypergrowth is driving multi-dimensional memory pressure, forcing design teams to innovate memory strategies

© 2026 Penguin Solutions, Inc. All rights reserved. |

Talk to Penguin

Evaluate the Penguin Solutions KV cache server

Discuss sizing and cluster design

Explore OriginAI AI factory reference designs



Kumar Natarajan
Product Manager, MemoryAI



PENGUIN SOLUTIONS



Thank You

PENGUIN SOLUTIONS

The AI Factory Platform Company

www.penguinsolutions.com