signal65

HPE

Maximizing GPU Utilization with HPE Alletra Storage MP X10000

High-performance KV-Cache offload over RDMA turns unused GPU cycles into real output at scale.

The Hidden Bottleneck in AI Inference

GPUs spend **significant time recomputing past tokens**, not generating new ones

KV-cache evictions force **full prefill recomputation**

As context grows, **memory becomes the limiter, not compute**

The Solution: A Petabyte-Scale KV-Cache

By offloading the KV-Cache to HPE Alletra Storage MP X10000, the memory limit is increased from gigabytes to petabytes.

Key Performance Breakthroughs

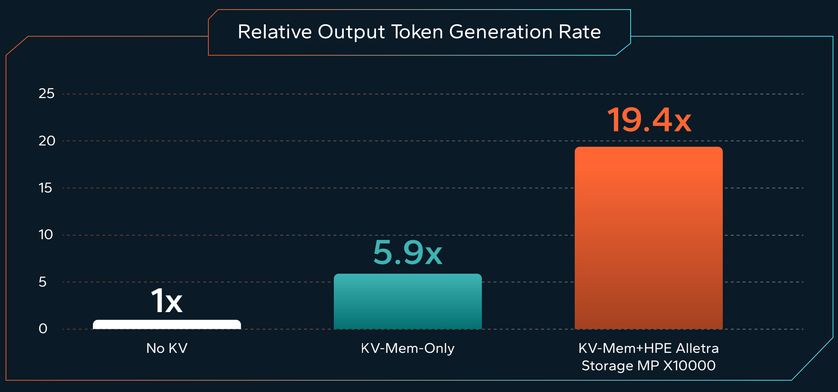
21.5x
Reduction Time to First Token (TTFT) vs no KV-Cache offload

19.4x
Increase in Output Token Generation Rates vs no KV-Cache offload

5.9x
Higher Throughput than system memory-only offloading

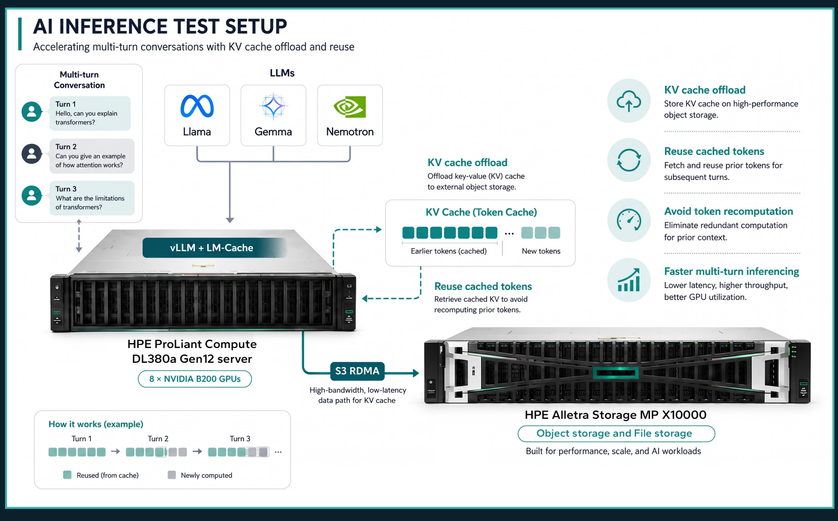
Performance Progression

Each tier of KV-Cache offload unlocks a step-change in throughput.



Test Configuration

Real-world agentic inference at scale - In partnership with Kamiwaza.



Up to 96 concurrent users running multi-turn agentic conversations

INFERENCE SERVER
HPE ProLiant DL380a Gen12 · 8× NVIDIA H200 GPUs · 1.5 TB system RAM

STORAGE TARGET
HPE Alletra Storage MP X10000 · 4× X10250 nodes

WORKLOAD
Kamiwaza KAMI agentic workload · vLLM + LMCache

DATA PATH
RDMA over 100 GbE · NVIDIA cuObject GPU Direct Storage

HPE Alletra Storage MP X10000 unlocks the full potential of GPU-Accelerated AI Inference.

Explore the full Signal65 Insights report >>