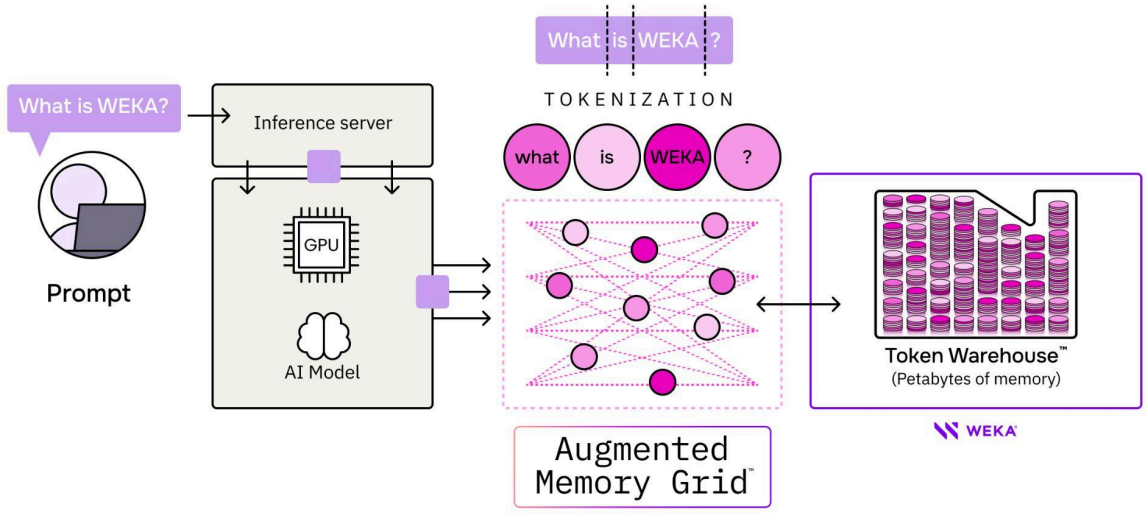
NEURALMESH WITH AUGMENTED MEMORY GRID

# Persistent GPU Memory for AI Inference at Scale

## Extend GPU memory into a token warehouse™ with petabytes of capacity and microsecond latencies for next-gen inference

Large language models (LLMs) and agentic AI workloads are rapidly outgrowing today's GPU memory capacity. Traditional infrastructures rely on volatile HBM or DRAM, which top out in terabytes—forcing GPU recomputation and driving up cost, latency, and energy consumption. WEKA's Augmented Memory Grid™ eliminates this bottleneck by extending GPU memory into a persistent, petabyte-scale token warehouse built on NeuralMesh™.




Augmented Memory Grid uses NVIDIA® Magnum IO® GPUDirect Storage and RDMA networking to move key-value (KV) cache data directly between GPUs and WEKA's token warehouse store at memory speed—bypassing the CPU and system DRAM. The result is a unified, high-speed memory hierarchy that scales linearly across GPU clusters and persists model context across sessions.

Augmented Memory Grid has been validated in production-grade environments to deliver:

- **1000x more KV-cache capacity**—Extends GPU memory far beyond DRAM limits with a persistent NVMe-backed token warehouse, eliminating memory bottlenecks and enabling continuous inference at scale.
- **6× faster time-to-first-token (TTFT)**—By reducing unnecessary prefill computations Augmented Memory Grid accelerates model responsiveness and low latency even as context windows and concurrent sessions grow.
- **4.2x higher token throughput per GPU**—By increasing sustained inference throughput across the prefill-to-decode pipeline, GPUs can process longer contexts, handle more concurrent users, and deliver more output.
- **Lower total infrastructure cost**—Maintaining high cache-hit rates and avoiding redundant prefill computation enables customers to meet the same workload demand with fewer GPUs and lower operational expense.

---

# Features

## Persistent Token Warehouse

NeuralMesh provides a persistent, NVMe-backed token warehouse that holds petabytes of KV-cache data, giving inference systems a durable, allowing GPUs to retrieve persistent KV-cache entries.

Augmented Memory Grid™ sits directly on top of NeuralMesh as the high-speed transport path, streaming KV-cache pages between the GPU and the token warehouse via NVIDIA® GPUDirect® Storage and RDMA. Together, they provide persistent inference memory 1000x beyond DRAM capacity at memory speed, preserving model context across sessions and users, eliminating redundant prefill, reducing latency, and dramatically increasing GPU utilization.

---

## Memory-Speed through RDMA and a GPUDirect Storage Path

Augmented Memory Grid uses RDMA and NVIDIA GPUDirect Storage to establish a direct, zero-copy data path between GPU HBM and NVMe storage.

By bypassing system components like CPU and DRAM, data moves at microsecond latency and memory-class throughput—providing KV-cache capacity and sustained, near-memory performance for large-scale inference.

## Accelerating AI Innovation Through Strategic Partnerships

WEKA continues to build meaningful collaborations across the AI ecosystem to advance Augmented Memory Grid. Through deep integrations with leading technology partners, we're enabling frictionless, memory-speed data movement between accelerated computing platforms and NeuralMesh—powering the next generation of large-scale inference.

---

## Integrated with the NVIDIA Ecosystem

Augmented Memory Grid is built to support and integrate with NVIDIA GPUDirect Storage, NVIDIA Dynamo, and NVIDIA NIXL, including a WEKA-developed open-source NIXL plugin.
This ensures seamless integration across inference pipelines and accelerated frameworks such as TensorRT-LLM.

---

Get the latest updates and insights by visiting our product page