



September 2025

HPC User Forum Fall 2025

Trends in AI for AI – Native Science: Towards the Convergence of HPC and AI

Kevin Tubbs, PhD
Field CTO, HPC & AI

AI is Evolving Fast



And Data Infrastructure Needs to Keep Up



Real-time
Responsiveness



Explosive
Data Growth



Unpredictable
Workloads

A New Approach Is Needed

Performance
That Scales

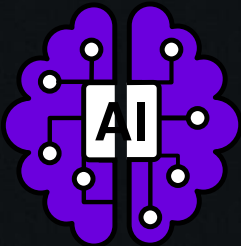
Flexibility Without
Tradeoffs

Efficiency
That Compounds

What do we mean “Radical”?

Reduced Model Training time by 90%

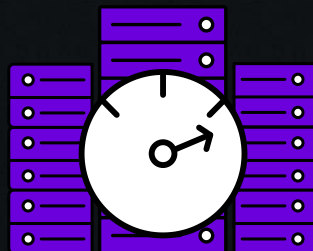
Achieve up to 90% improvement in deep learning epoch time, dramatically speeding up AI model development*



* Source: Proven Business Benefits, Slide 20

Accelerate Storage Performance by 5X

Achieve five times faster storage throughput compared to legacy storage solutions**



** Source: cs-Cetner for AI Safety□Final

Expanded Research Capacity by 6x

Support six times more research projects on the same infrastructure footprint**



** Source: cs-Cetner for AI Safety□Final

Speed Up Time to First Token by 41x

Deliver dramatically faster inference with 41x quicker time to first token for AI applications***



*** Source: WEKA Sets the Standard for AI Inference, slide 12


Business Outcomes



Large Autonomous Vehicle Manufacturer:

Moved from a major All-Flash NAS to WEKA and reduced Autonomous Training Workload from 80 hours to 4 hours...

Genomics England



72x speed up in genome pipeline compared to CPU

75% reduction in storage cost per genome



sphere

4 x 16K video displays

160K sqft LED display

16K x 16K video resolution

402 GB/s of video streaming data

Data Infrastructure Needs to Keep Up



WEKA Powers The Most Demanding Workloads

Generative AI

stability.ai

synthesia

contextual.ai

ElevenLabs

upstage

Midjourney

AI / NeoClouds

CoreWeave

N3XGEN
CLOUD

N

ampz

|||||

smc

IrisEnergy

SIAM.AI
CLOUD

YOTTA

YTL

YTL AI Cloud
YTL GROUP

AI / HPC

DOD
HPC
MODERNIZATION PROGRAM

OAK
RIDGE
National Laboratory

SLAC

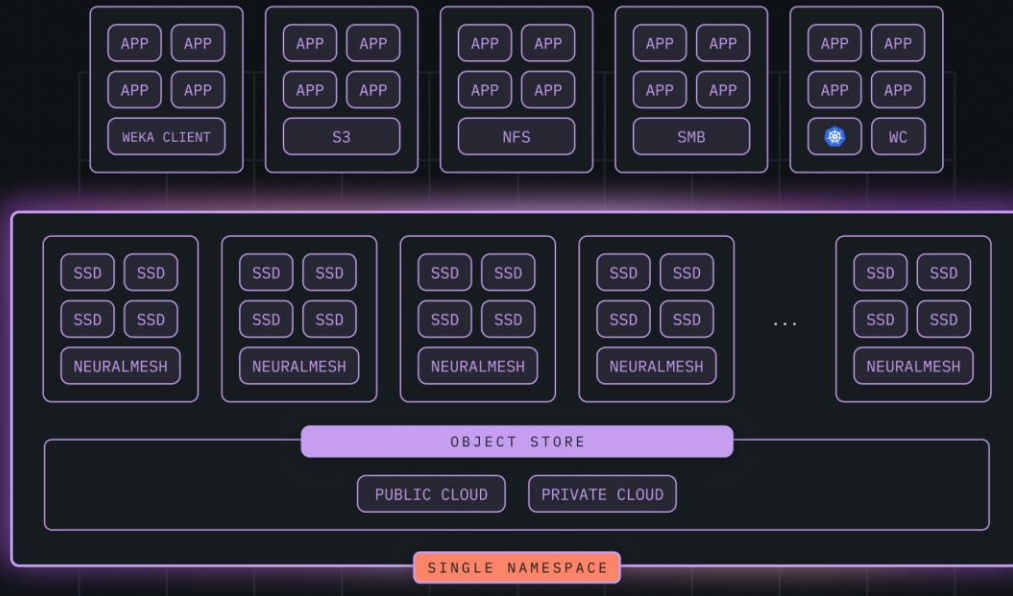
NATIONAL
ACCELERATOR
LABORATORY

NIST

More than 300 Customers Globally
<https://www.weka.io/customers/>

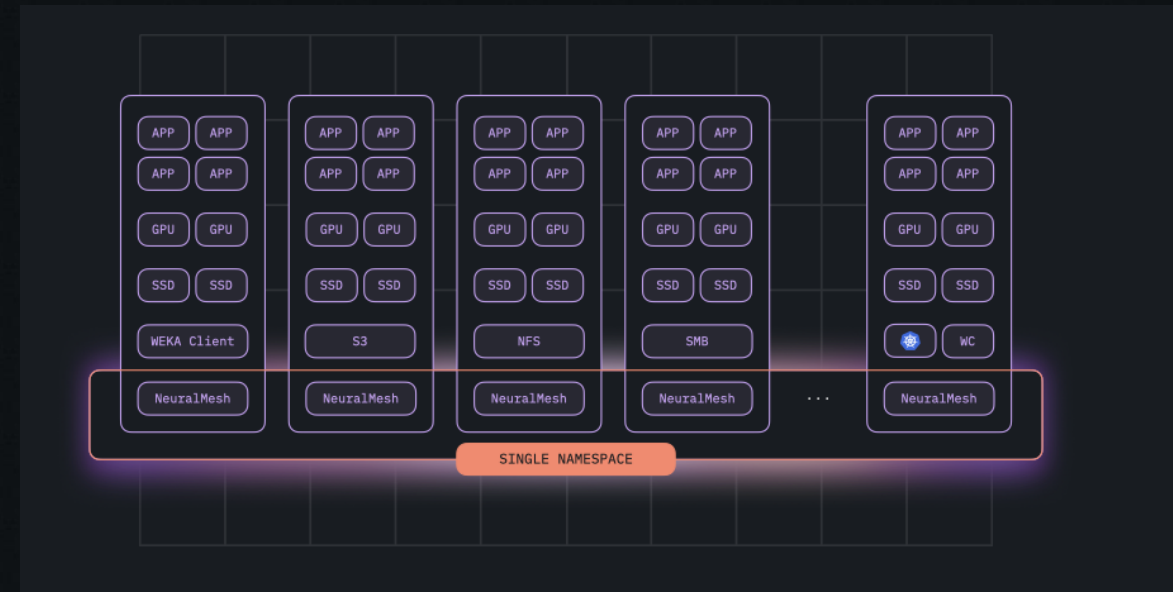
How we do it?

NeuralMesh™ Dedicated



Separate high-perf storage layer. Leverages servers with NVMe
Multi-protocol access (POSIX, S3, NFS, SMB, MC).
Single namespace
Exabyte Scale
Advanced Multi-Tenancy

NeuralMesh™ Axon™



Embedded in compute/GPU nodes.
POSIX Only
Single namespace
Exabyte Scale
Single Tenant

Current view of Inference

Industry reality today:

Rapid adoption of LLMs across **enterprises, research labs, and cloud providers.**

Context windows expanding (4k → 128k+, moving toward 1M) → exponential KV cache growth.

GPU utilization often stuck at 30–55%, not because of lack of GPUs, but because KV cache storage is the bottleneck.

Model deployment patterns:

Smaller fine tuned models (7B–13B) widely used for cost-sensitive apps (chat, customer service).

Mid large models (30B–70B) dominate production coding, RAG, and reasoning workloads.

Agentic systems (multi-turn workflows, tool use, planning chains) are accelerating adoption of persistent KV cache.

At scale (examples in production):

5,000+ GPUs, 20PB+ storage, 700+ concurrent users

Models: Llama-3.1-70B, Qwen-2.5-32B, DeepSeek-16B

Key takeaway:

Whether running a **dozen GPUs or thousands**, the problem is the same:

KV cache outgrows GPU memory, slowing inference in both single-shot and multi-turn agentic workloads.

POV with a Large Cloud Provider

Context:

- Leading CSP exploring **Inference-as-a-Service** for enterprise customers.
- Evaluating multiple frameworks (e.g., vLLM, Modal, etc.) to deliver **scalable, multi-tenant inference**

Challenges observed (similar to Cohere, Stability.ai):

- **High TTFT (Time to First Token)** at long contexts (e.g., 128k tokens → 20s+).
- **GPU underutilization (30–50%)** due to KV cache overwhelming GPU memory.
- **Concurrency bottlenecks** → throughput drops sharply as active users increase.
- **KV cache persistence** missing for multi-turn or agentic workflows (forces recompute).
- **Storage throughput ceiling** when scaling across many GPU nodes.

Key question the CSP asked:

“How do we break past the GPU memory wall and deliver inference at cloud scale without just throwing more GPUs at the problem?”

Augmented Memory Grid™ by WEKA

Open-Source Contribution

WEKA released the **GDS integration plugin**, with LMCache.

Petabytes of Persistent Storage for KV Cache

Unlocks up to **1000× more memory** than fixed DRAM, scaling LLM inference into petabytes.

Optimizations for Inference Infrastructure

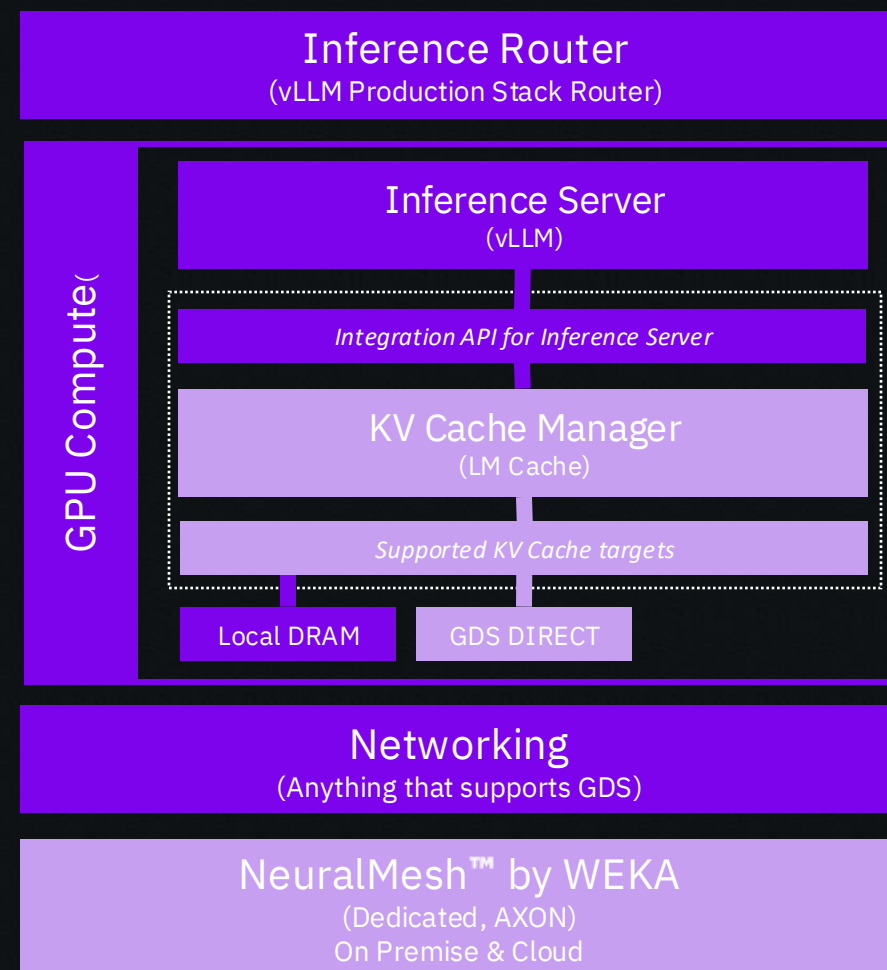
Removes the need to over-provision GPUs when memory is full, balancing **speed, accuracy, and cost**.

Dynamic Resource Reallocations

Offloads KV Cache from HBM DRAM so GPUs focus on critical compute tasks, boosting performance.

WEKA Advantage

While open for all, only AMG & NeuralMesh by WEKA delivers **petabyte-scale capacity, microsecond latency, and enterprise-grade reliability**



POV Environment

GPU Compute

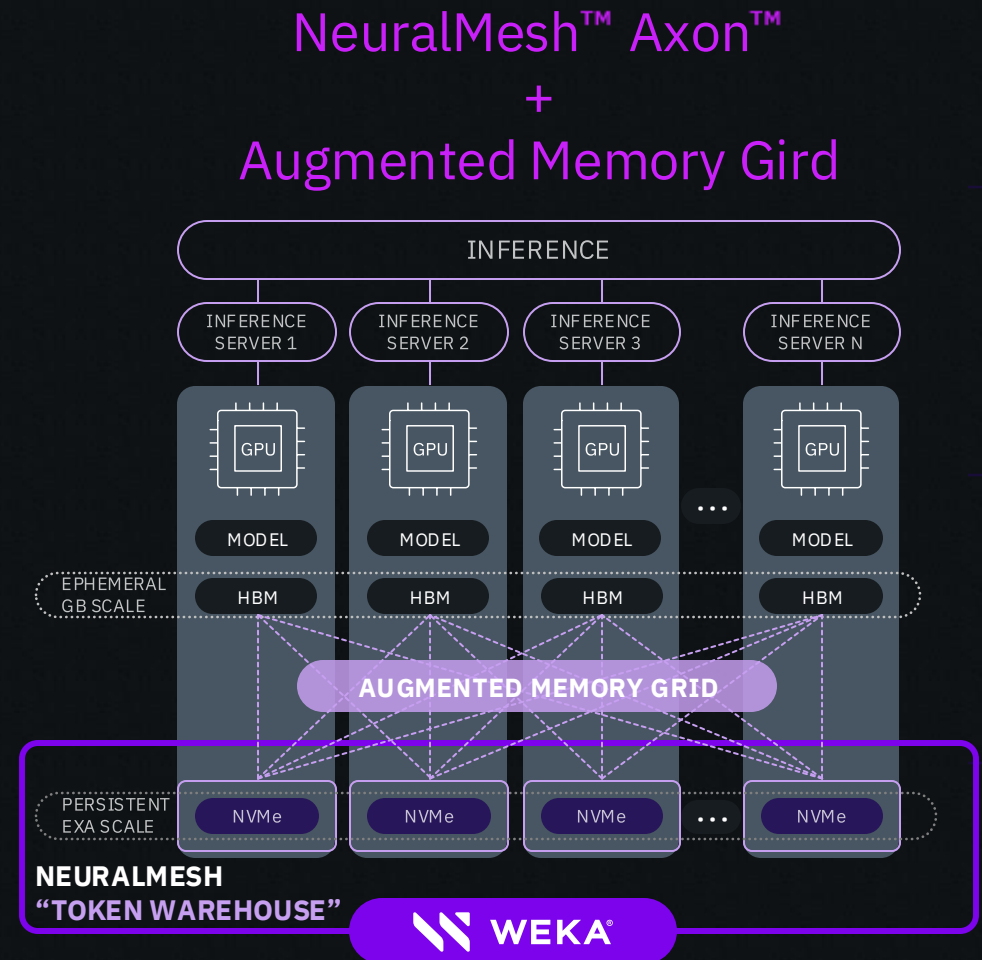
- 8x H100 HGX server
- 8-way NVIDIA H100
- E-W networking is 8x CX7 (400Gb) in IB mode
- 8x NVMe Gen4 7.6 TB drives

NeuralMesh Axon

Each Server:

- (6) Gen4 NVMe's
 - Bandwidth was NVMe limited
- (1) host had (5) NVMe's
- WEKA Cores:
 - (6) Drives
 - (6) Frontend
 - (18) Compute – Over provisioned (did not saturate)

- Memory: ~140 GB memory
- (4) 400 Gb interfaces being used by WEKA (client using all 8)



POV Benchmarks

Many tests were run to evaluate AMG & NeuralMesh Axon with vLLM. For today's session, we will focus on **three key results** that best illustrate the value in production:

Single-shot performance

Show AMG with vLLM performs on par with vanilla vLLM when generating new tokens.

Concurrent requests on a single host

Demonstrate that with multiple users and requests, single-shot performance differences matter less.

Concurrent requests across multiple hosts

Prove that AMG sustains performance at scale, efficiently handling bursty KV-cache workloads.

Graph 1 – Single-Shot Inference Performance (Qwen3-Coder-30B-A3B-Instruct)

Why we ran this test

- To validate whether AMG adds overhead when generating tokens compared to DRAM.
- Focused on **time-to-first-token (TTFT)** and **generation speed** at different context lengths.

What the results show

- AMG performance closely tracks DRAM across all context lengths.
- Baseline (no offload) degrades sharply beyond 32K tokens, while AMG and DRAM remain efficient.

What this proves

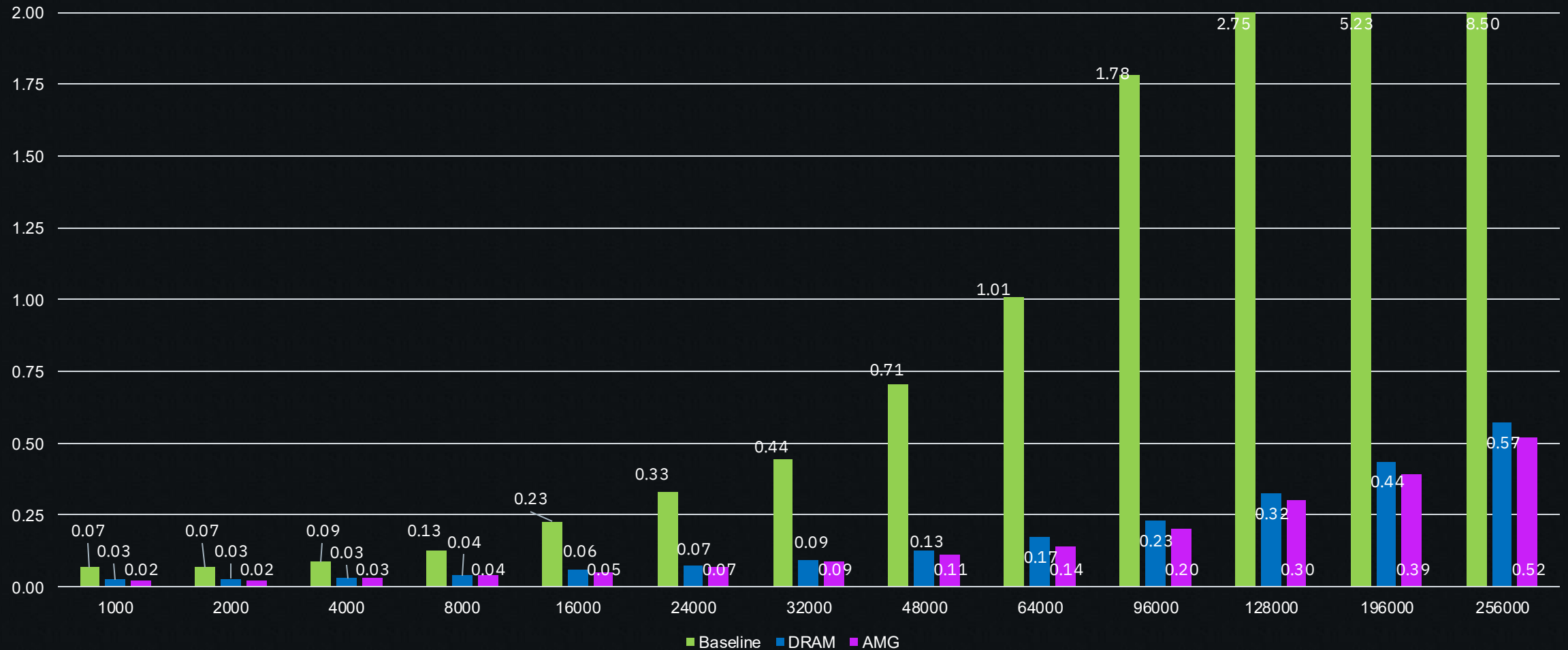
- **AMG delivers near-DRAM performance** for single-shot prompts.
- KV cache offloading to AMG does not create meaningful latency penalties.

What this means in production

- Enterprises can safely adopt AMG to handle **larger contexts** without compromising responsiveness.
- Unlocks **scalability and cost efficiency** while keeping latency competitive with DRAM

Inference performance with single shot prompts

Results for Qwen3-Coder-30B-A3B-Instruct as a medium MOE model



*5 runs, we took the average

Graph 2 – Concurrent Inference Performance (Llama-3.3-70B, 96K Context)

Why we ran this test

- To evaluate AMG performance when **multiple users issue requests simultaneously**.
- Stresses both **time-to-first-token (TTFT)** and **throughput (T/s)** under concurrency.

What the results show

- **AMG and DRAM track very closely** across 1–8 concurrent requests.
- Minor variance between runs, but no meaningful performance degradation with AMG.
- Output throughput scales linearly until saturation, with AMG maintaining parity.

What this proves

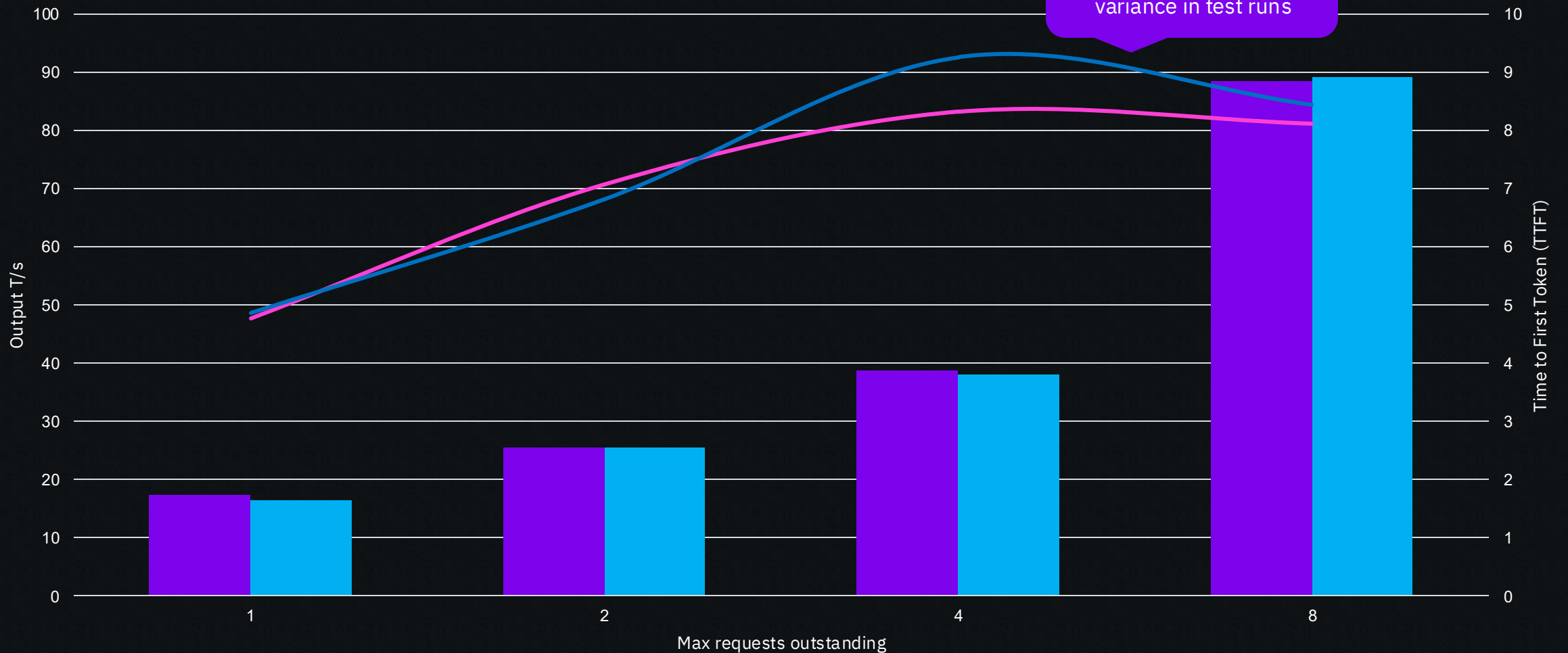
- **Concurrency overhead is negligible** with AMG.
- AMG sustains low-latency response and throughput even under user load.

What this means in production

- AMG can reliably support **multi-user inference workloads** on a single host.
- For AI services with many users (chatbots, copilots, coding assistants), AMG with NeuralMesh will behave **indistinguishably from DRAM** in responsiveness.

Inference performance with concurrent users

DRAM and AMG comparison with Llama-3.3-70B at 96000 context



Both DRAM and AMG are comparable. Some variance in test runs

Graph 3 – Inference Performance with Large Cache (Llama-3.3-70B, 96K Context)

Why we ran this test

- To examine AMG's performance when **working set size grows beyond available DRAM cache capacity**.
- Simulates **real-world multi-host / production scale** where context windows and concurrent requests often exceed GPU memory + DRAM.

What the results show

- At smaller working set sizes, AMG and DRAM remain comparable.
- Once cache capacity is exceeded, **DRAM performance drops sharply** (higher TTFT, reduced throughput).
- AMG sustains **low TTFT and higher throughput**, even under large working sets.

What this proves

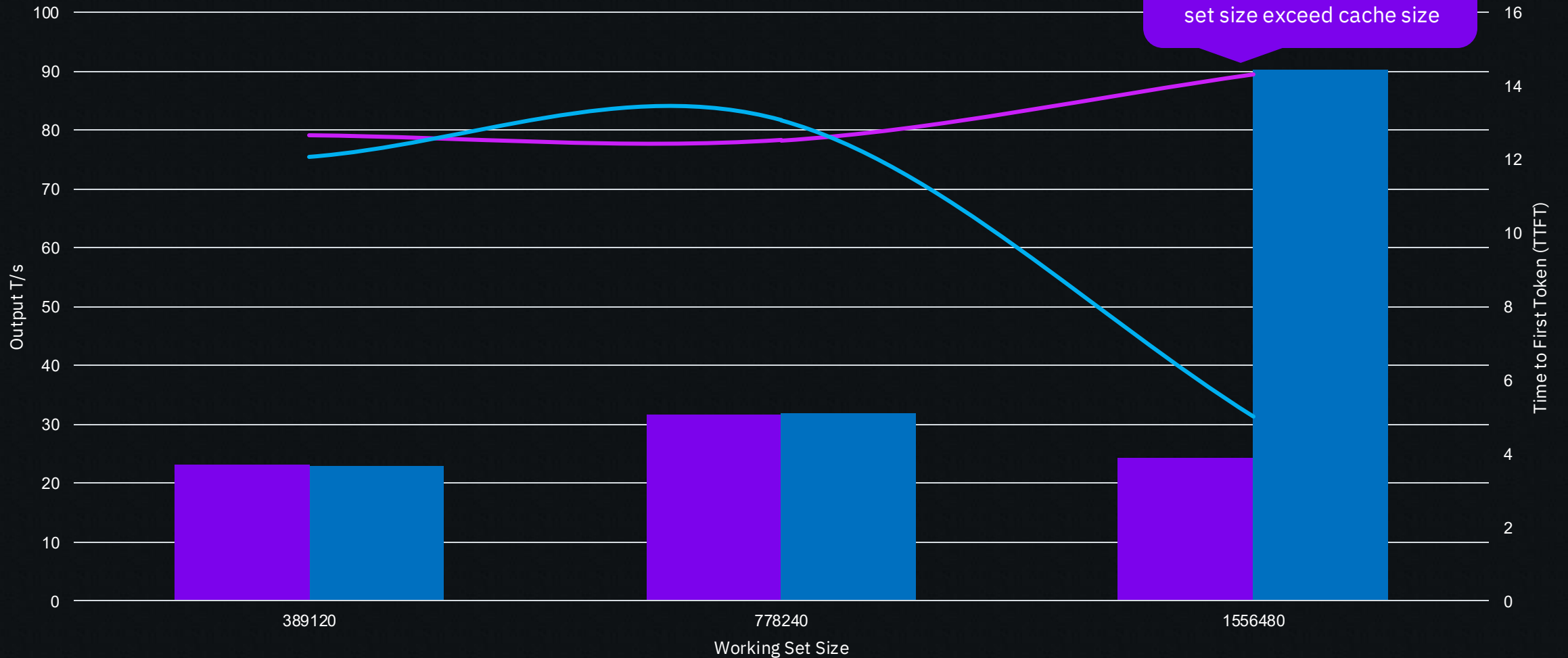
- AMG is not constrained by DRAM limits, making it more resilient under **large-scale, memory-intensive workloads**.
- AMG provides a **clear performance advantage when scaling beyond local cache**.

What this means in production

- Inference services can scale to **longer contexts and more concurrent requests** without collapsing under memory pressure.
- AMG ensures predictable performance and **removes cache-size as a limiting factor**, enabling larger, more complex LLM deployments.

Inference Performance with large cache

DRAM and AMG comparison with Llama-3.3-70B at 96000 context



Significant advantage in TTFT and Output T/s once working set size exceed cache size

POV Takeaways

Seamless integration: AMG works with vLLM out of the box without degrading single-shot inference.

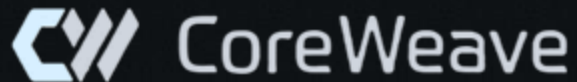
Scales with concurrency: Under multi-user, multi-request workloads, AMG maintains near-DRAM performance.

Production ready: As systems scale to many GPU hosts, AMG absorbs bursty KV-cache loads, ensuring stability and predictable throughput.

Future proof: Larger context sizes and bigger working sets benefit most from AMG's Token Warehouse, providing long-term performance headroom.

AMG enables high-performance, cost-efficient KV-cache management essential for real-world AI inference at scale.

NeuralMesh Axon **Customer Outcomes**



- GPU utilization improvements from ~30% to over 90%
- Rack space reduction of ~30–50%
- Up to 20x faster time-to-first-token (TTFT) for inference-heavy workloads (NeuralMesh Axon + AMG)
- Infrastructure and operational cost reductions of up to 33%



WEKA[®]