

What Top-Performing AI Clouds Built Differently with WEKA

What 4.2x more tokens, 10x faster checkpointing, and a \$10B financing round have in common (plus the decisions that made them possible).

Table of Contents

Introduction	3
The problem we kept finding inside the best AI clusters	4
CoreWeave: 4.2x more tokens from the same 64 GPUs	5
Oracle Cloud Infrastructure: 10x concurrent users, 7x more tokens — same hardware	6
Cohere: when the infrastructure change shows up in operations	7
Firmus: when infrastructure efficiency becomes a financing thesis	8
What CoreWeave, OCI, Cohere, and Firmus have in common	8
Where we think this goes next	10
The DRAM ceiling will become visible for more AI clouds	10
The re-architecture trap will claim its first high-profile casualties	10
Sovereignty will split the market	10
Token economics will become a public metric	11

Introduction

WEKA's infrastructure runs inside some of the world's leading AI clouds and hyperscalers: CoreWeave, Cohere, Firmus, and Oracle Cloud Infrastructure. Over the past two years, we've worked alongside these teams, managing, testing, and benchmarking what happens when AI inference workloads hit real scale: context windows stretch to 100,000 tokens and beyond, concurrent users multiply, and multi-agent workflows replace single-shot queries. And we keep finding the same thing: the infrastructure decision that looked fine at 10 petabytes starts failing at 50. The GPU utilization number stays high. The token economics quietly deteriorate.

This isn't a best-practices guide. It's a brief based on direct feedback, anecdotes, and learnings from the AI clouds running these workloads in production; including what they built, what they measured, and what changed. We've added what we think comes next, because the pattern we've seen across these deployments points somewhere specific.

The pattern we kept finding inside the best AI clusters

Every large language model inference request generates KV cache — the record of key-value pairs computed during a session. Retaining it means the model doesn't have to re-compute context it's already seen. The problem is that KV cache grows with context length, and GPU high-bandwidth memory (HBM) runs out. Even the newest Blackwell GPUs don't close that gap: a B200 carries 192GB of HBM, a B300 288GB — up from 80GB on the H100. Models are also getting more KV-cache efficient. But neither trend will keep pace with what production inference demands: 100,000-token context windows, thousands of concurrent sessions, and agentic workflows that keep context alive far longer than a single query. Run enough of those sessions at once, and even Blackwell's added headroom disappears.

When HBM fills up, the cache evicts. The system re-prefills from scratch. Latency spikes. Throughput drops. The GPU utilization number stays high...because the GPU is busy. It's just that it's busy redoing work it has already done. Every eviction is a tax: on GPU cycles, on latency, on user experience, and on the cost of every token served. For long-context and agentic workloads, where inputs routinely run to 100,000 tokens or more, that tax isn't a rounding error. It's the real ceiling on what the infrastructure can do.

DRAM was not built to serve as a KV-cache tier. It's designed for high-speed, sequential, transactional work, not overflow storage. DRAM in a standard GPU tops out around 8–9 TiB, and at \$5–10 per GB, it's too expensive to scale to petabyte-class context volumes. Worse, DRAM is strictly node-local. It can't share KV state across the cluster, so each inference instance becomes an isolated cache island. A request lands cold on any instance that hasn't seen that context, even if another instance cached it seconds ago, and every new autoscaled instance starts empty, paying full prefill tax until it warms up. Offloading overflow duty to WEKA's NVMe tier removes that ceiling and frees DRAM to do the job it was designed for.

The failure cliff is invisible until you hit it

Below DRAM saturation, systems look fine. Above it, cache misses compound, TTFT climbs, and throughput falls all while GPU utilization stays high.

- Cache evictions force re-prefill and GPU cycles spent redoing work already done
- TTFT becomes unpredictable under load, making SLO commitments harder to keep
- Adding DRAM delays the ceiling, it doesn't remove it

CoreWeave: 4.2x more tokens from the same 64 GPUs

CoreWeave, ranked #1 on [SemiAnalysis's ClusterMAX AI cloud rating](#), was running production inference workloads on 64 H100 GPUs when we measured what was actually happening to token economics as context length increased. The results quantified the cost of the memory wall precisely.

With DRAM-only caching, once working sets exceeded DRAM capacity, throughput dropped to 55 tokens per second and TTFT climbed to 10.9 seconds. With **WEKA Augmented Memory Grid**, which extends KV cache storage 1,000x beyond HBM and DRAM using NVMe over RDMA via NVIDIA GPUDirect Storage, throughput reached 231 tokens per second and TTFT held at 1.8 seconds. Same 64 GPUs. Same cluster. 4.2x more usable output and 6x faster response. 🙌 [Read the full methodology right here.](#)

4.2x

More token output per GPU —
CoreWeave H100 production

6x

Faster time-to-first-token — same
CoreWeave cluster

231

Tokens/sec with AMG vs. 55
DRAM-only — CoreWeave

NeuralMesh™ Axon™ — which fuses WEKA's distributed file system directly into GPU servers using the NVMe drives already on compute nodes — is what makes the data path possible. No separate storage infrastructure. No additional rack space or power draw. The storage is already in the cluster. NeuralMesh Axon activates it. On CoreWeave's production infrastructure, NeuralMesh Axon delivers more than 30 GB/s read and 1 million IOPs to a single GPU server at microsecond latency.

weka.io/product/neuralmesh-axon

Partner solutions like WEKA's NeuralMesh Axon deployed with CoreWeave provide a critical foundation for accelerated inferencing while enabling next-generation AI services with exceptional performance and cost efficiency. Embedding ultra-low latency NVMe storage close to the GPUs unlocks more bandwidth and extends available on-GPU memory.

Marc Hamilton, VP, Solutions Architecture & Engineering, NVIDIA

Oracle Cloud Infrastructure: 10x concurrent users, 7x more tokens, same hardware

In May 2026, Oracle published production-scale benchmark results on its [AI & Data Science blog](#) validating WEKA Augmented Memory Grid on OCI bare-metal H100 infrastructure — nine nodes, 72 GPUs, 100,000-token context windows, thousands of concurrent users. The test was designed specifically to expose how systems behave when pushed past steady-state and the conditions that reveal the real production ceiling.

DRAM-only configurations hit a hard ceiling at approximately 600 concurrent users. Augmented Memory Grid scaled past 5,000. In the same one-hour, 2,400-user test, DRAM-only served 700 million tokens. Augmented Memory Grid served 5 billion. Token throughput: under 200,000 tokens per second for DRAM versus approximately 2 million tokens per second with Augmented Memory Grid — a 10x gain. The cache pool: 8.64 TiB of DRAM versus 287 TiB of usable NVMe.

The hardware was identical. What changed was the architecture.

10x

More concurrent users — OCI H100,
100K-token context

7x

More tokens served — same 1-hour
test window on OCI

5B

Tokens served with AMG vs. 700M
DRAM-only — OCI

Oracle's team noted the pattern clearly in their published results: the performance gap between DRAM-only caching and Augmented Memory Grid "does not appear gradually. It appears at the point where DRAM cache saturates." That's the operational reality: below the ceiling, the difference can be invisible. Above it, one configuration keeps scaling and one doesn't. Oracle has positioned the results as a validated reference architecture for long-context inference on OCI, and NeuralMesh with Augmented Memory Grid is now available on the Oracle Cloud Marketplace — with OCI as WEKA's exclusive cloud launch partner.

NeuralMesh with Augmented Memory Grid running on OCI brings orders of magnitude more tokens to customers in an extremely cost-efficient way.

Liran Zvibel, CEO, WEKA

Cohere: when the infrastructure change shows up in operations

The CoreWeave and OCI results measure efficiency at the infrastructure layer — tokens per GPU, users per cluster. Cohere's experience with **NeuralMesh Axon** shows what the same architecture change looks like in day-to-day AI cloud operations. Inference deployments that previously took five minutes complete in 15 seconds. Checkpointing — the operation that saves model state during training and protects against costly restarts — accelerated by 10x. Both outcomes come from the same structural change: WEKA's file system running directly on the NVMe already inside the GPU servers, removing the data path overhead of a separate storage tier.

Embedding WEKA's NeuralMesh Axon into our GPU servers enabled us to maximize utilization and accelerate every step of our AI pipelines. The performance gains have been game-changing: Inference deployments that used to take five minutes can occur in 15 seconds, with 10 times faster checkpointing.

Autumn Moulder, VP Engineering, Cohere

Firmus: when infrastructure efficiency becomes a financing thesis

[Firmus](#) built its entire AI cloud on the premise that energy efficiency is a business model, not an engineering goal. In production, WEKA delivered 6.5x higher tokens per second and 4x faster TTFT at scale. Firmus then raised \$10 billion in financing — anchored, in material part, by its ability to demonstrate infrastructure economics that competing neoclouds couldn't match. The efficiency story became the investor story. The investor story became the growth story.

The Firmus outcome illustrates something we've seen develop across multiple AI cloud deployments: infrastructure efficiency compounds. The AI cloud that can demonstrate tokens-per-watt, sustained performance under multi-tenant load, and SLO stability at scale isn't just more profitable on a per-GPU basis — it attracts different capital, different customers, and different contract structures. Efficiency that starts as a technical decision becomes a competitive moat.

What CoreWeave, OCI, Cohere, and Firmus have in common

The infrastructure architecture in each case is the same: extend KV cache storage beyond local GPU memory into a high-performance token warehouse — persistent, petabyte-scale, accessible across the cluster at memory-class speed. Decoupling cache from GPU-coupled DRAM removes session stickiness, enables safe GPU memory oversubscription, and increases tenant density without adding hardware. The workloads are different. The business models are different. The architecture decision is the same.

There's also a supply chain dimension worth noting. Because **NeuralMesh Axon** uses NVMe drives already in GPU compute nodes, the storage capacity upgrade doesn't require a separate hardware procurement cycle. For neoclouds building fast — or responding to demand faster than purpose-built storage can arrive — WEKA deploys on existing bare metal or Kubernetes infrastructure from any hardware vendor. **WEKApod™** provides a pre-integrated appliance path for providers who need turnkey deployment with predictable lead times, particularly relevant for EMEA neoclouds navigating sovereign AI mandates and long procurement cycles.

The architecture decision that runs through all four

Every deployment extended KV cache beyond GPU-coupled DRAM into a persistent, cluster-wide token warehouse. The workloads differed. The decision didn't.

- Decoupled KV cache allows safe GPU memory oversubscription and higher tenant density
- Persistent cache eliminates session stickiness: any host can serve any session with cache hits intact
- Storage activates inside existing GPU servers requiring no separate infrastructure, no separate procurement

Where we think this goes next

Based on what we've seen across these deployments, here's where AI cloud infrastructure is heading — and why the decisions made now compound in one direction or the other.

The DRAM ceiling will become visible for more AI clouds

Most AI clouds today haven't hit the DRAM saturation point yet. They will. Context windows are expanding. Agentic workloads requiring multi-turn, multi-step, persistent-context are moving from pilot to production. Each step multiplies the KV cache pressure: a reasoning model needs roughly 10x more context memory than a chat model; an agentic workflow needs 10x more than a reasoning model. The AI clouds building infrastructure that handles 100,000-token context windows today will have a meaningful head start when 1-million-token contexts become standard.

The re-architecture trap will claim its first high-profile casualties

We've already seen one leading AI cloud that is now operating more than 140 petabytes go through a full storage re-architecture after an early infrastructure choice hit its performance ceiling at scale. The migration cost exceeded the original storage investment. That pattern will repeat, and it will become more visible as more AI clouds approach the 50–100 petabyte threshold. The providers who locked in the right architecture at day one won't face that event. The ones who optimized for initial cost will.

Sovereignty will split the market

EMEA enterprise buyers are moving from preferring data locality to requiring it. National AI frameworks, government procurement requirements, and sovereign cloud mandates are becoming selection criteria before technical evaluation. AI cloud providers who can demonstrate jurisdictional control of AI workloads — where model weights are stored, where KV cache lives, which legal systems govern access — will be shortlisted for contracts that others won't even see. This is a structural market split, not a temporary regulatory cycle.

Token economics will become a public metric

Tokens per watt, tokens per dollar, and concurrency density per GPU will move from internal benchmarks to published differentiators — the way PUE moved from engineering metric to sales asset over the past decade in hyperscale data centers. The AI clouds that measure these numbers now, build infrastructure that optimizes for them, and can document the methodology will have a credibility advantage when buyers start asking. The ones that don't will be forced into the conversation reactively, and from a weaker position.

Want to see these results in your environment?

Talk to WEKA about running a production benchmark on your infrastructure visit weka.io/contact.

Benchmark data reflects production testing environments. CoreWeave results: 64-GPU H100 cluster. OCI results: 72-GPU bare-metal H100 cluster, 100K-token context, published Oracle AI & Data Science blog, May 2026. Firmus results: production environment. Performance varies by workload configuration.