WEKA

AI INFRASTRUCTURE ECONOMICS:

Speed Your Time to First Token (and Save!)

WEKA Delivers Dramatically Faster Time to First Token & Optimized Token Processing.

As model sizes and context window sizes expand, the demand for efficient and large-scale memory is increasing beyond the fixed amounts available today.

Optimizing token generation is the key to AI adoption, growth, and organizational success.

Don't let hitting the memory bandwidth wall slow down your time to first token. ↓

Don't let hitting the memory wall:

- Limit the capabilities of AI applications
- Erode engagement rates for end users due to slow prompt results
- Require unnecessarily expensive GPUs and infrastructure

MEMORY MYTHS

Inferencing is limited to the memory in a GPU

More memory requirements = more GPUs

→ **If these are true, scaling becomes a nightmare and wildly expensive.**

WEKA Breaks Through Memory Myths.

"Just as breaking the sound barrier unlocked new frontiers in aerospace innovation, WEKA's Augmented Memory Grid is shattering the AI memory barrier, expanding GPU memory and optimizing token efficiency across the NVIDIA AI Data Platform."

–Nilesh Patel, Chief Product Officer @ WEKA

WEKA's Augmented Memory Grid™ blasts through the memory wall optimizing inference infrastructure and supercharging token efficiency.

WEKA extends memory hierarchy to cache prefixes or key value (KV) pairs—around **three orders of magnitude (1000x)** more than today's fixed DRAM increments of single terabytes.

HOW

WEKA Supercharges Token Efficiency

Lower Latency

Latency is the key to optimizing token generation.

WEKA's modern GPU-optimized architecture, featuring NVMe SSD acceleration and high-speed networking, enables token processing at **microsecond latencies**.

→ And no one, **seriously no one**, does ultra-low latency better than WEKA.

Prove it.

We put the WEKA® Data Platform to the test, leveraging NVIDIA Magnum IO™ GPUDirect Storage (GDS) and a high-performance 8-node WEKApod.

Our tests demonstrated a staggering **41x reduction in time to first token prefill time on LLaMA3.1-70B** for 105K tokens, **dropping from 23.97 seconds to just 0.58 seconds**.

41X reduction in AI inference latency

"By extending GPU memory and maximizing utilization across our Shakti Supercloud fleet, WEKA will help us deliver improved AI performance, faster inference, and better cost efficiency to our customers."

–Sunil Gupta, co-founder, Managing Director & CEO @ Yotta Data Services, an NVIDIA Cloud Partner

Fewer GPUs

The secret to breaking through the wall of high-bandwidth memory limitations is to decouple GPU compute from memory.

For the very first time in the market, WEKA has made it possible to extend memory for inference workloads outside of the GPU.

→ No. It's not a hole in the space time continuum.

It's WEKA's **Augmented Memory Grid**.

Explain how WEKA does it.

WEKA seamlessly extends GPU HBM beyond its current limitations. **That's right.** We found ultra-low latency storage that functions like memory.

We treat microsecond latency NVMe storage and the GPUDirect RDMA as an adjacent tier of memory.

WEKA saves 24 seconds of compute time per request.

WEKA's Augmented Memory Grid

makes it possible for GPUs to deliver **200% more capacity** than they are currently delivering, so you can **run your workloads with 30% fewer GPUs**.

Reduced Costs

Every millisecond saved in token inference translates to reduced infrastructure overhead and efficiency gains.

WEKA enables optimized inference workloads so businesses can scale their AI applications cost-effectively while maintaining high levels of efficiency and accuracy.

Show me the money.

When you need fewer GPUs for the same workload, you can **cut your AI inference costs by 30%**.

WEKA's Augmented Memory Grid

inferencing clusters can achieve higher cluster-wide output tokens, **lowering the cost of token throughput by up to 24% for the entire inference system**.

As enterprises scale their AI deployments, these optimizations make LLMs and LRMs much more profitable, ensuring cost efficiency without sacrificing accuracy or performance.

WEKA Focuses On Less.
Lower Latency. Fewer GPUs. Less Cost.

So Your AI Projects Can Do More.

→ **MORE** Tokens Faster.
→ **MORE** Customers Satisfied.
→ **MORE** Revenue Realized.

Do More With WEKA

weka.io

WEKA