WEKA®

Persistent Agent Swarms

Use-cases & Inference Implications

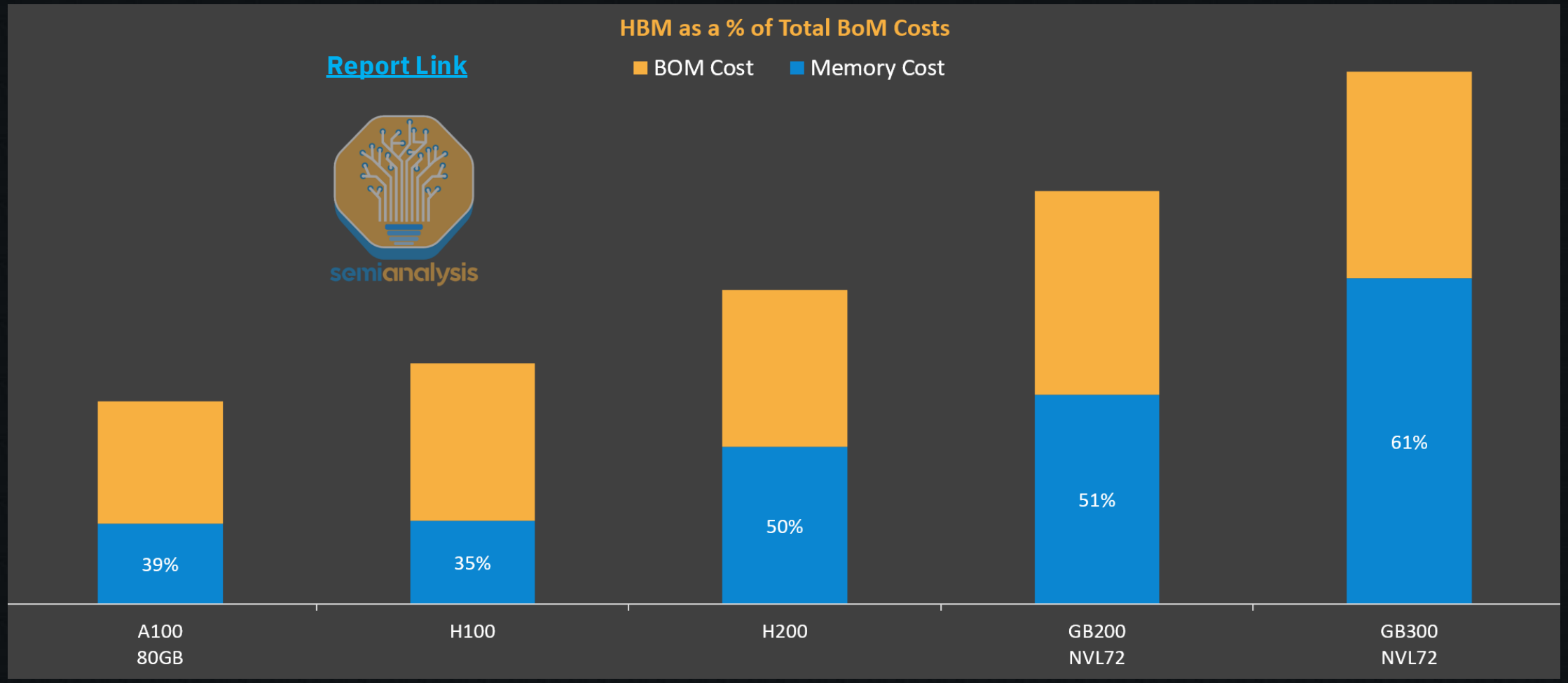
Val Bercovici

Chief AI Officer

AI Future = Fusion of Storage + Memory

WEKA®

GPU Server Bill of Materials

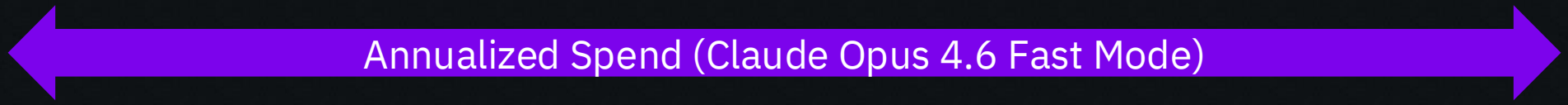


March 2026 (GTC)

Inference Inflection Arrives: 10,000x ChatGPT Compute

NVIDIA

Remember Tokenmaxxing?



$0.00 $250,000 $743,000 $5,000,000

Higher Order Co

WEKA®

THE WALL STREET JOURNAL.

SUBSCRIBE SIGN IN

NEWSLETTERS • WSJ TECHNOLOGY

Tokenmaxxing Maxes Out

Plus, AI topples an 80-year-old math problem, the Brockmans sit for an interview and physical AI gets its due

By Bradley Olson Follow

May 31, 2026 9:30 am ET

Aa Listen (1 min)

CNBC WATCH LIVE

TECH

Tokens or humans? The new corporate trade-off

PUBLISHED FRI, MAY 29 2026 2:24 PM EDT
UPDATED SAT, MAY 30 2026 8:16 AM EDT

Deirdre Bosa @DEE_BOSA | Jasmine Wu @JASWU_

WATCH LIVE

KEY POINTS

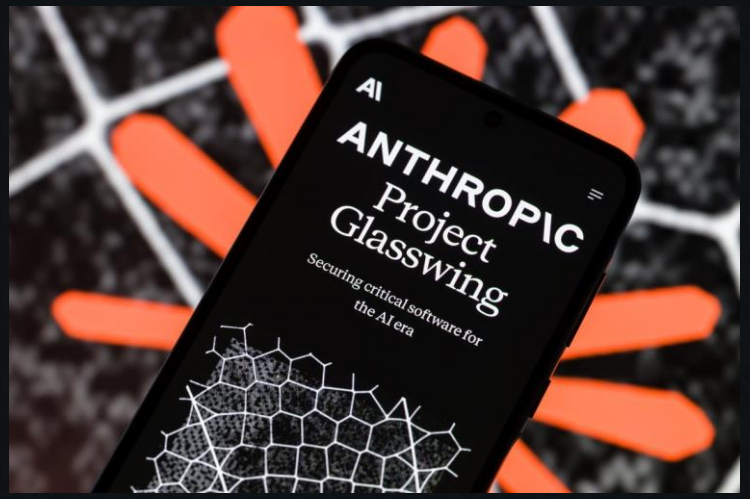
- AI is turning out to be more expensive than enterprises expected, and CFOs are now trading future headcount for tokens.
- Roughly 95% of enterprise AI still runs on the priciest frontier models even for simple tasks, and routing that work to cheaper tiers can deliver hefty savings.
- The AI trade assumes demand stays enormous and indifferent to cost, but Fortune 500 buyers are becoming more sensitive to price.


End of AI Innocence

Post Mythos

AI/ML Timeline

AlexNet GPT-2 ChatGPT DeepSeek R1 **Mythos/GPT 5.5** **DeepSeek v4**



deepseek v4

WEKA®

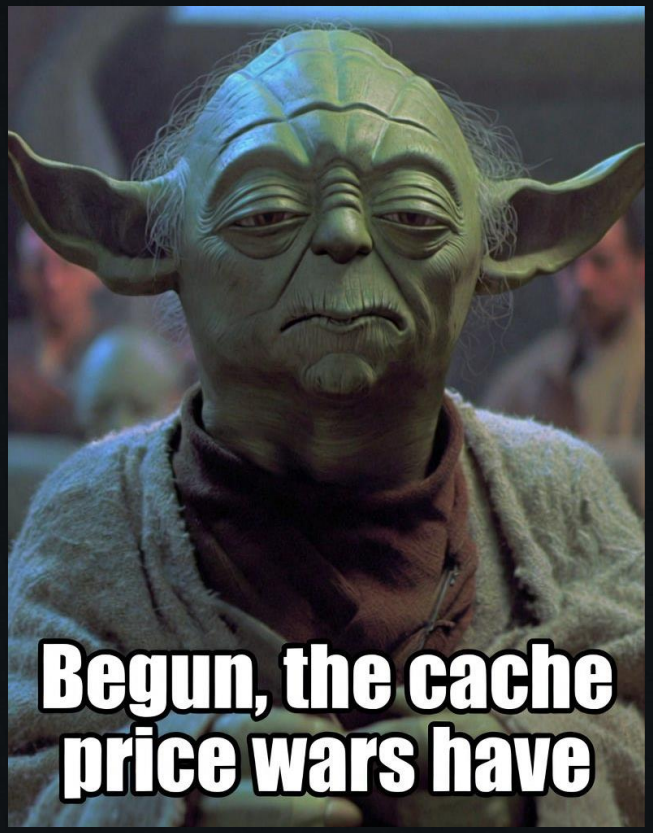
Price per Billion Tokens

DeepSeek-V4-Pro
75% OFF - Now Permanent!

Model	Input (cache hit)	Input (cache miss)	Output
DeepSeek-V4-Pro	$ 0.003625 ~~$ 0.0145~~	$ 0.435 ~~$ 1.74~~	$ 0.87 ~~$ 3.48~~

- After the extended offer ends, the DeepSeek-V4-Pro API price will remain unchanged. It is officially set to 1/4 of the original price!

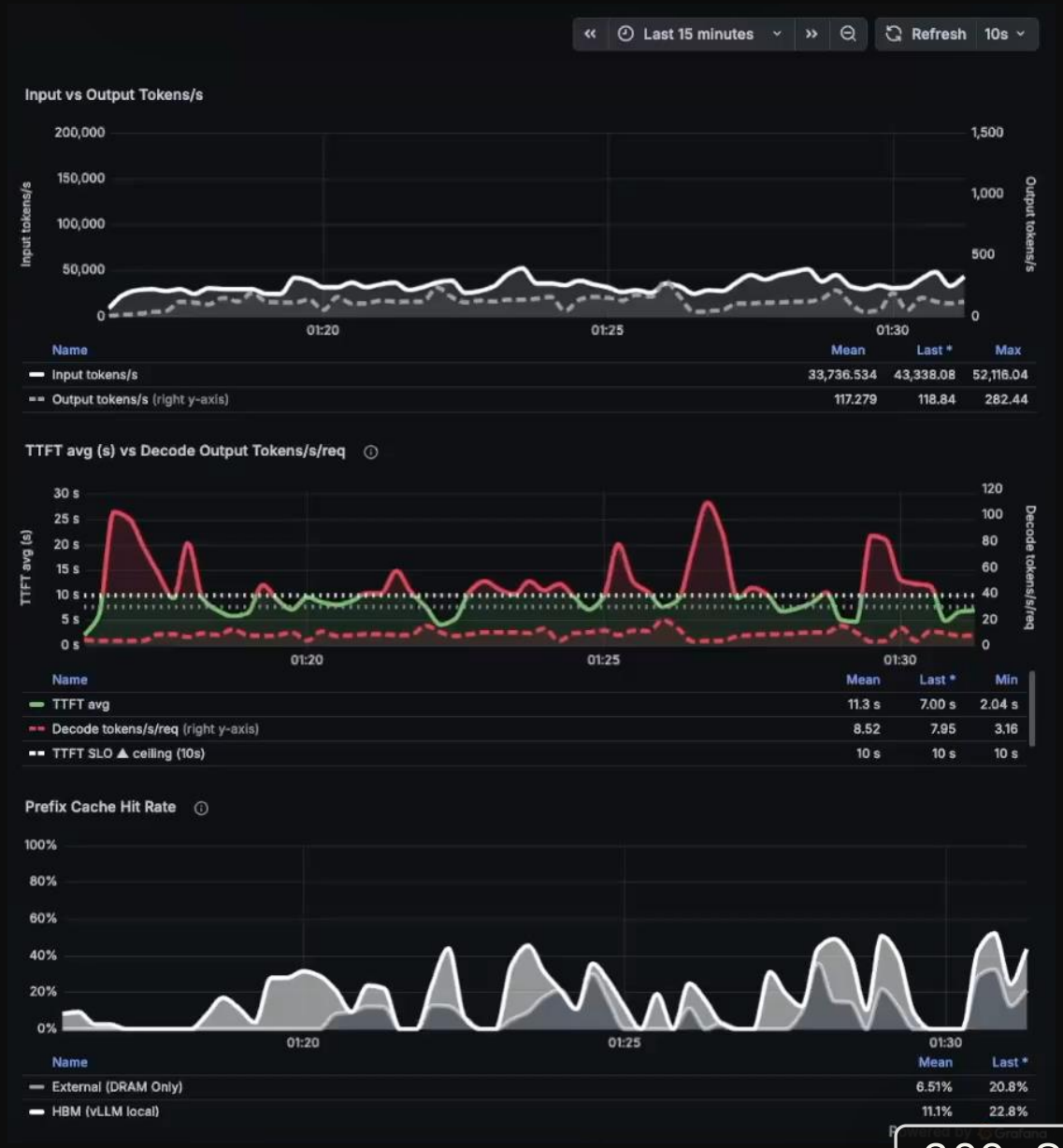
deepseek



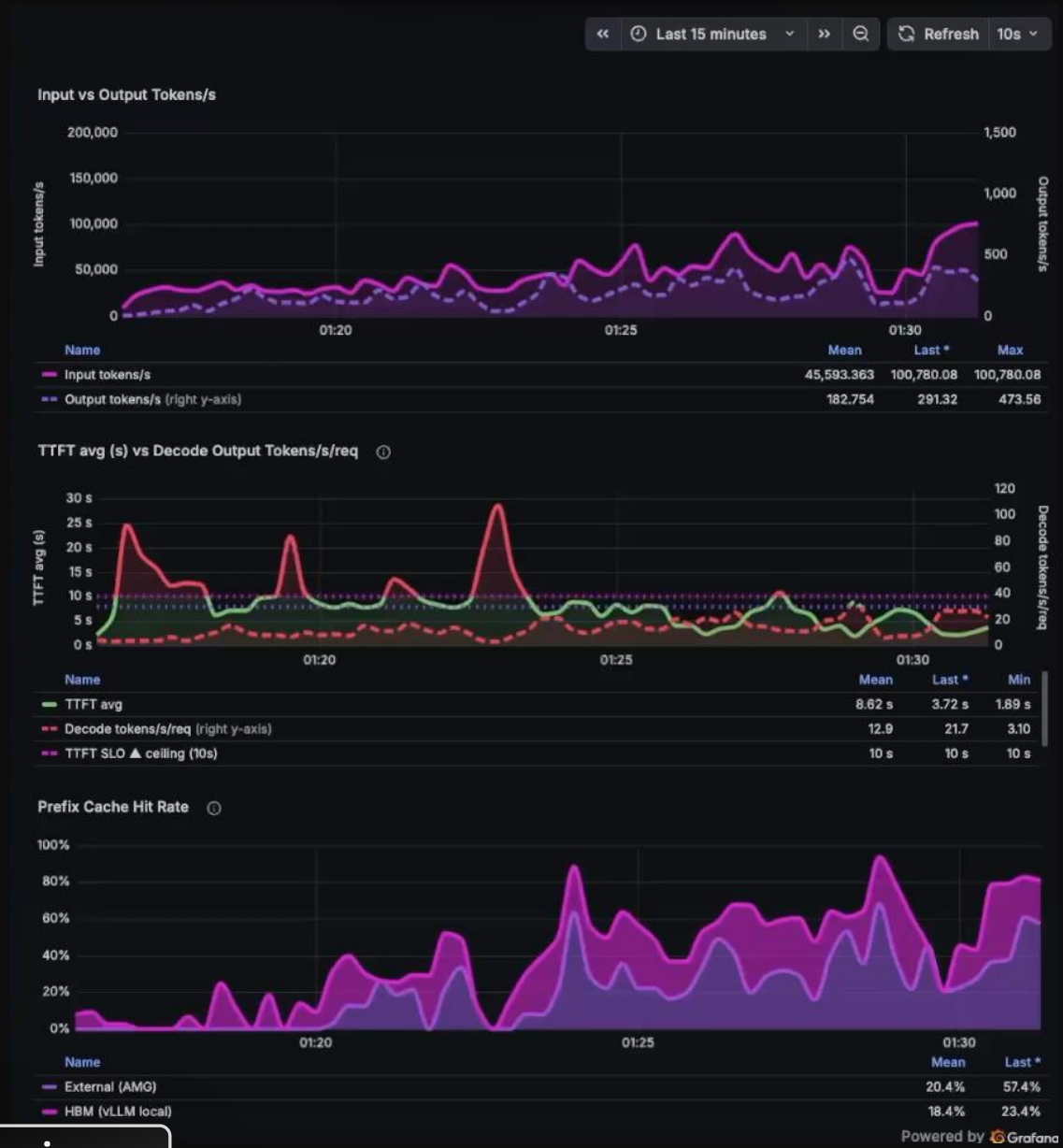
WEKA®

WEKA

Blue Agent Swarm



Red Agent Swarm



300 Sessions

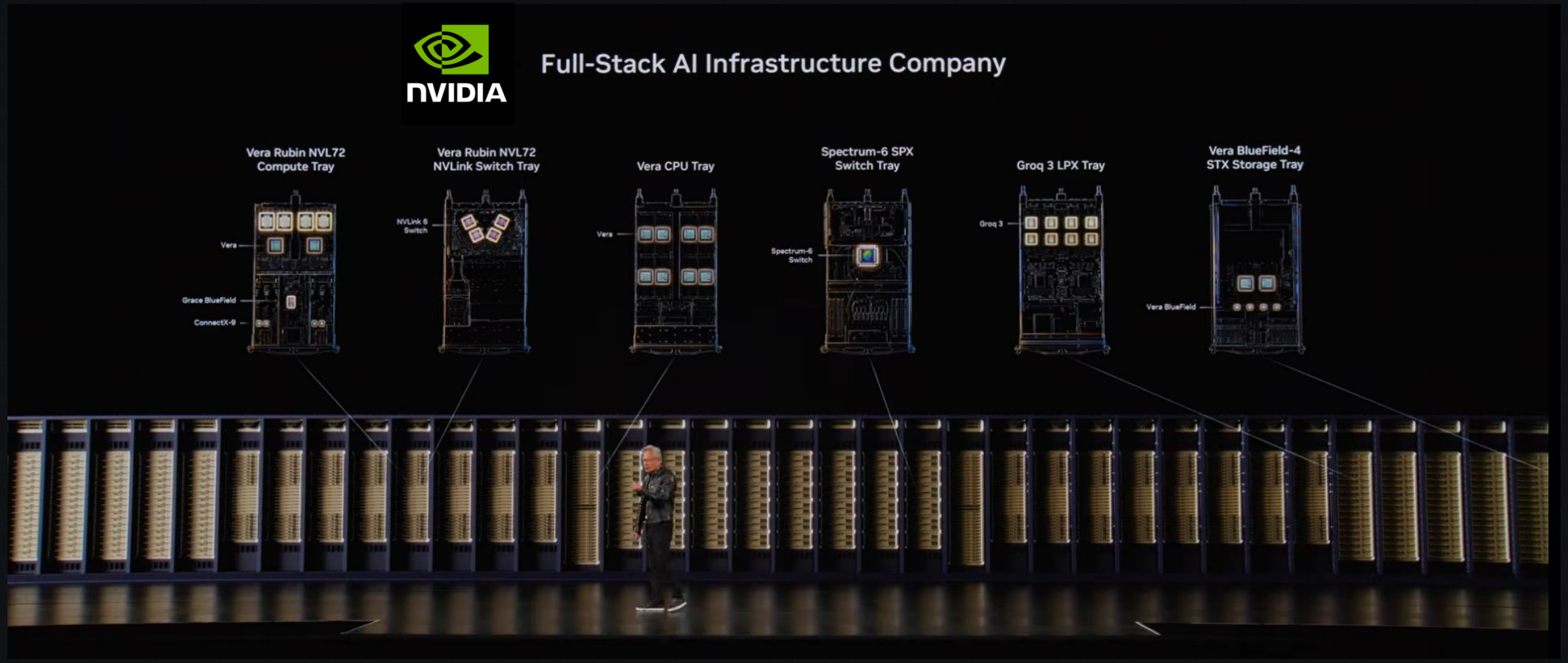
2026

No Agents

Only Swarms

WEKA®

Monday, June 1st (Computex)

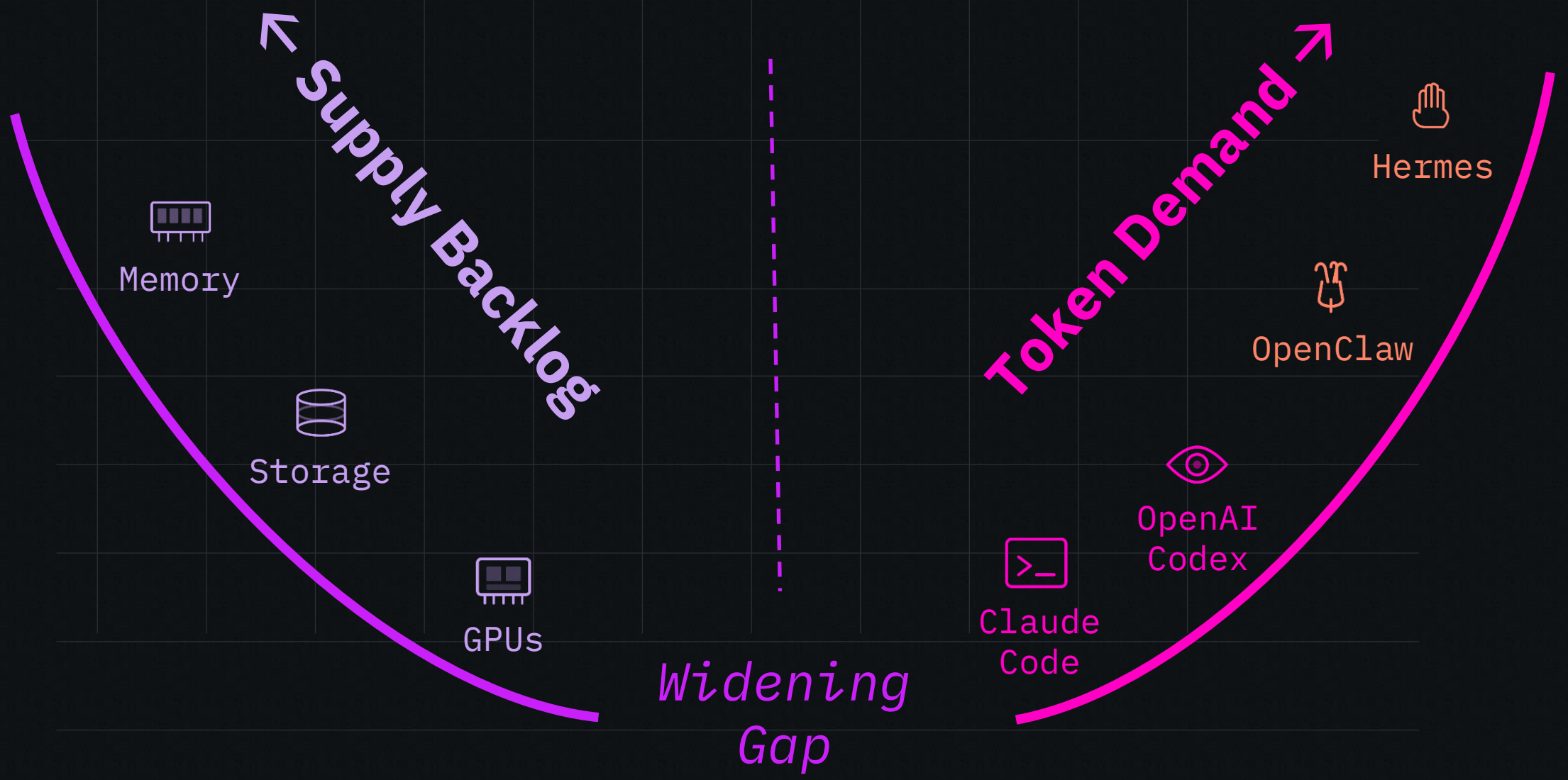


Hopper → Blackwell → Vera Rubin

WEKA®

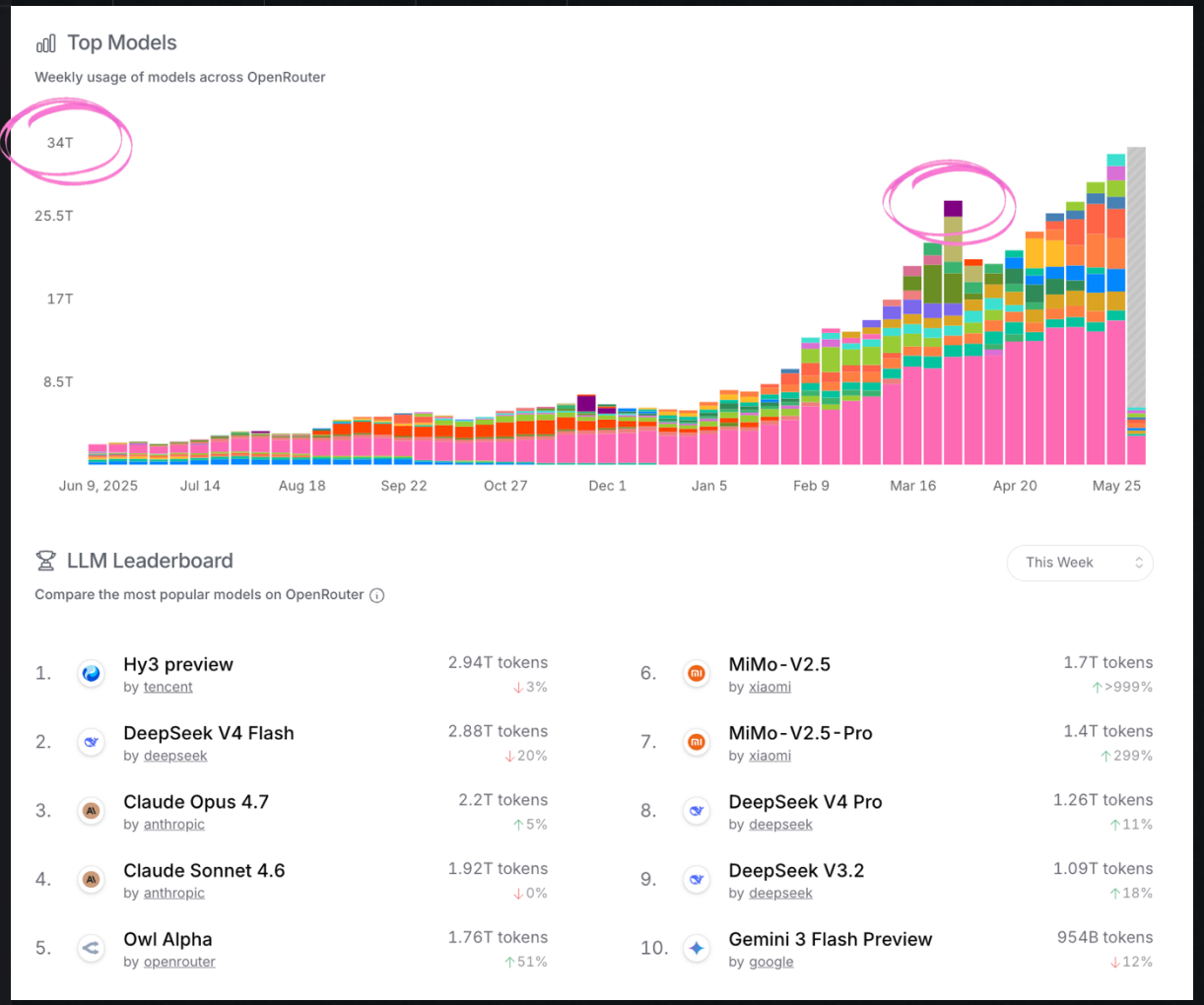
AI Industry Divergence

WEKA



WEKA®

OpenRouter as market proxy



WEKA®

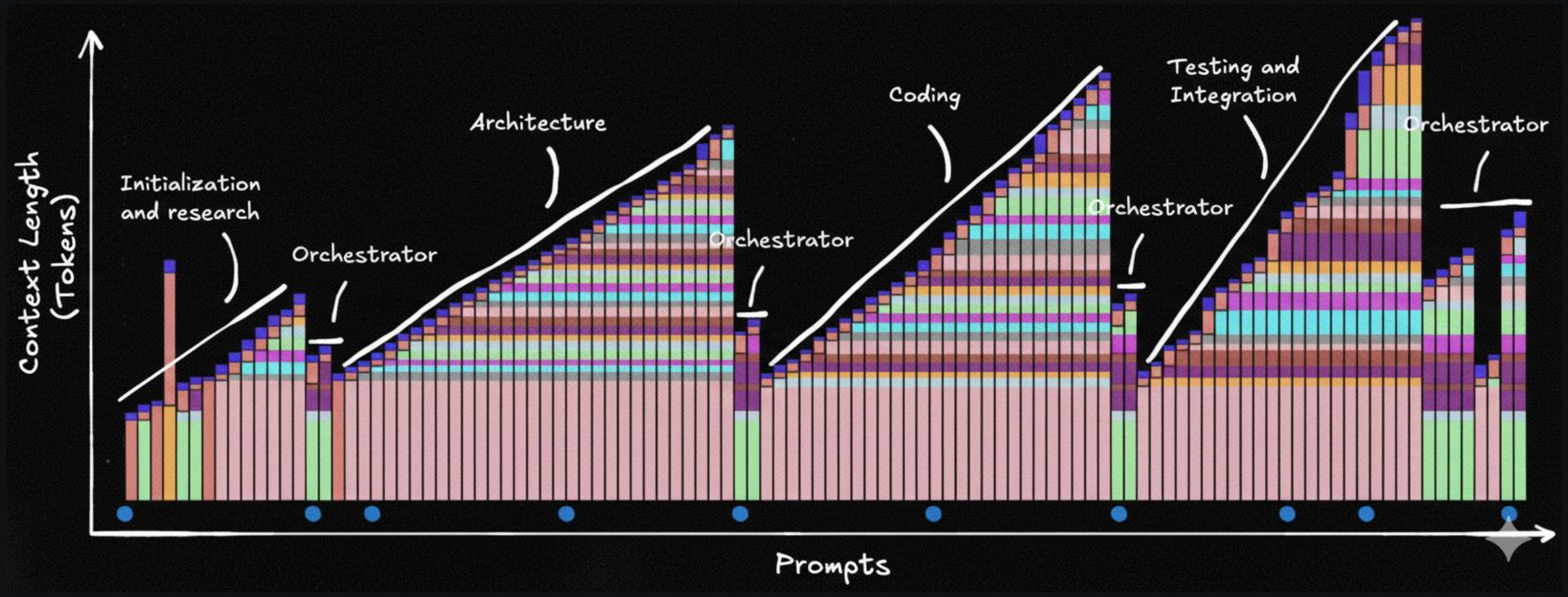
Can We Measure Swarms?

WEKA®

https://www.startuphub.ai/ai-news/ai-video/2025/maximizing-kv-cache-hit-rates-wekas-open-source-context-platform-engineering

[KV Cache Observability 101]
[KV Cache Observability 201] (open source repo by SemiAnalysis InferenceX & Nvidia Dynamo AIPerf)
[KV Cache Observability 301] (Inference Provider Infra)

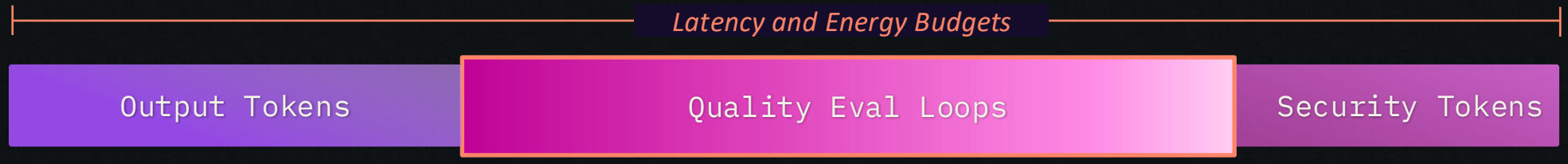
/goal
Claude Workflows



WEKA®

Enterprise Agents ‘3x Rule’

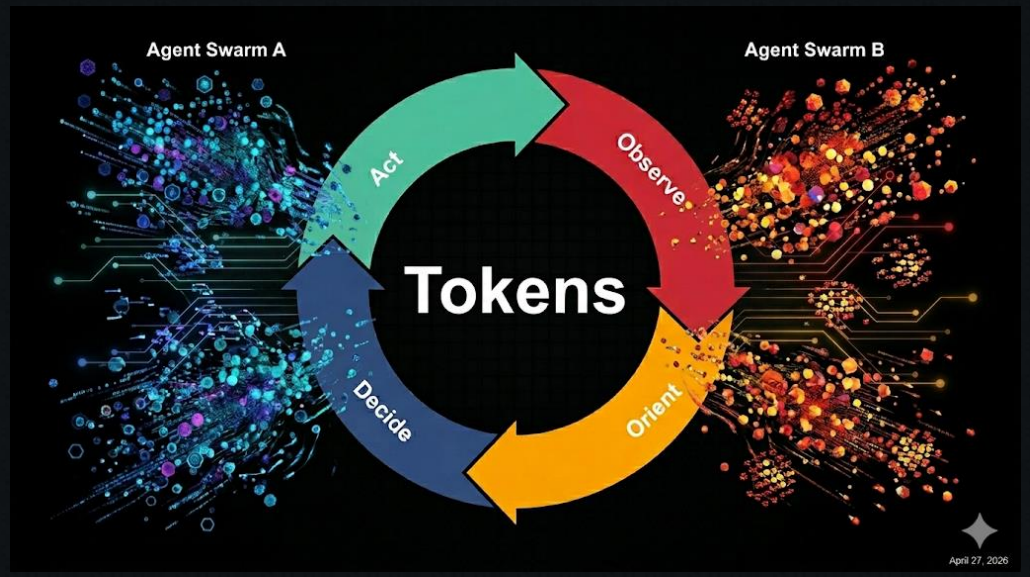
Output tokens are not enough. Quality and Security within Latency and Energy Budgets are the new goal



WEKA®

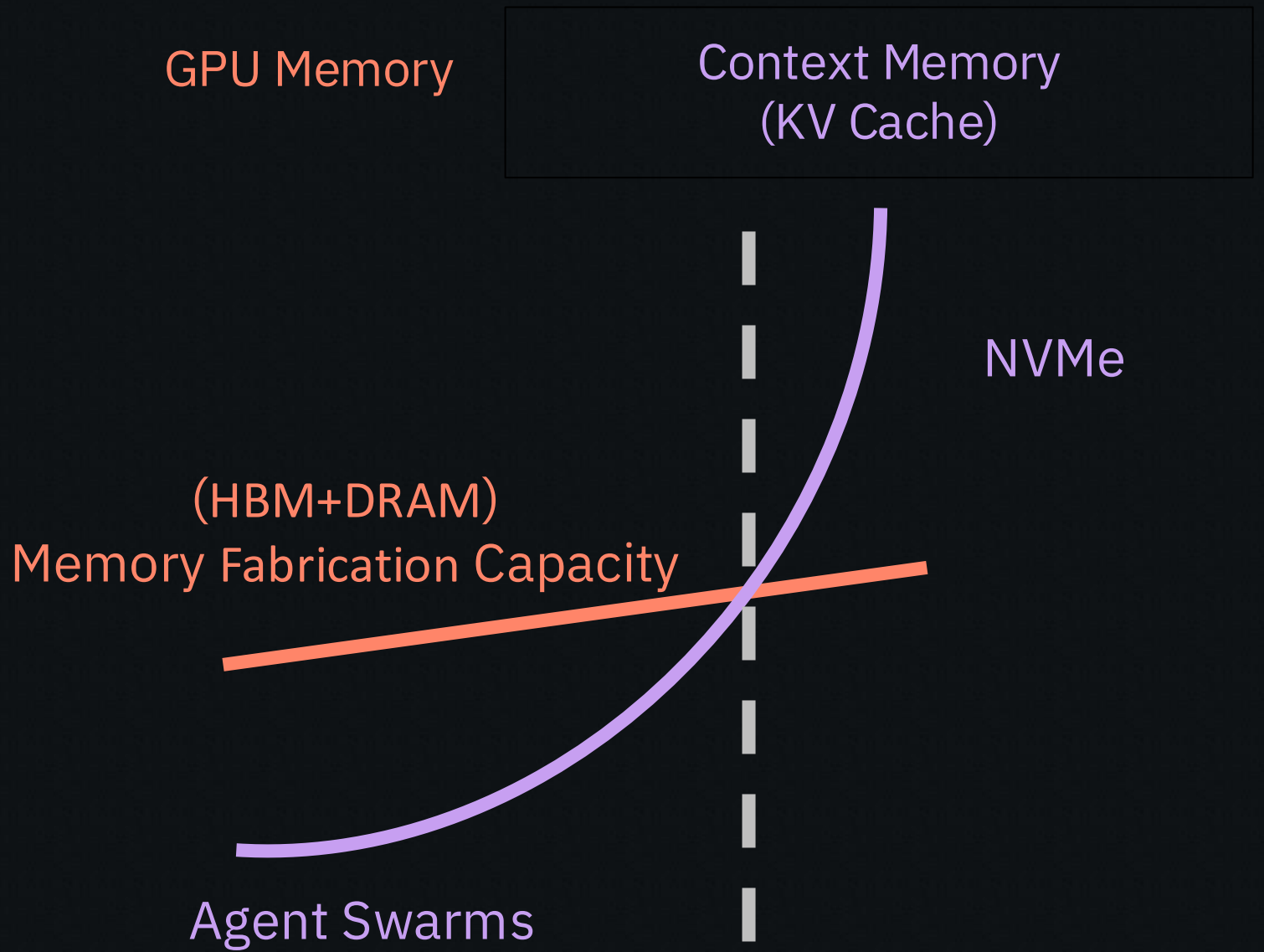
Can We Afford Persistent AI?

Annualized Spend (Claude Opus 4.8 / GPT 5.6?)



WEKA®

The AI Memory Wall



WEKA®

Context Memory Revolution

Context Memory ‘Storage Must be Rearchitected’

"For storage, that is a completely unserved market today. This is a market that never existed, and this market will likely be the largest storage market in the world, basically holding the working memory of the world's AIs."

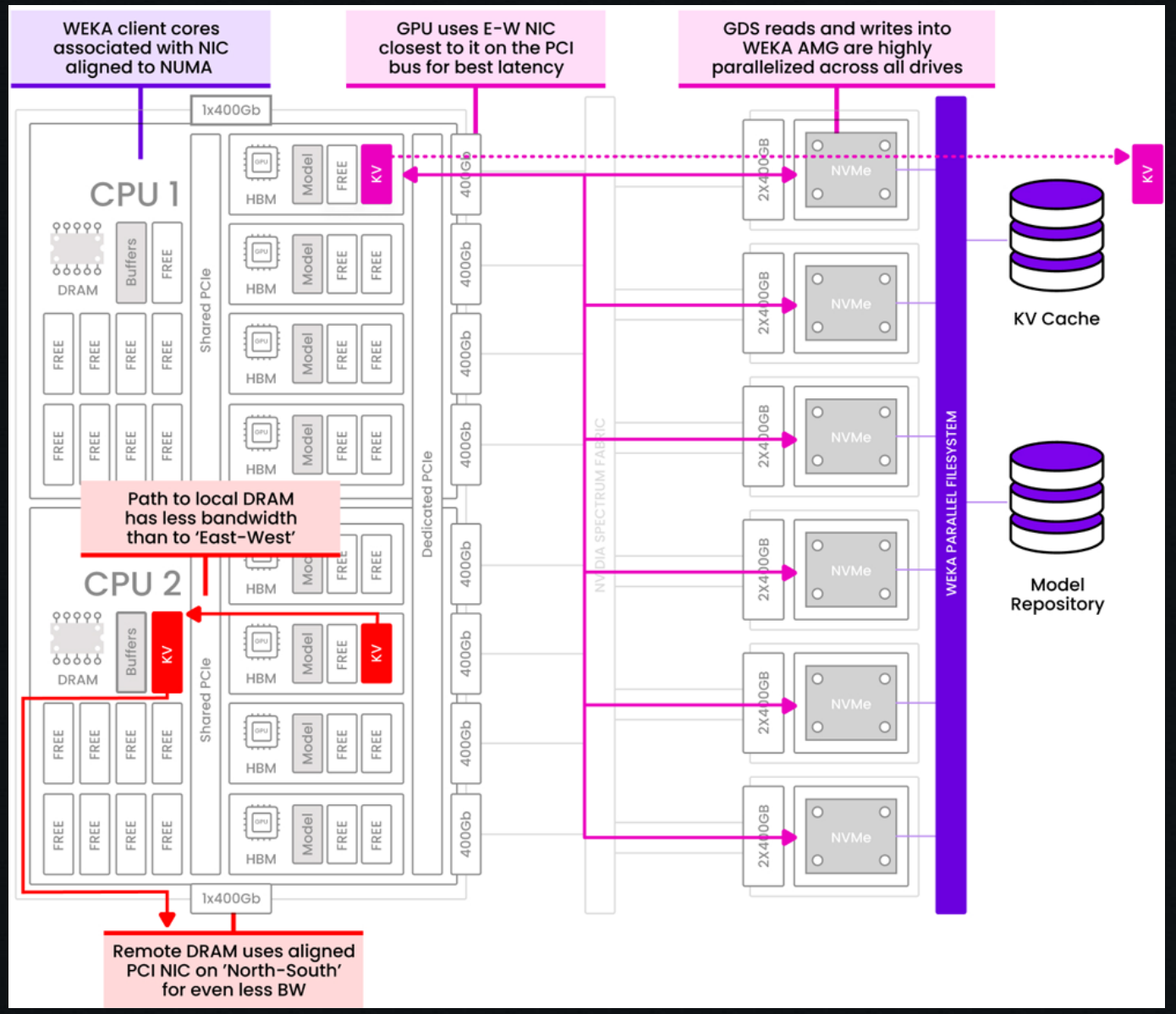
-Jensen Huang @ CES 2026

WEKA®

Line Rate Performance for All Networks

RDMA over E-W network > DRAM Bandwidth

- Single host / mount performance of **~380GB/s** KV Cache reads
- More detail in AMG Blog



WEKA®

Technical Blog

Search blog

How are storage providers taking advantage of KV Cache Offload?

Vast tested a high-performance integration between NVIDIA Dynamo and the Vast AI OS to enable persistent KV Cache movement between GPU and storage. Using the GPU Direct Storage (GDS) plugin in Dynamo, Vast achieved **35GB/s** throughput to a single NVIDIA H100 GPU, demonstrating full GPU saturation and confirming that storage was not a performance bottleneck.

In a separate test, Vast validated the impact of persistent KV Cache reuse using vLLM and LMCache on an NVIDIA DGX H100 system. Running the Qwen3-32B model with a 130K-token prompt, the system loaded precomputed KV cache from Vast storage rather than recomputing it, reducing Time to First Token (TTFT).

WEKA conducted lab testing to evaluate high-performance KV Cache movement between GPU and storage using NVIDIA Dynamo and a custom NIXL plugin developed and open-sourced by WEKA. The tests demonstrated that the WEKA's Augmented Memory Grid can stream KV Cache from its token warehouse to GPUs at near-memory speeds, reducing TTFT and improving overall token throughput for inference workloads.

Testing was performed using a DGX system with eight H100 GPUs. The setup achieved read throughput up to **270GB/s** across eight GPUs, validating that WEKA's RDMA-based, zero-copy data path can meet the demands of disaggregated inference without becoming a bottleneck.

These test results highlight the potential of KV Cache offload to storage in supporting large-context, high-throughput generative AI workloads in distributed environments.

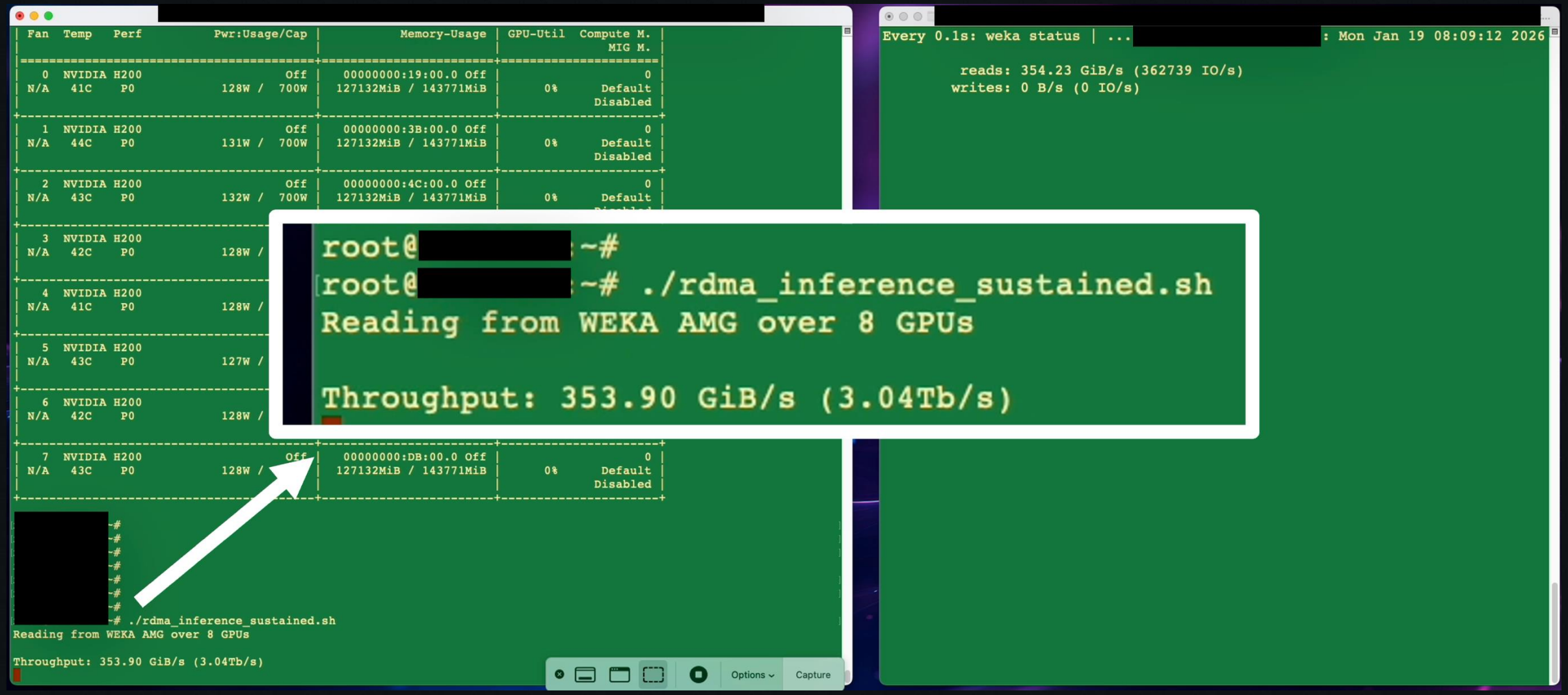
NVIDIA DEVELOPER

Dynamo NIXL

https://developer.nvidia.com/blog/how-to-reduce-kv-cache-bottlenecks-with-nvidia-dynamo

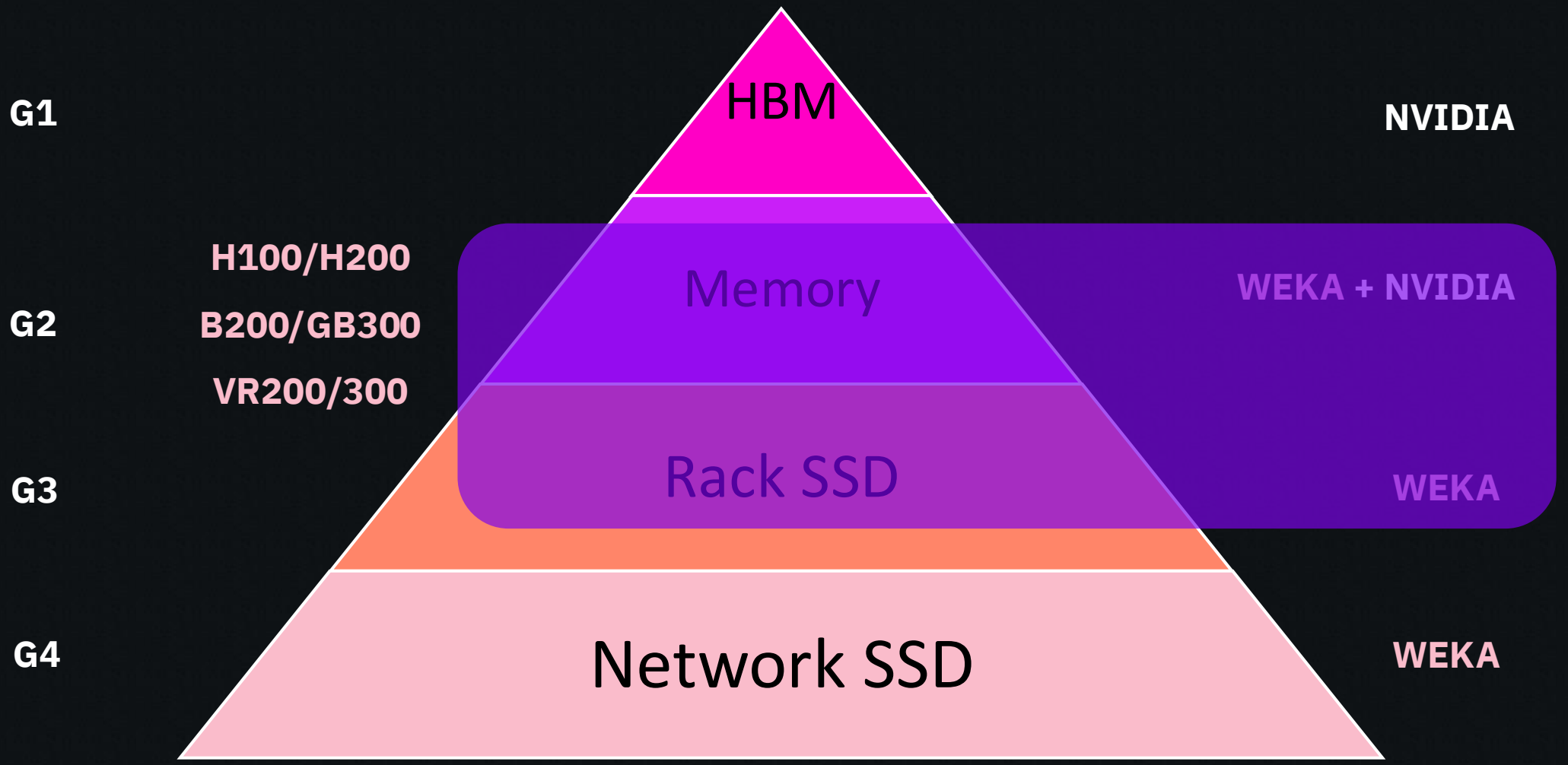
WEKA®

RAW RDMA Performance



WEKA®

Memory Storage Hierarchy

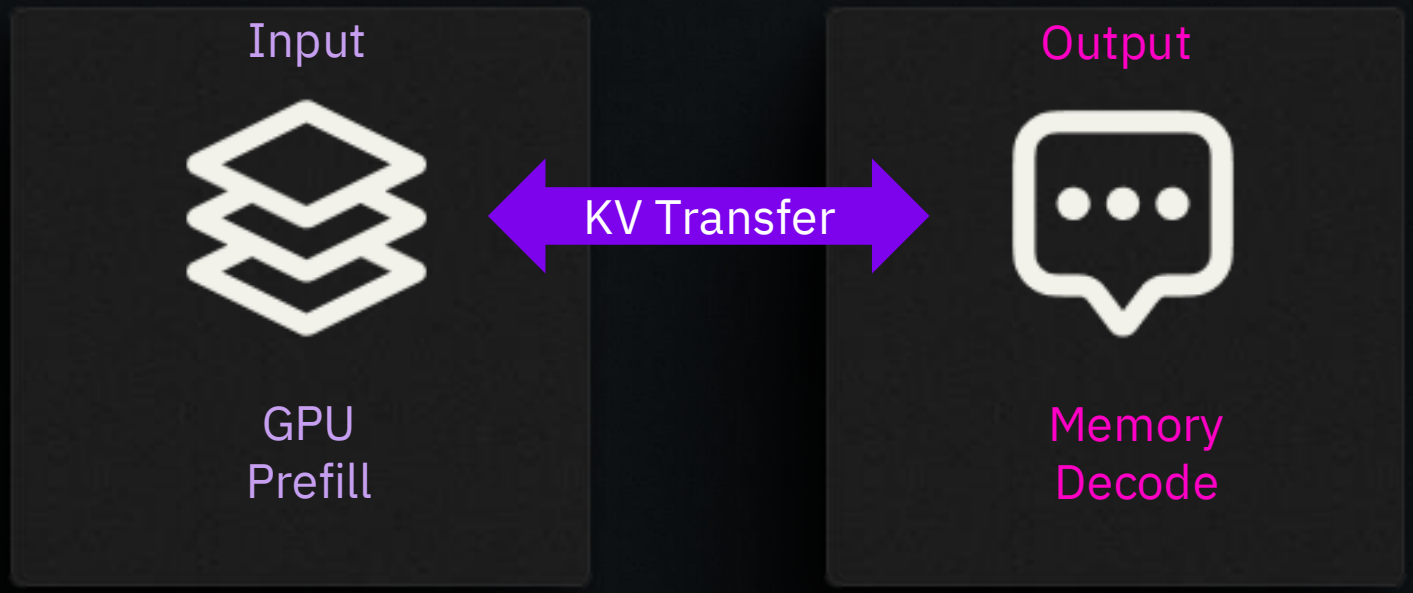


WEKA

Token Warehousing

WEKA®

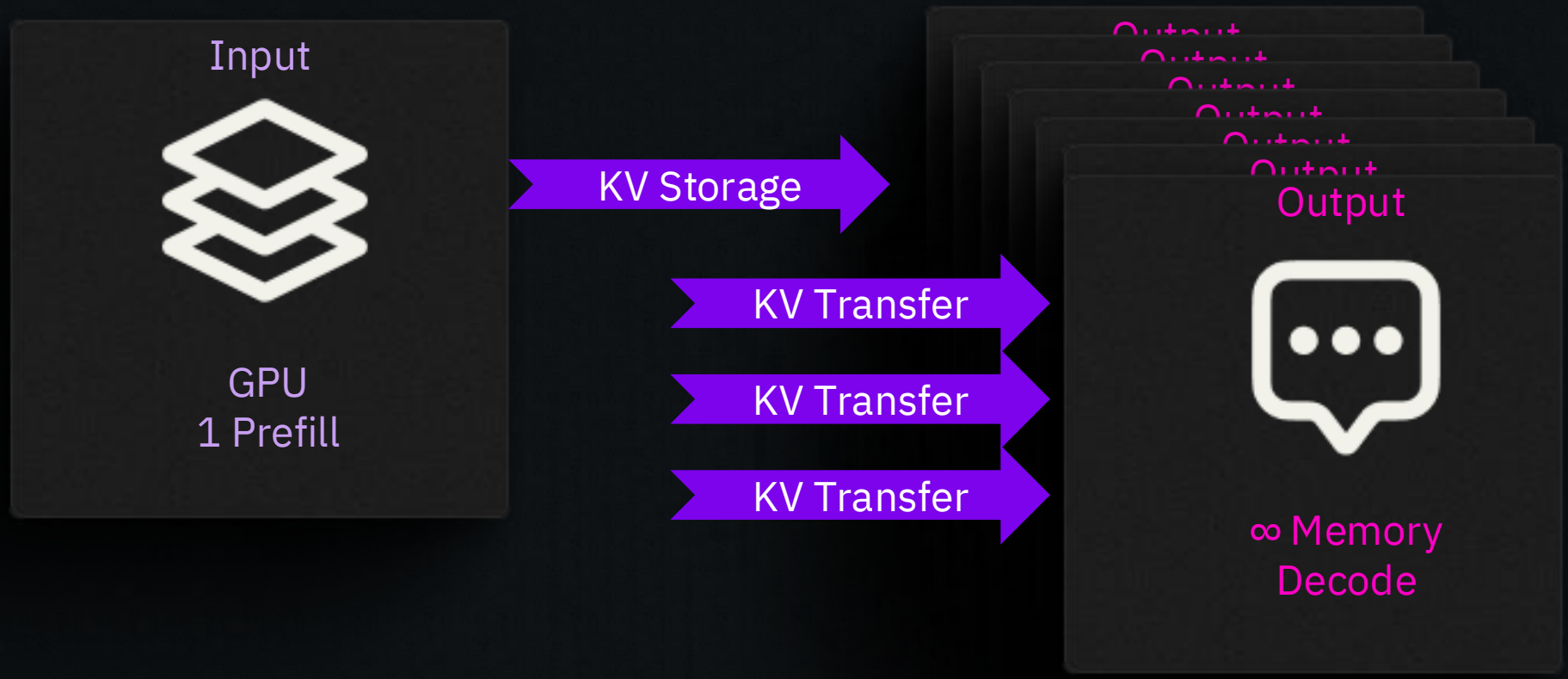
Factory → Consumer?



WEKA®

WEKA®

Attention Warehouse → Output

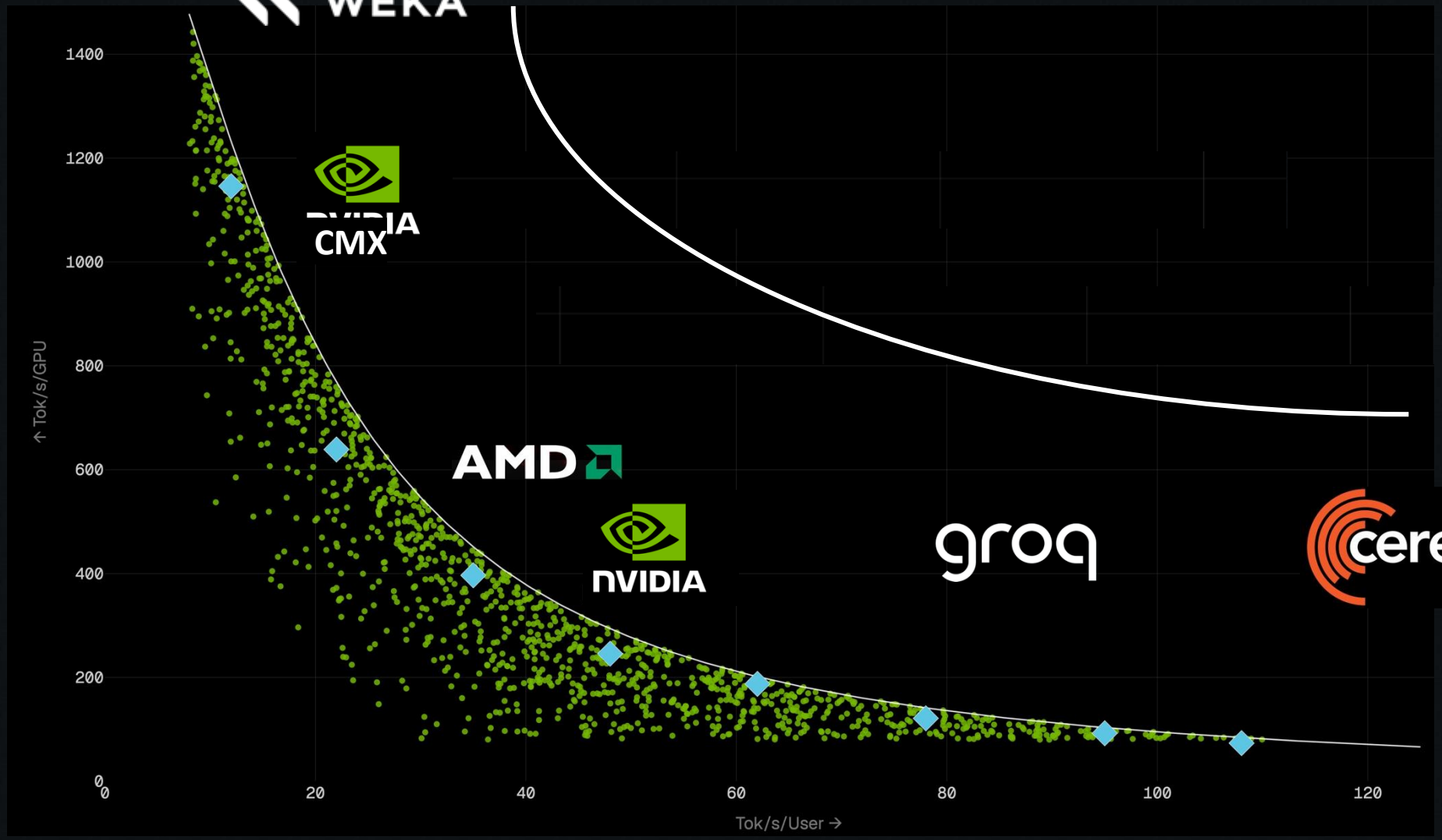


WEKA

Attention

WEKA

Pareto Frontier



Output

Independent Proof Points

WEKA

Oracle AI & Data Science Blog

Scaling Long-Context Inference on OCI with WEKA's Augmented Memory Grid

Metric	DRAM Baseline	Augmented Memory Grid	Improvement
Max concurrent users	~600	5,000+	10X
Requests completed, 2,400-user test	~6,700	47,000+	7X
Tokens served	700M	5B	7X
Token throughput	<200K tokens/sec	~2M tokens/sec	10X

WEKA®

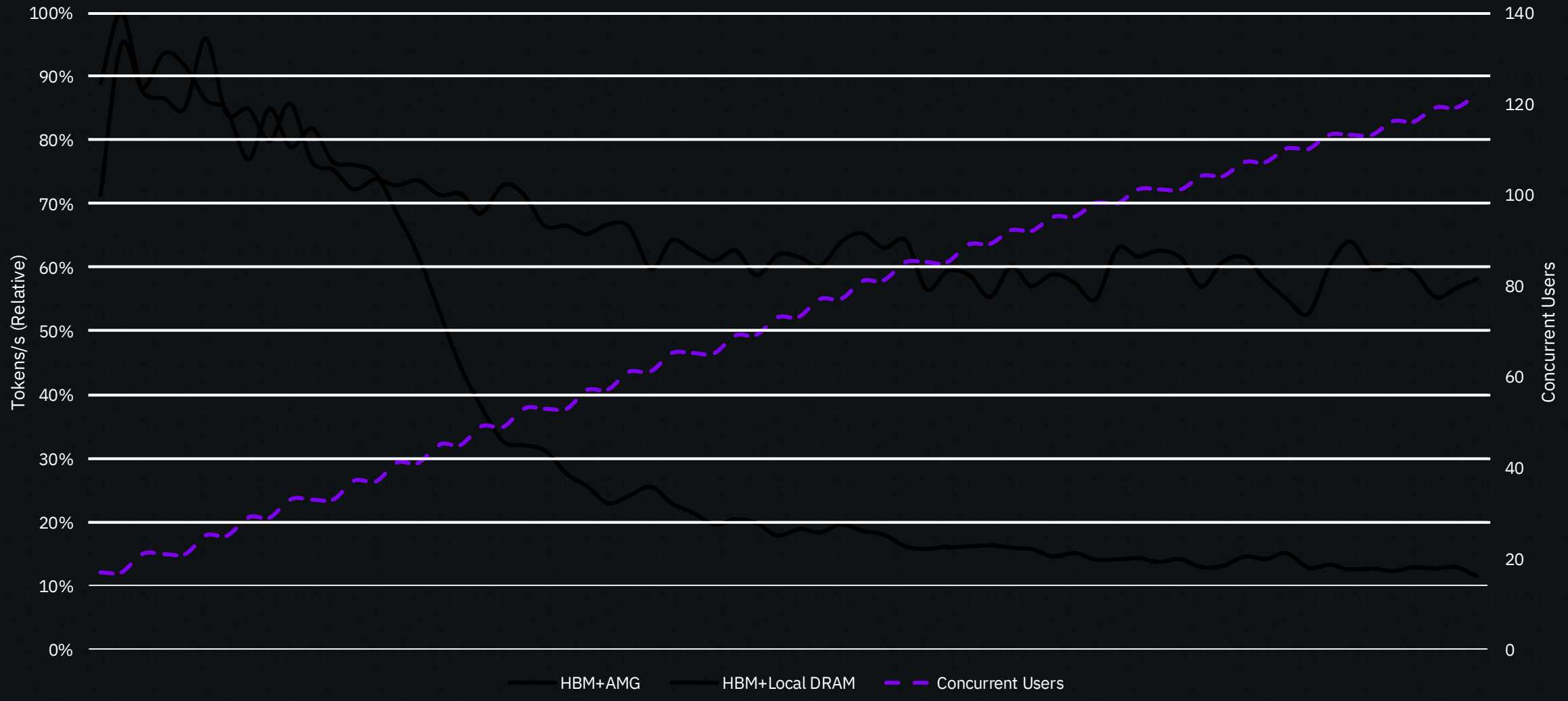
Performance Data

Real-World Insights

WEKA

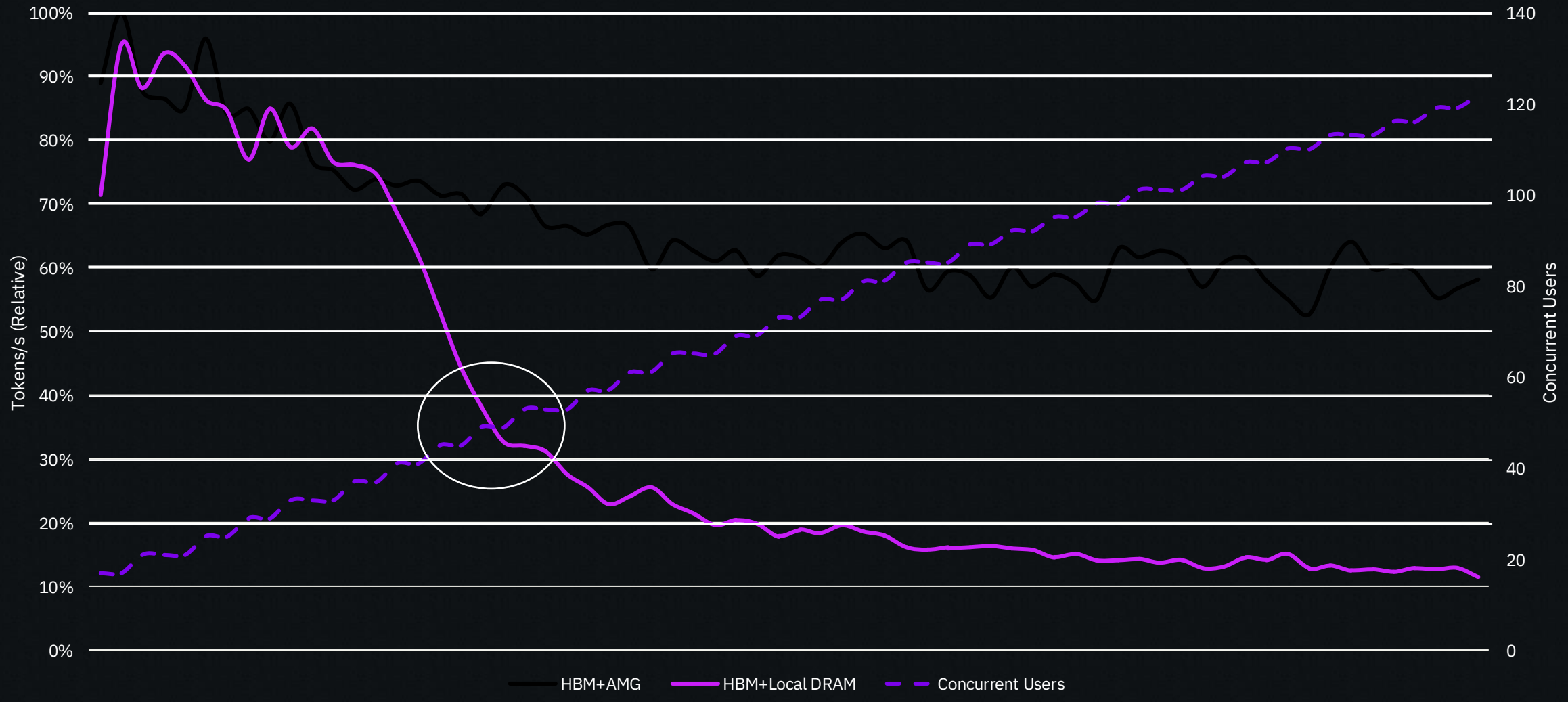
18-125 Concurrent Session

Qwen3-Coder-480B Inference Performance



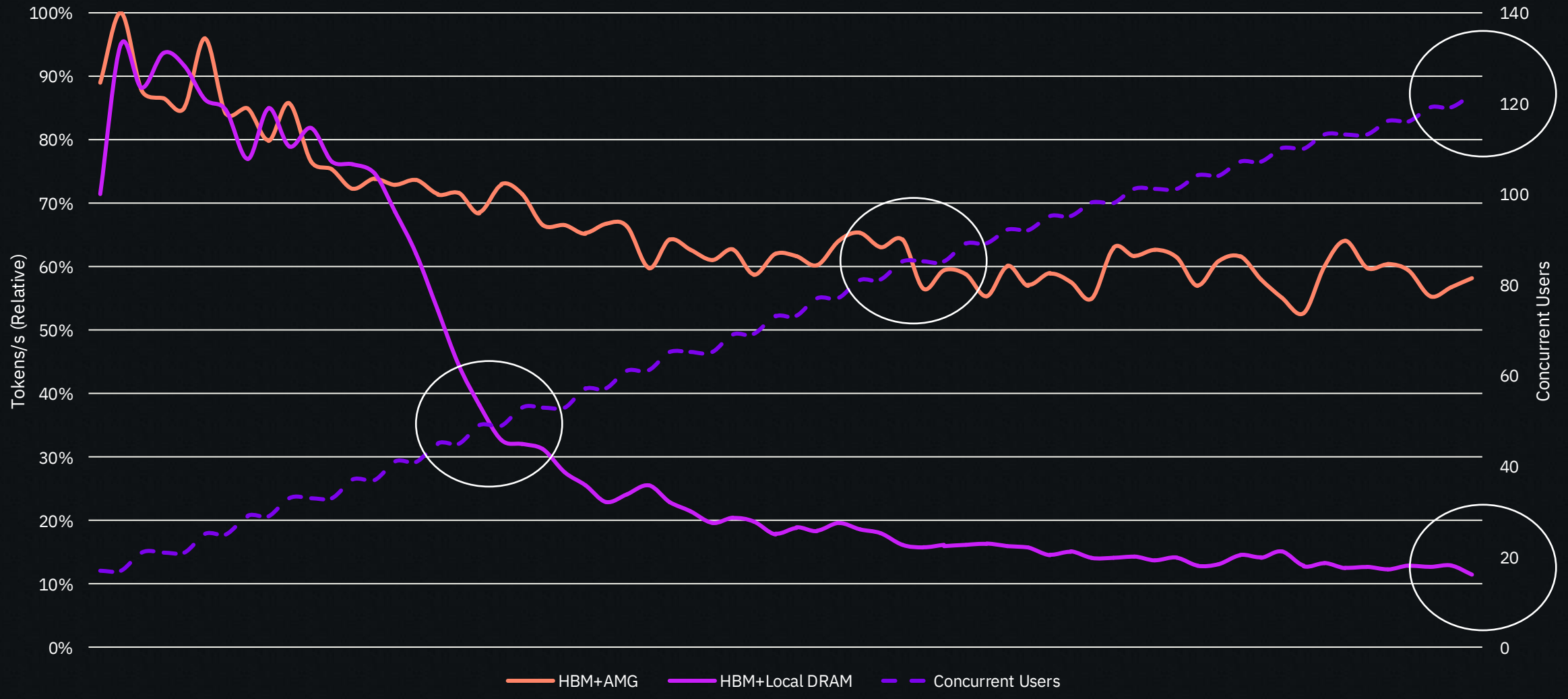
WEKA®

GPU HBM/VRAM + CPU DRAM = Performance Cliff



© WEKA 2026

GPU HBM/VRAM + WEKA Augmented Memory Grid =
New Performance Plateau



WEKA®

5-10 Instant NEW Data Centers

Physical

Data Center

WEKA

Virtual Data Center

WEKA

Virtual Data Center

WEKA

Virtual Data Center

WEKA

Virtual Data Center

What it Takes to Win

- More Swarm Tokens
- 99% KV Cache Hit Rate
- Prefill Once, Decode Forever
- Abundant Quality & Safety Tokens (Every Agent Step)

THANKS FOR YOUR TIME

Learn How to Maximize Your AI Token Production

WEKA®


WEKA®