



Persistent Agent Swarms

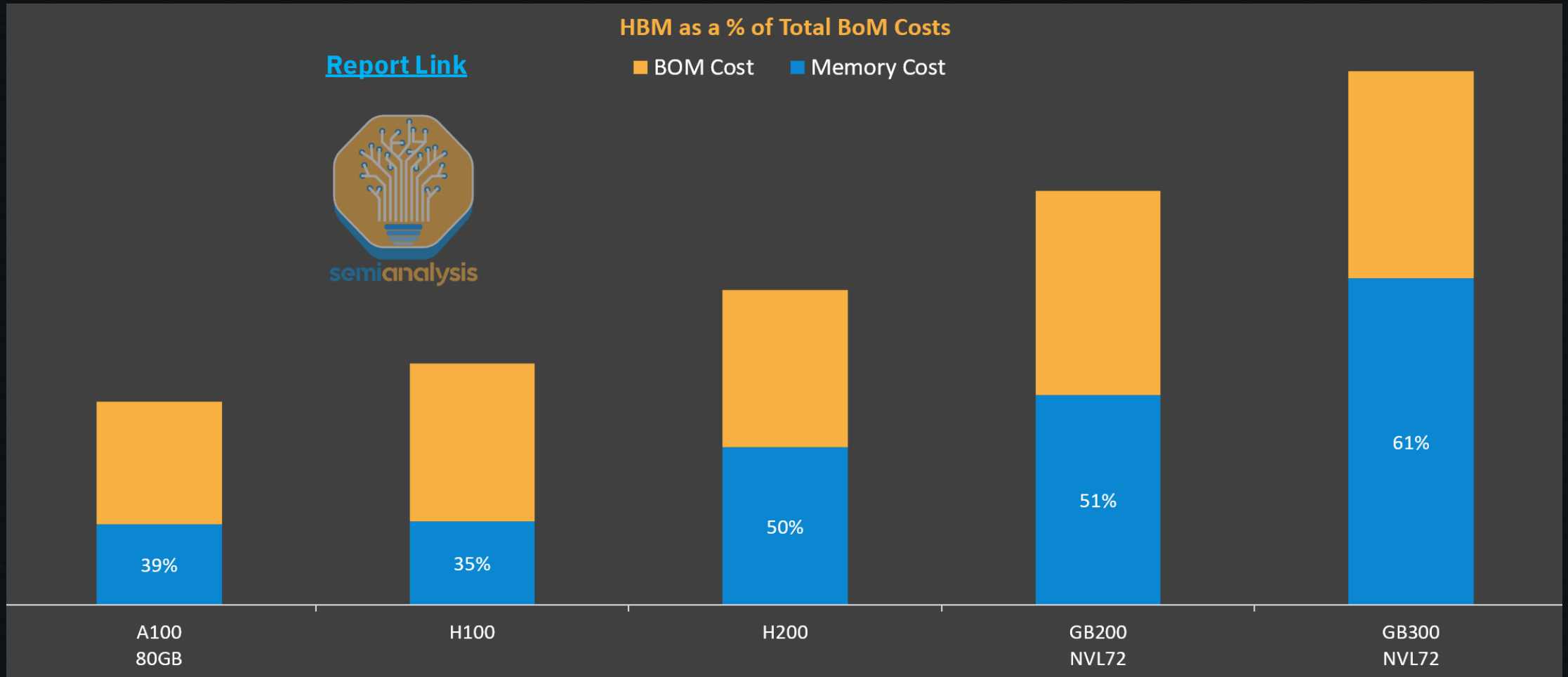
Use-cases & Inference Implications



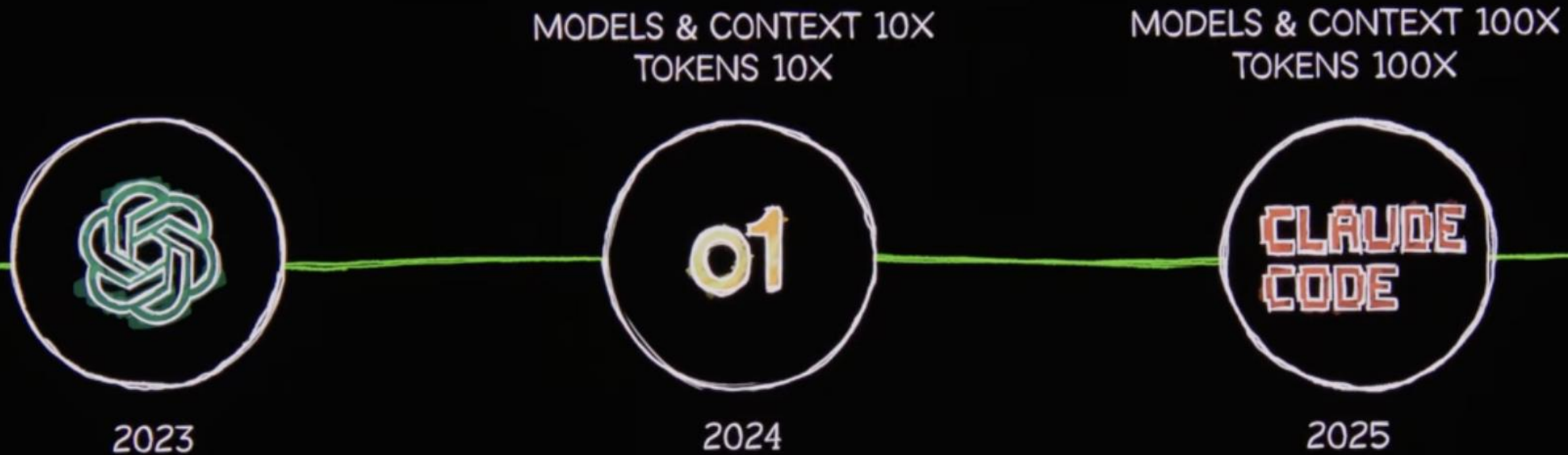
Val Bercovici								
Chief AI Officer								

AI Future = Fusion of
Storage + Memory

GPU Server Bill of Materials



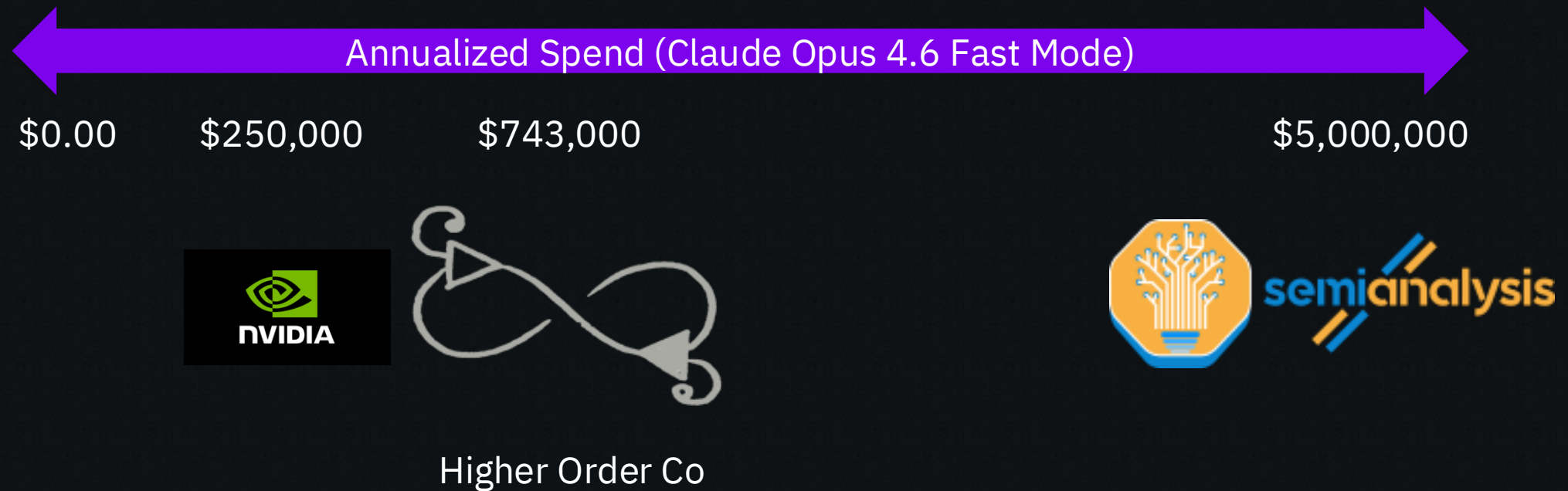
March 2026 (GTC)



Inference Inflection Arrives: 10,000x ChatGPT Compute



Remember Tokenmaxxing?





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)



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_](#)

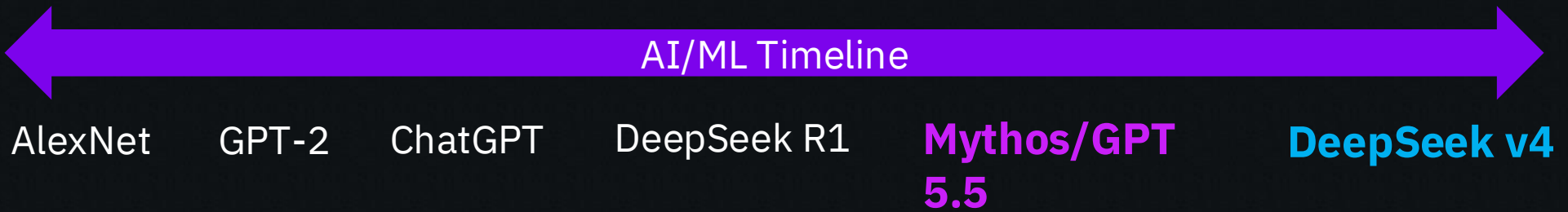
WATCH LIVE

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

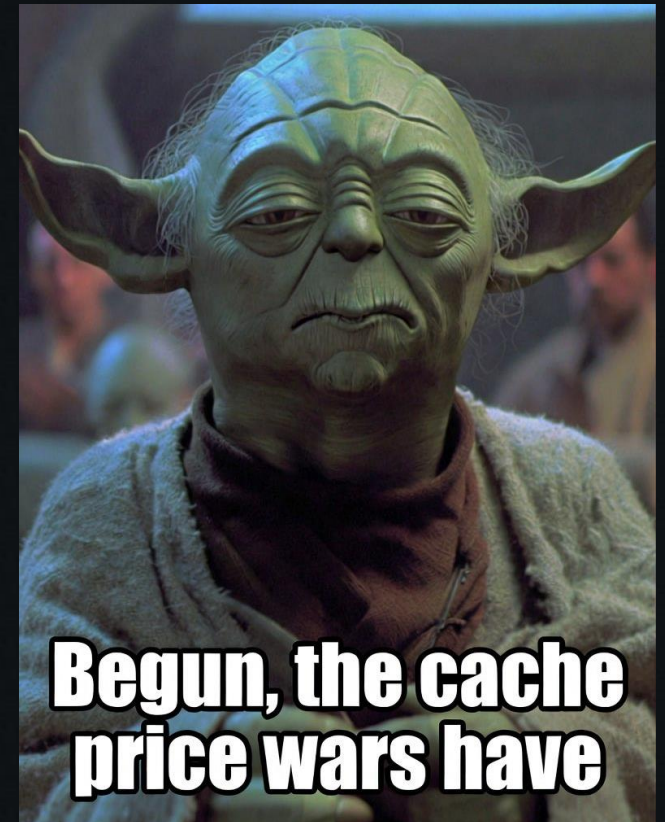


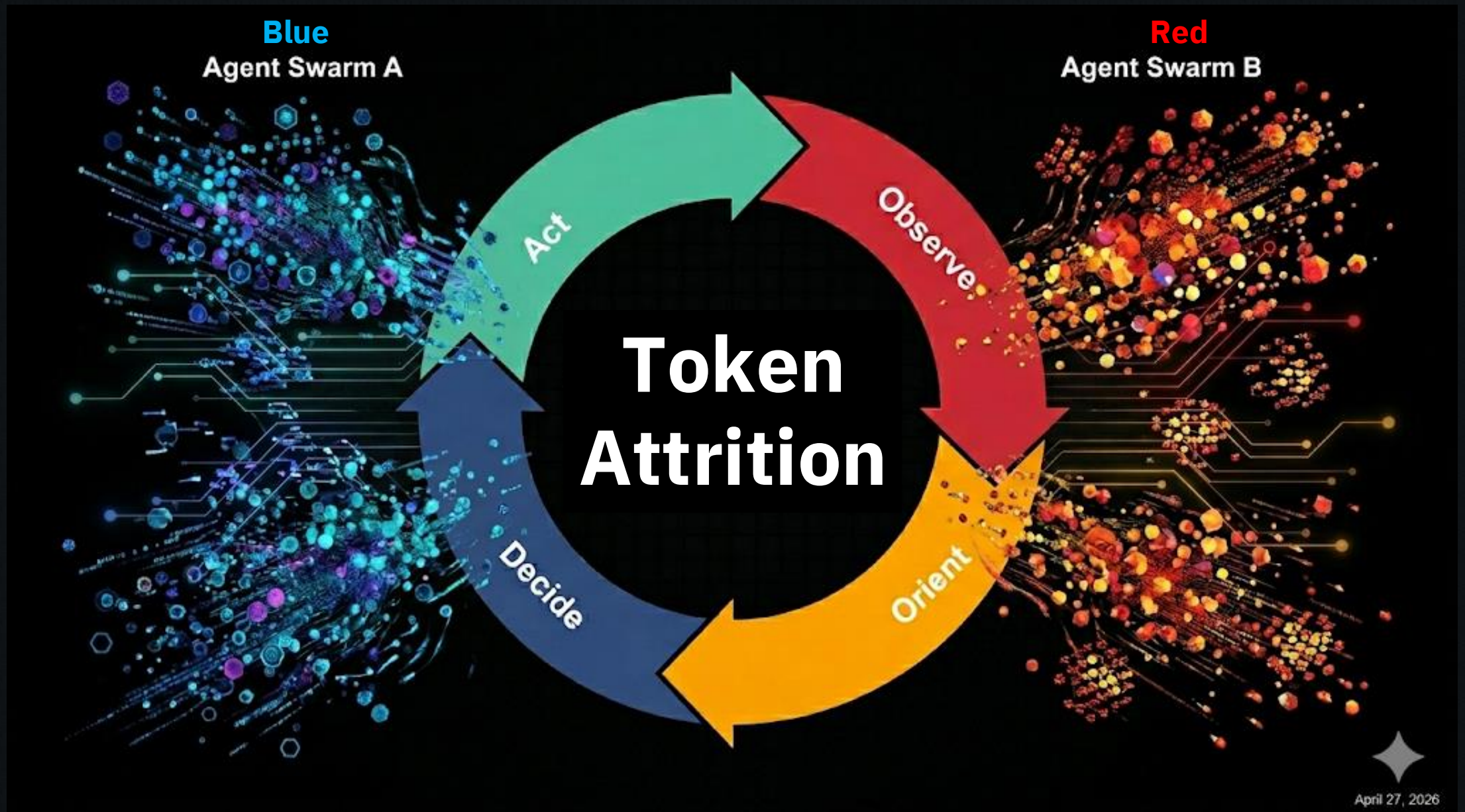
Price per Billion Tokens

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!



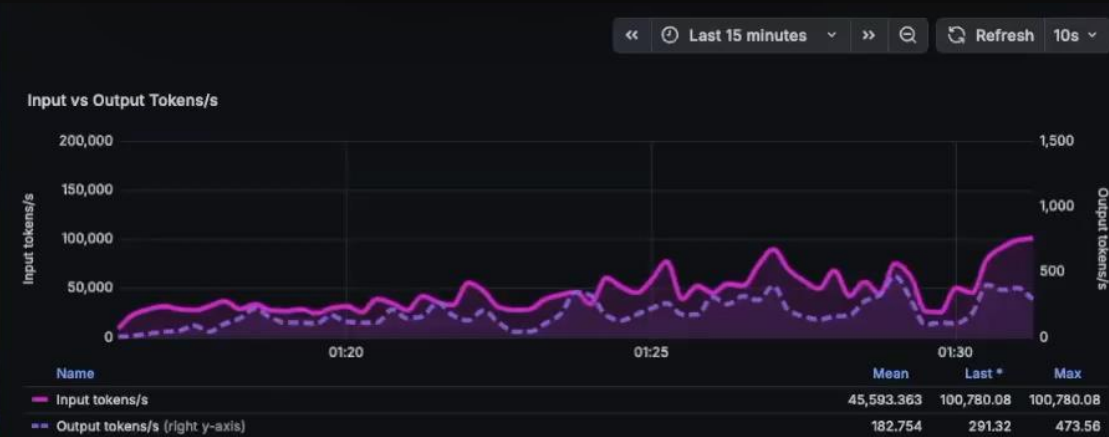


April 27, 2026

Blue Agent Swarm



Red Agent Swarm



300 Sessions



2026

No Agents

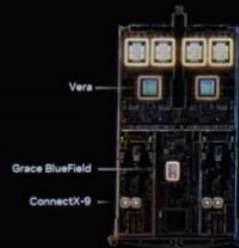
Only Swarms

Monday, June 1st (Computex)



Full-Stack AI Infrastructure Company

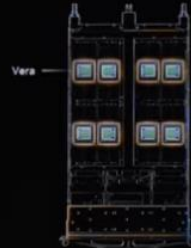
Vera Rubin NVL72
Compute Tray



Vera Rubin NVL72
NVLink Switch Tray



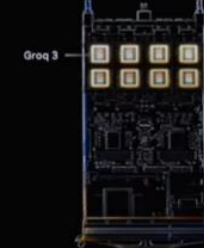
Vera CPU Tray



Spectrum-6 SPX
Switch Tray



Groq 3 LPX Tray



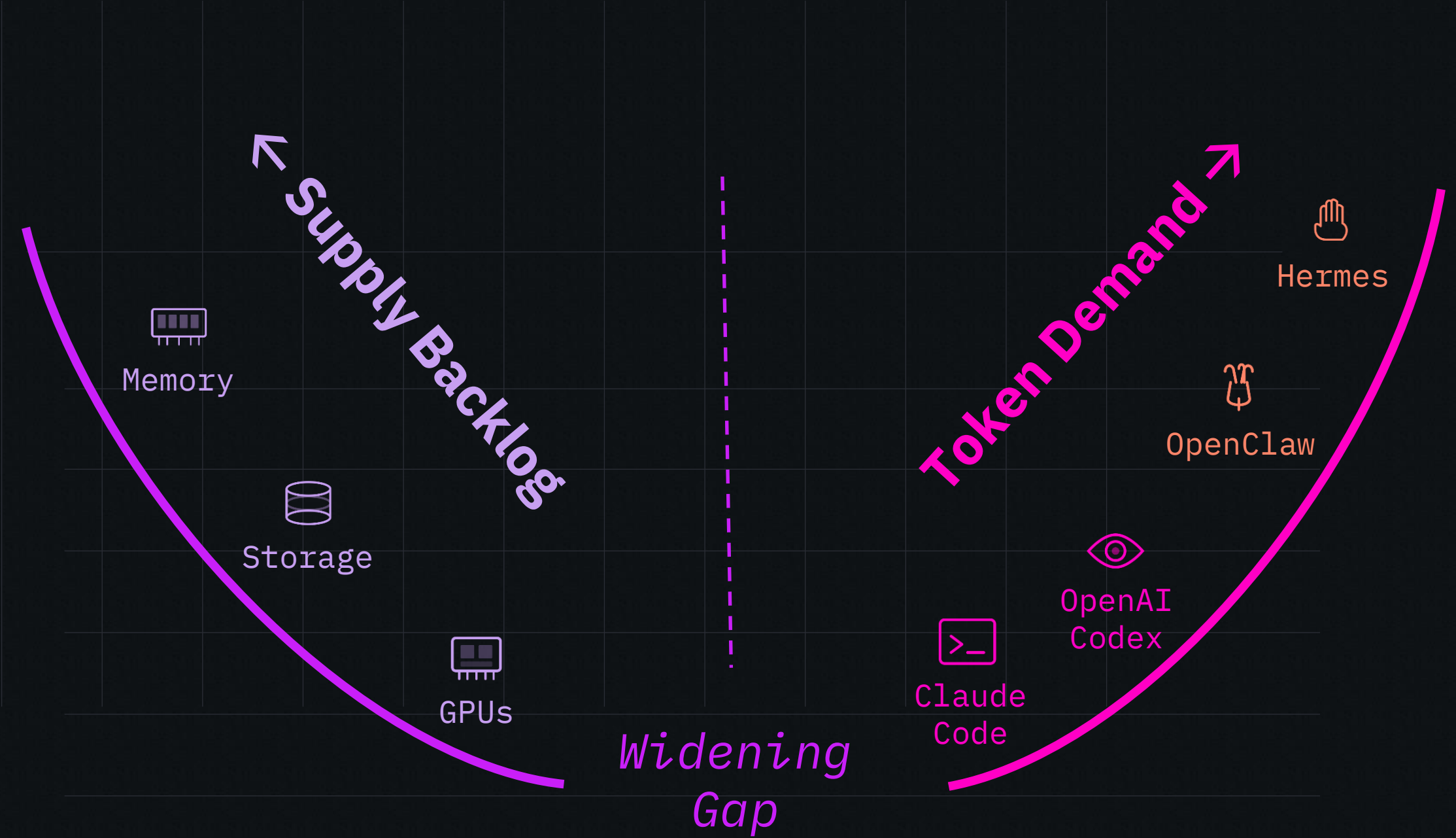
Vera BlueField-4
STX Storage Tray



Hopper → Blackwell → Vera Rubin

The background of the slide features a dark, grayscale image of financial charts. The top half shows a candlestick chart with a white line, possibly a moving average, that trends upwards and then sharply downwards. The bottom half shows a bar chart with many vertical bars of varying heights, representing price movements over time. The overall aesthetic is technical and data-driven.

AI Industry Divergence



OpenRouter as market proxy

< OpenRouter

Search

Models

Chat

Rankings

Apps

Enterprise

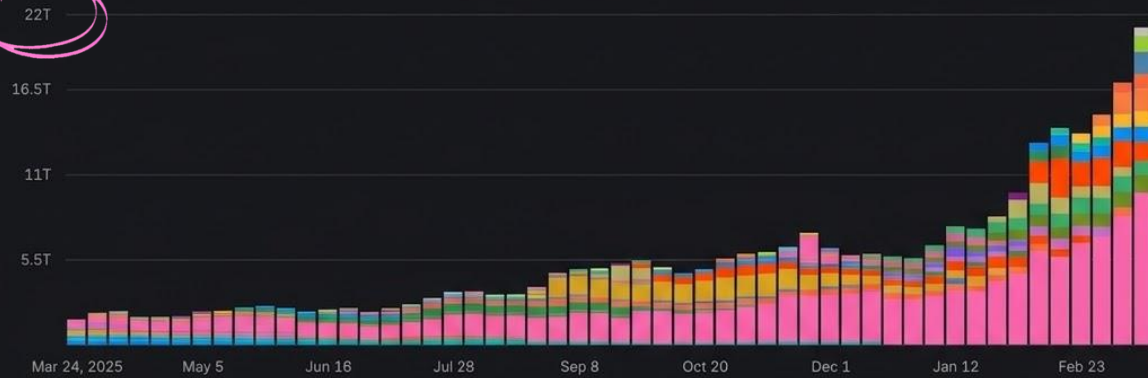
Pricing

Docs

Sign Up

Top Models

Weekly usage of models across OpenRouter



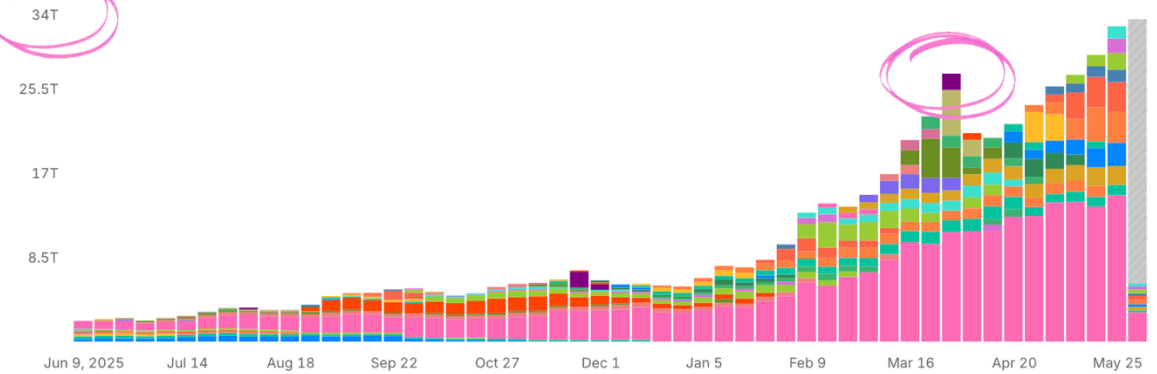
LLM Leaderboard

Compare the most popular models on OpenRouter

1.	Step 3.5 Flash (free) by stepfun	1.47T tokens ↑13%	6.	MiMo -V2-Pro by xiaomi	1.02T tokens new
2.	MiniMax M2.5 by minimax	1.36T tokens ↓25%	7.	Claude Opus 4.6 by anthropic	974B tokens ↑21%
3.	Hunter Alpha by openrouter	1.15T tokens ↑164%	8.	Gemini 3 Flash Preview by google	938B tokens ↓7%
4.	DeepSeek V3.2 by deepseek	1.12T tokens ↑11%	9.	GLM 5 Turbo by z-ai	890B tokens new
5.	Claude Sonnet 4.6 by anthropic	1.03T tokens ↑18%	10.	Gemini 2.5 Flash by google	584B tokens ↑5%

Top Models

Weekly usage of models across OpenRouter



LLM Leaderboard

Compare the most popular models on OpenRouter

1.	Hy3 preview by tencent	2.94T tokens ↓3%	6.	MiMo -V2.5 by xiaomi	1.7T tokens ↑>999%
2.	DeepSeek V4 Flash by deepseek	2.88T tokens ↓20%	7.	MiMo -V2.5 -Pro by xiaomi	1.4T tokens ↑299%
3.	Claude Opus 4.7 by anthropic	2.2T tokens ↑5%	8.	DeepSeek V4 Pro by deepseek	1.26T tokens ↑11%
4.	Claude Sonnet 4.6 by anthropic	1.92T tokens ↓0%	9.	DeepSeek V3.2 by deepseek	1.09T tokens ↑18%
5.	Owl Alpha by openrouter	1.76T tokens ↑51%	10.	Gemini 3 Flash Preview by google	954B tokens ↓12%

Can We Measure Swarms?

AI Engineer

Code Summit



Context Platform Engineering

TO REDUCE TOKEN ANXIETY

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



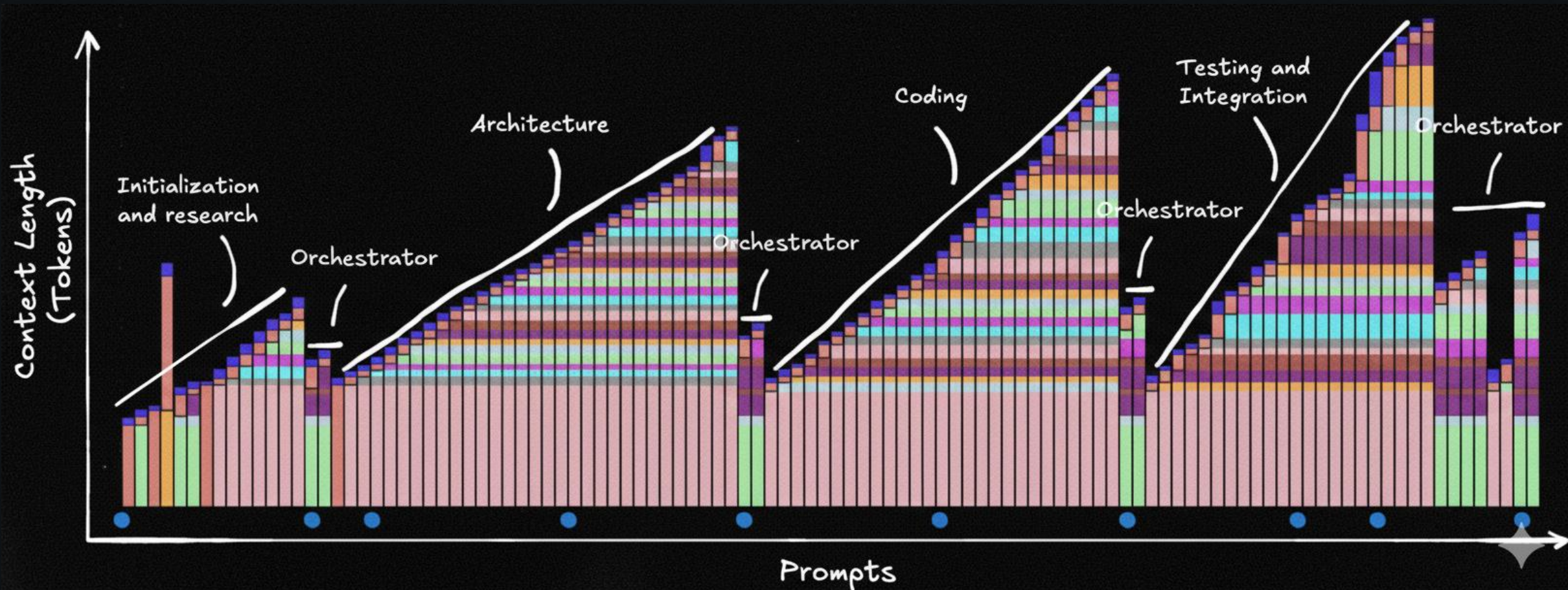
[KV Cache Observability 101](#)

[KV Cache Observability 201](#) (open source repo by SemiAnalysis InferenceX & Nvidia Dynamo AIPerf)

[KV Cache Observability 301](#) (Inference Provider Infra)

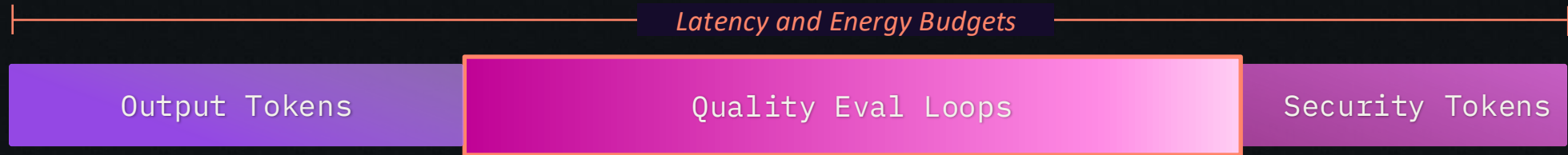
/goal

Claude Workflows



Enterprise Agents '3x Rule'

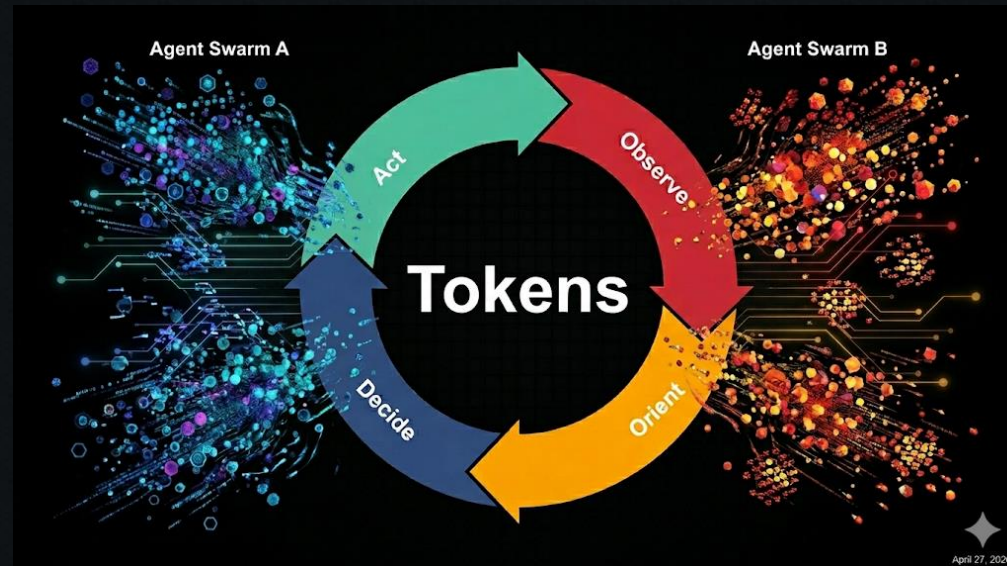
Output tokens are not enough. Quality and Security within Latency and Energy Budgets are the new goal



Can We Afford Persistent AI?

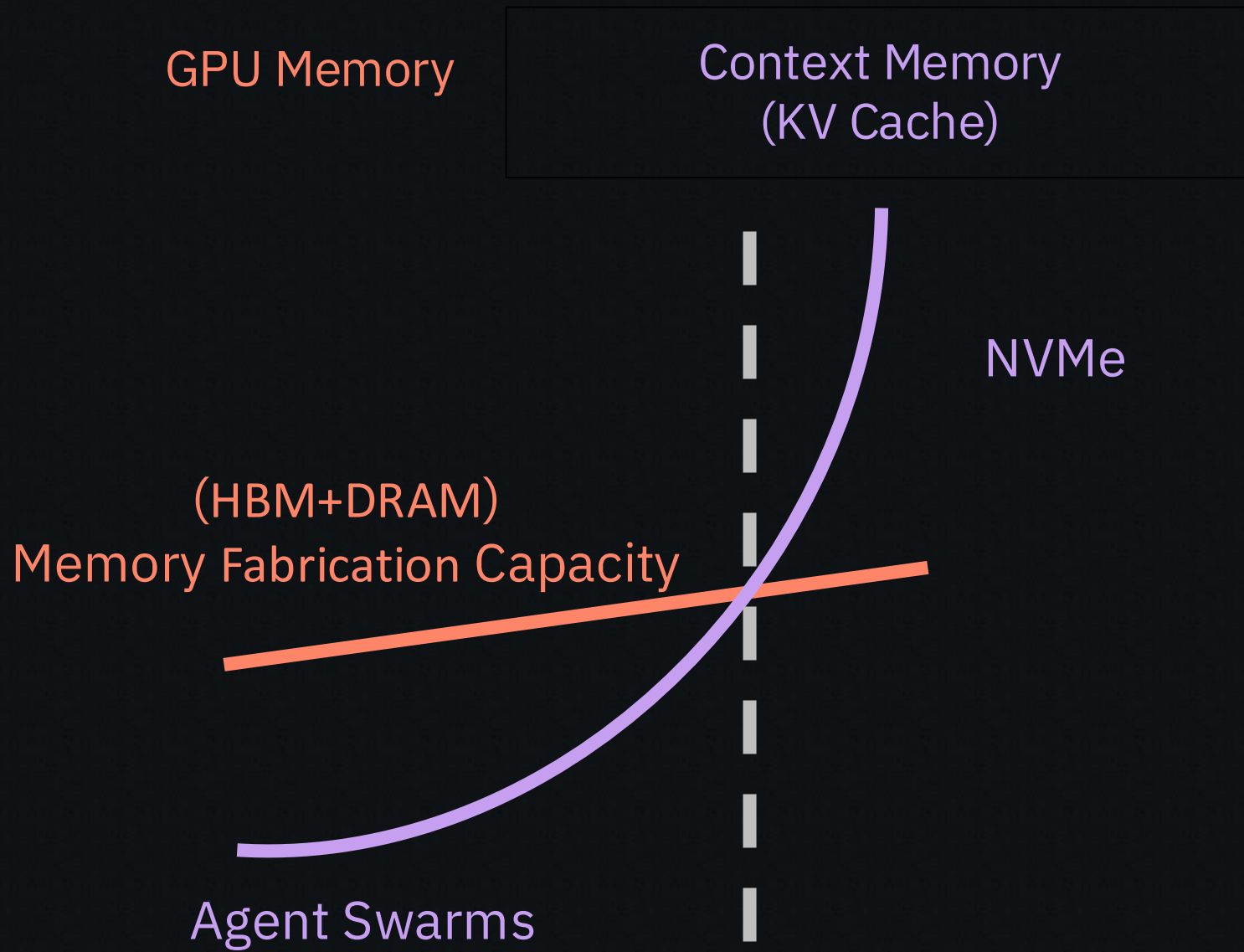


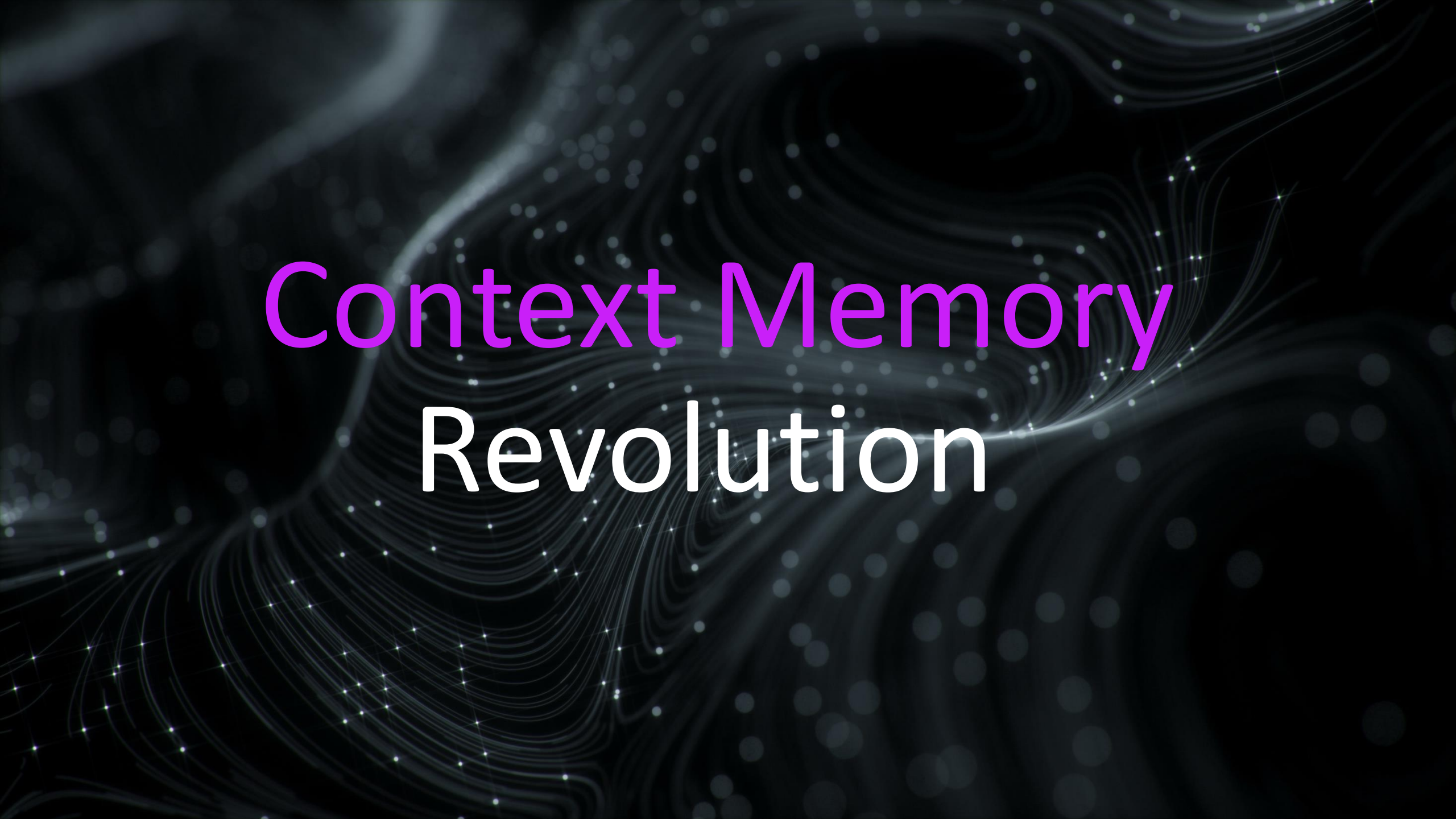
Annualized Spend (Claude Opus 4.8 / GPT 5.6?)



April 27, 2026

The AI Memory Wall



The background is a dark, almost black, space filled with intricate, glowing patterns. These patterns consist of numerous thin, curved lines that swirl and flow across the frame, creating a sense of dynamic movement. Interspersed among these lines are many small, bright white dots and larger, fainter circular bokeh-like shapes, which contribute to a futuristic and high-tech aesthetic. The overall effect is reminiscent of a complex data visualization or a stylized representation of a neural network or energy field.

Context Memory Revolution

Context Memory 'Storage Must be Rearchitected'



Context is the New Bottleneck -
Storage Must be Rearchitected

- HBM
- Memory
- Rack SSD
- Network SSD

Model Size Growing
Context Length Growing
Multi-Turn Conversations - Context Accumulates
More Concurrent Users and Sessions

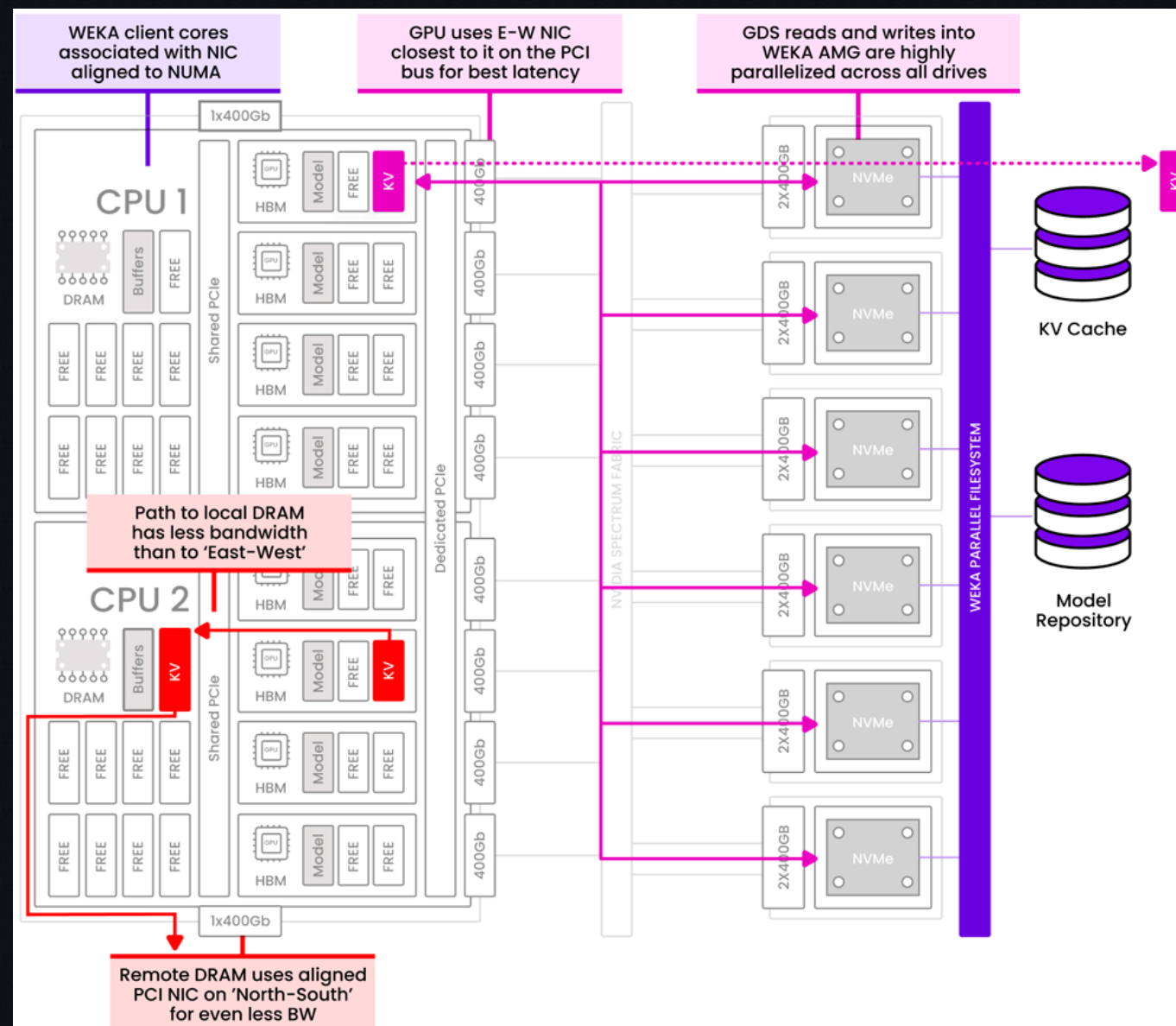
"For storage, that is a completely unserved market today. This is a market that never existed, and this market will likely be **the largest storage market in the world**, basically holding the working memory of the world's AIs."

-Jensen Huang @ CES 2026

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](#)



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.



Dynamo NIXL

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

RAW RDMA Performance

Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util	Compute M.
0	NVIDIA	H200				MIG M.
N/A	41C	P0	128W / 700W	00000000:19:00.0 Off 127132MiB / 143771MiB	0%	0 Default Disabled
1	NVIDIA	H200				
N/A	44C	P0	131W / 700W	00000000:3B:00.0 Off 127132MiB / 143771MiB	0%	0 Default Disabled
2	NVIDIA	H200				
N/A	43C	P0	132W / 700W	00000000:4C:00.0 Off 127132MiB / 143771MiB	0%	0 Default Disabled
3	NVIDIA	H200				
N/A	42C	P0	128W /			
4	NVIDIA	H200				
N/A	41C	P0	128W /			
5	NVIDIA	H200				
N/A	43C	P0	127W /			
6	NVIDIA	H200				
N/A	42C	P0	128W /			
7	NVIDIA	H200				
N/A	43C	P0	128W /	00000000:DB:00.0 Off 127132MiB / 143771MiB	0%	0 Default Disabled

```
Every 0.1s: weka status | ... : Mon Jan 19 08:09:12 2026

reads: 354.23 GiB/s (362739 IO/s)
writes: 0 B/s (0 IO/s)
```

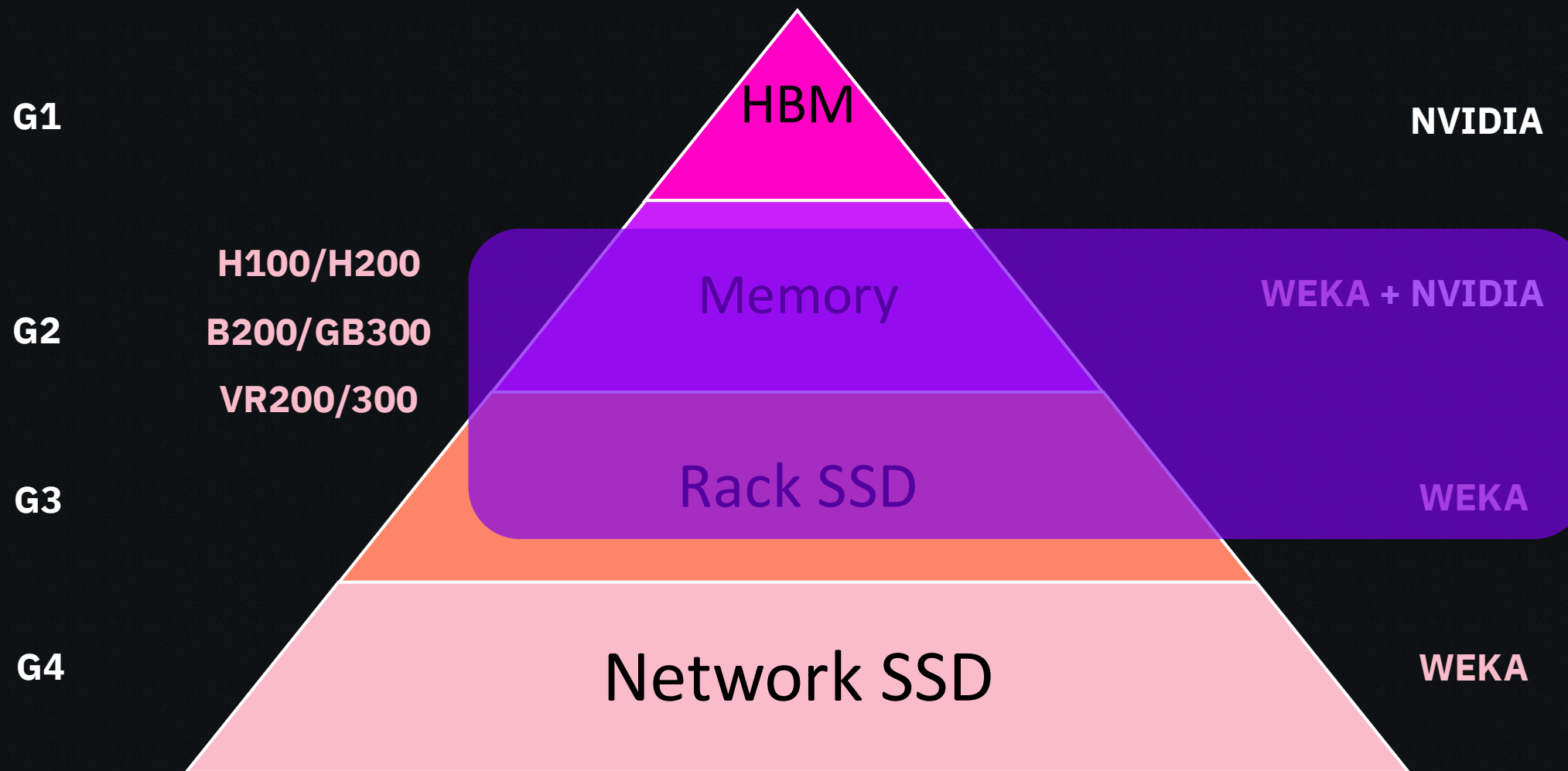
```
root@ [REDACTED] ~#
root@ [REDACTED] ~# ./rdma_inference_sustained.sh
Reading from WEKA AMG over 8 GPUs

Throughput: 353.90 GiB/s (3.04Tb/s)
```

```
#
#
#
#
#
#
#
# ./rdma_inference_sustained.sh
Reading from WEKA AMG over 8 GPUs

Throughput: 353.90 GiB/s (3.04Tb/s)
```


Memory Storage Hierarchy

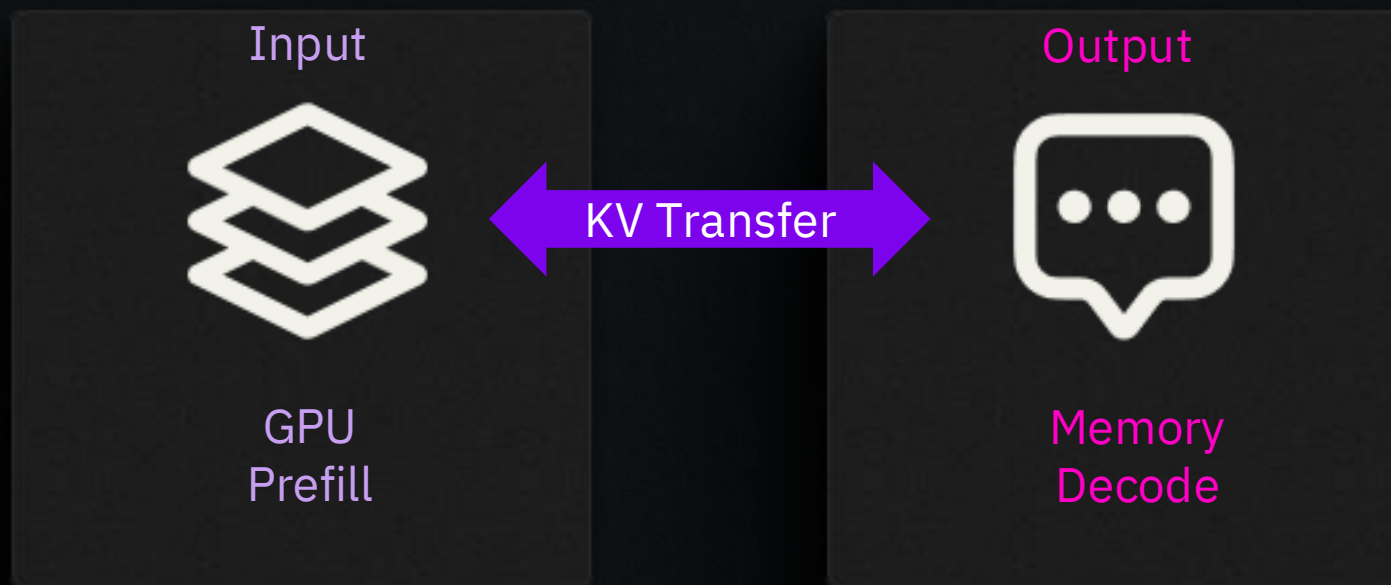




GPU
Prefill

Token Warehousing

Factory → Consumer?

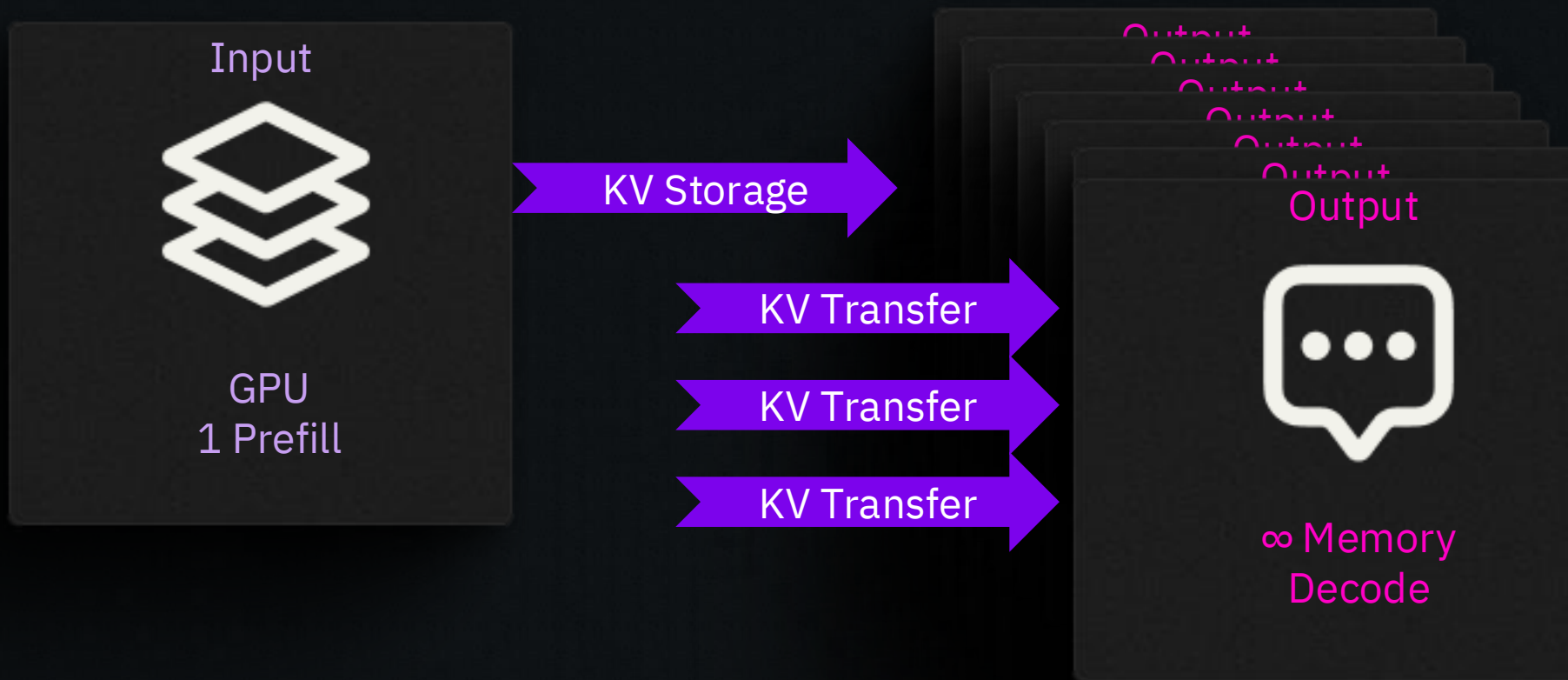




GPU
Prefill

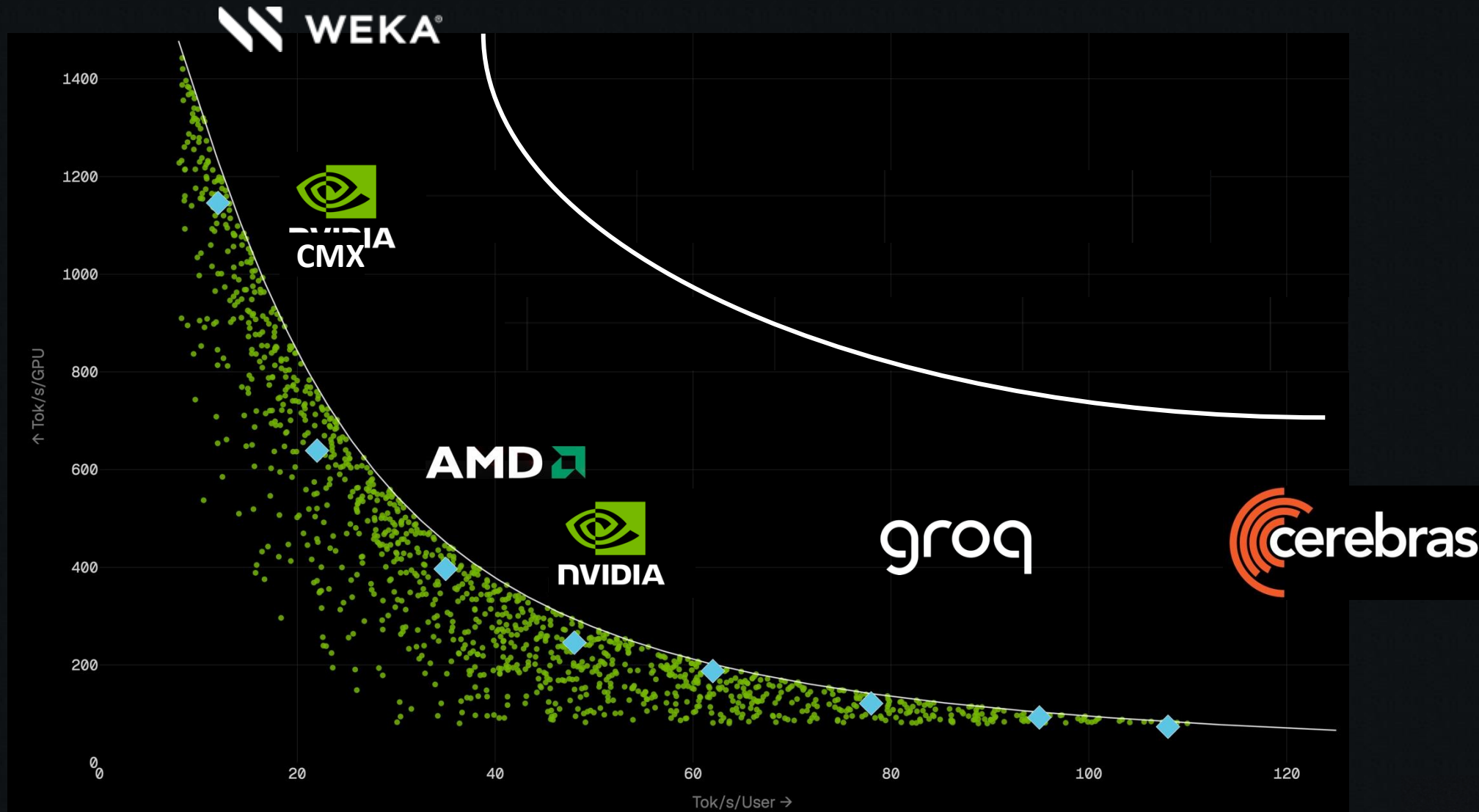
Warehousing Attention

Attention Warehouse → Output



Attention 

Pareto Frontier




Output



Independent Proof Points



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

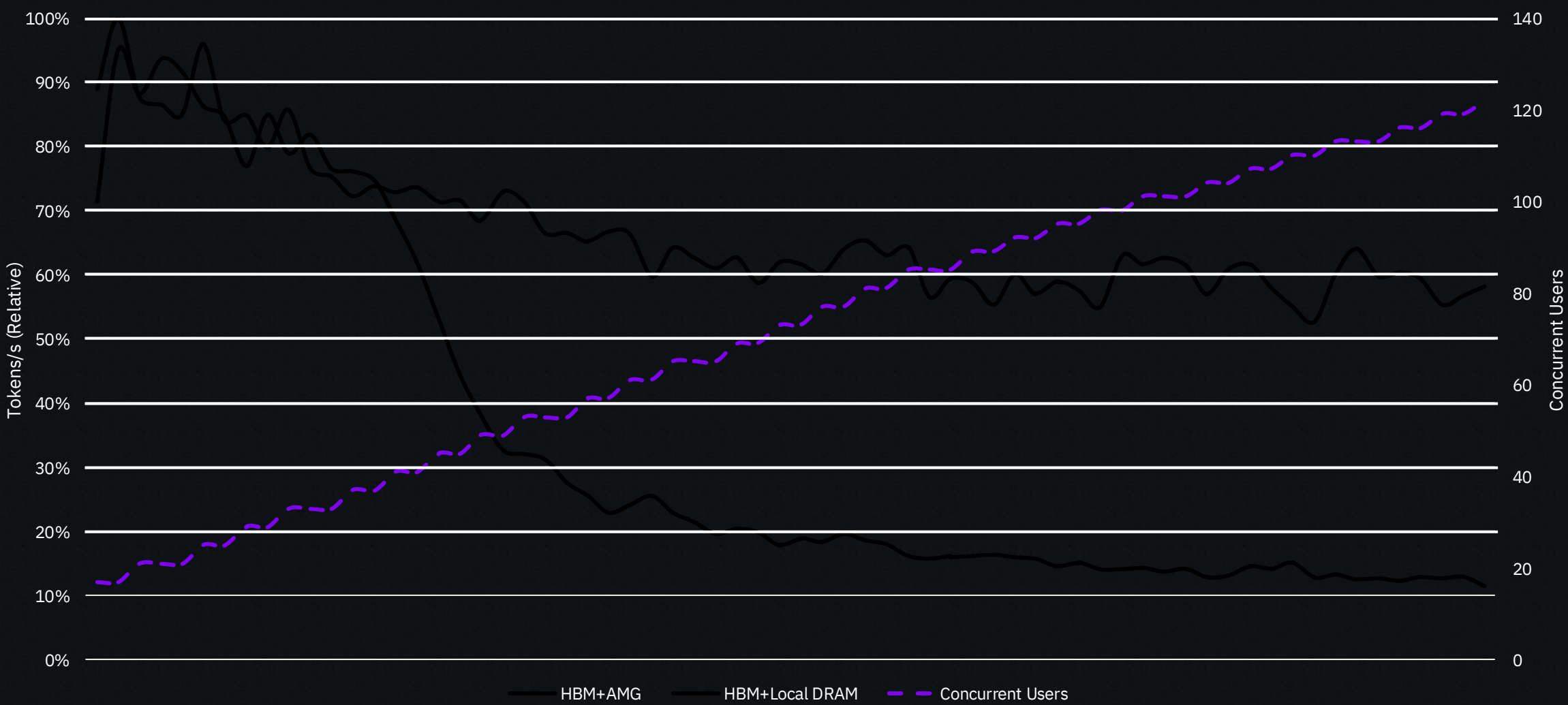
Metric	DRAM Baseline	Augmented Memory Grid	Improvement
Max concurrent users	~600	5,000+	10X
Requests completed, 2,400-user test	~6,700	47,000+	7X
Tokens served	700M	5B	7X
Token throughput	<200K tokens/sec	~2M tokens/sec	10X

The background of the slide features a dark, semi-transparent overlay of financial data. At the top, there is a candlestick chart with several price bars and a moving average line. Below this, a bar chart with numerous vertical bars of varying heights is visible, representing time-series data. The overall aesthetic is professional and data-driven.

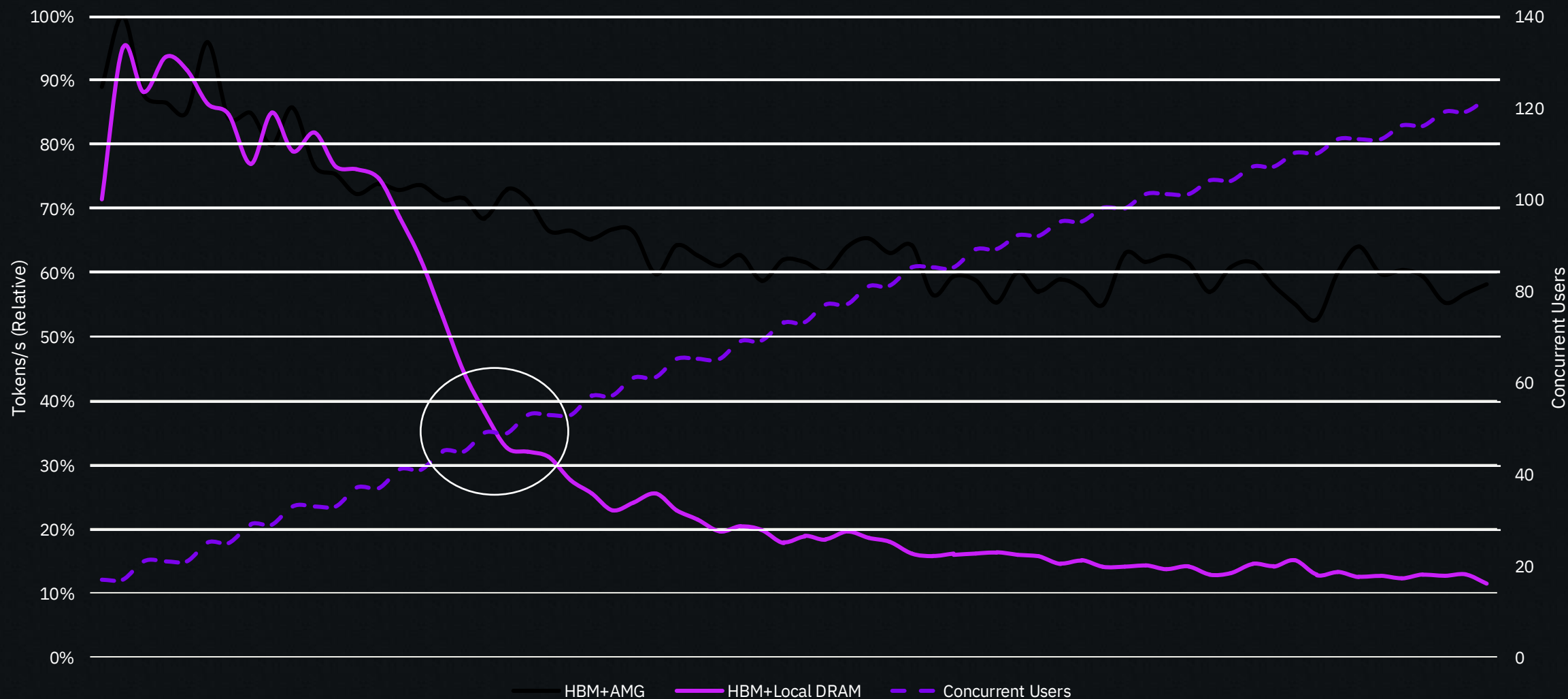
Performance Data Real-World Insights

18-125 Concurrent Session

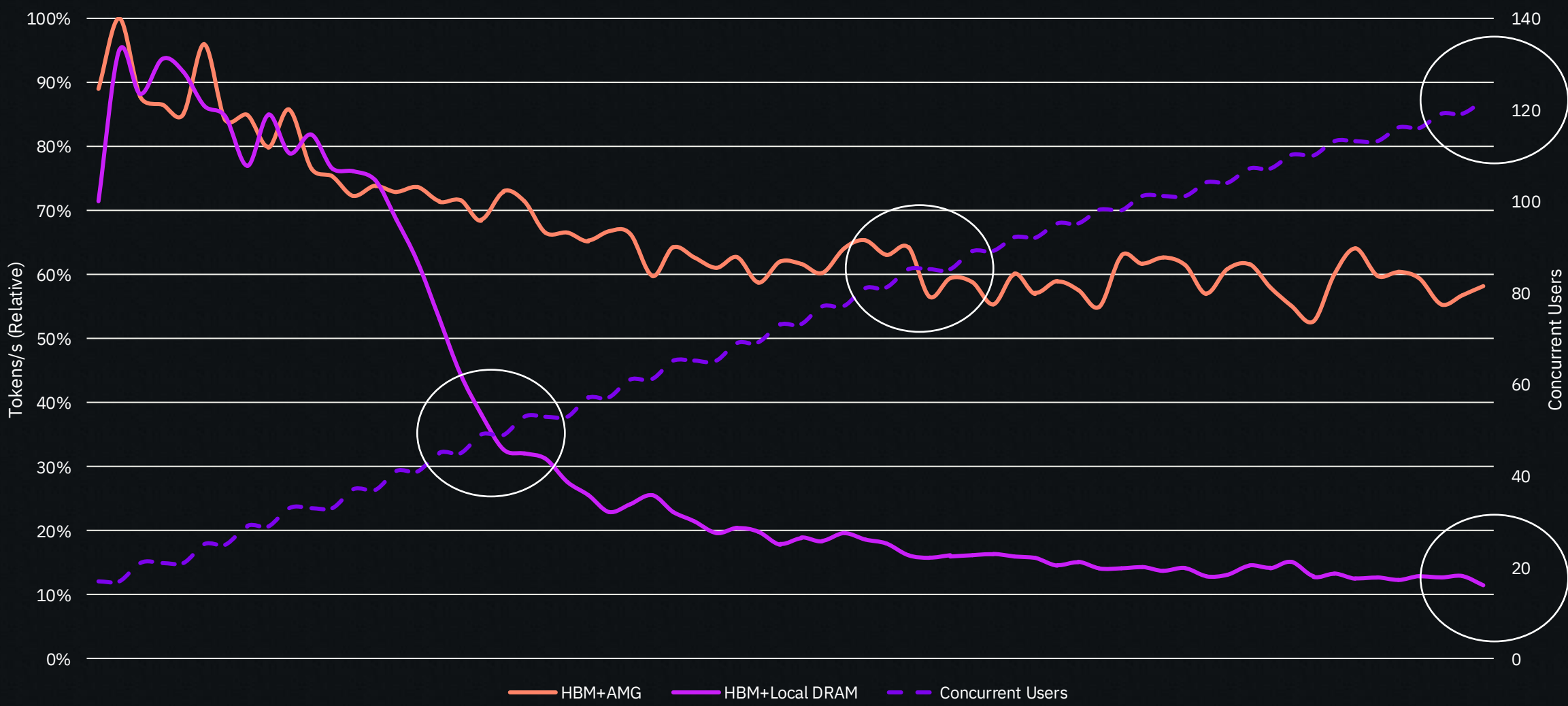
Qwen3-Coder-480B Inference Performance



GPU HBM/VRAM + CPU DRAM = Performance Cliff



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



5-10 Instant **NEW** Data Centers

Physical



Data
Center

WEKA



Virtual Data
Center

WEKA



Virtual Data
Center

WEKA

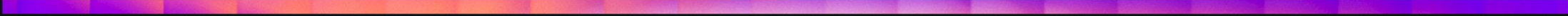


Virtual Data
Center

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

