The Token Economy:
Performance as the Profit Driver in AI

HPC-AI Swiss Conference, April 21

Axel Rosenberg
Systems Engineering DACH

WEKA®

WHERE ARE YOU ON YOUR AI JOURNEY?

From connectivity to cognition — match your infrastructure posture to your AI ambition.

Internet Era 1992–2002 | Virtualisation 2002–2012 | Cloud Era 2012–2022 | AI Training 2020–Now | AI Inference & Agentic AI

MAJORITY

Exploring AI

Traditional IT still dominant.
AI use-cases under evaluation.

INFRASTRUCTURE PROFILE

- SAN / NAS storage
- VMware / on-prem servers
- AWS / Azure Deployments
- First GPU pilots (1–4 GPUs)
- Evaluating LLM APIs & tools

KEY FOCUS

Latency of first AI results

WEKA: AI-ready storage entry point in the Cloud or On-Prem

EARLY ADOPTERS

Scaling AI

GPU clusters in production.
Model training & inference active.

INFRASTRUCTURE PROFILE

- GPU fabric (10s–100s of GPUs)
- Local NVM or legacy shared storage
- Kubernetes / containerised
- Internal model serving & RAG

KEY FOCUS

IOPS · Throughput · Token cost

WEKA: no-copy pipelines, high-IOPS

INNOVATORS

Leading with AI

GPU at scale. Agentic AI &
token economies in production.

INFRASTRUCTURE PROFILE

- Exa-scale GPU infrastructure
- Agentic AI workflows
- Token warehouse / vector stores
- Safety, governance & FinOps

KEY FOCUS

Agent autonomy · Token economy

WEKA at scale: multi-tenant, KMS, S3

Latency — inference SLO · IOPS — NVMe pipelines · Token Economy — cost/call · Agentic AI — autonomous ops · Governance — multi-tenant KMS

WEKA

The AI Journey

Where are you today? Where do you need to be?

1

Cloud & Infra-as-Code

Modernised infrastructure

2

Data Preparation

ETL, vectorisation & tagging

3

AI Training

Hyperscalers & model builders

4

AI Inferencing

Production AI services at scale

5

Data Services & Insights

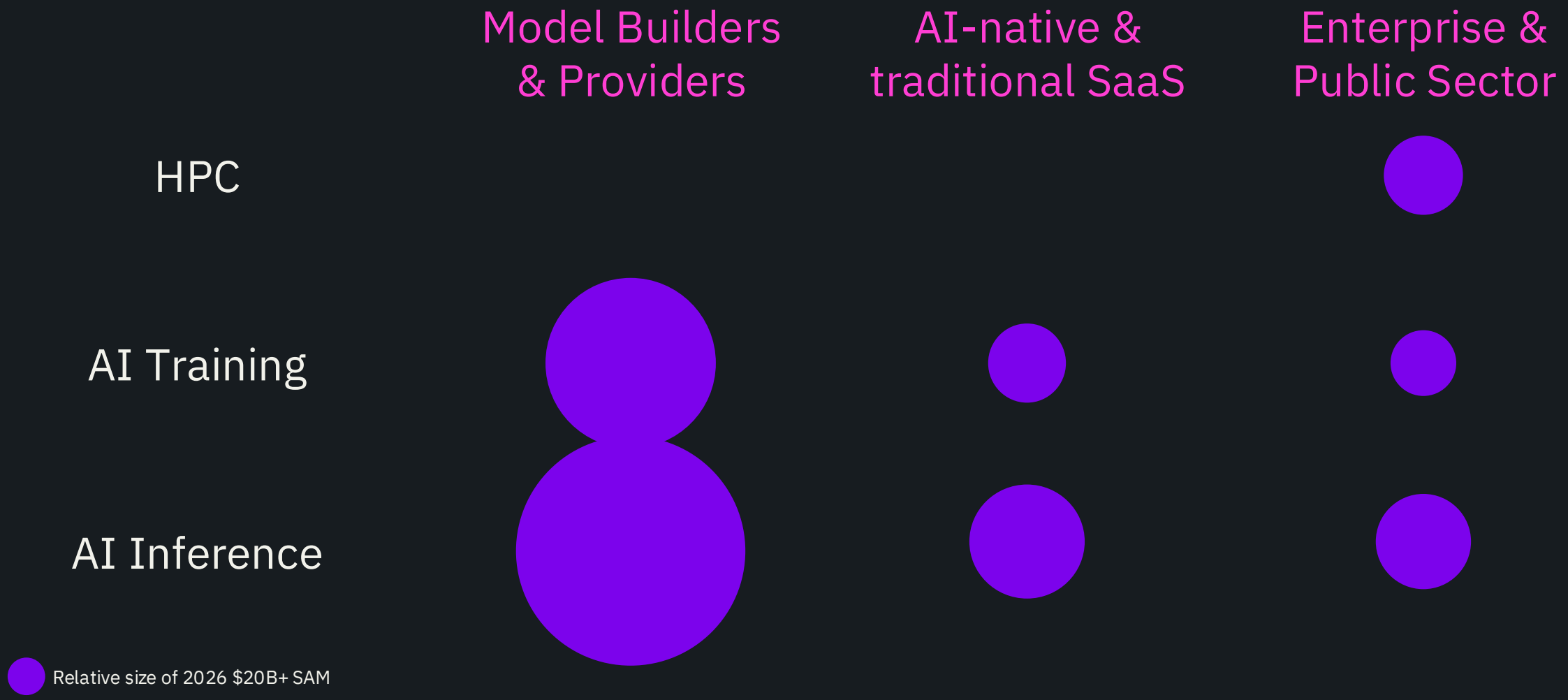
Value-added AI outcomes

Skipped by most enterprises

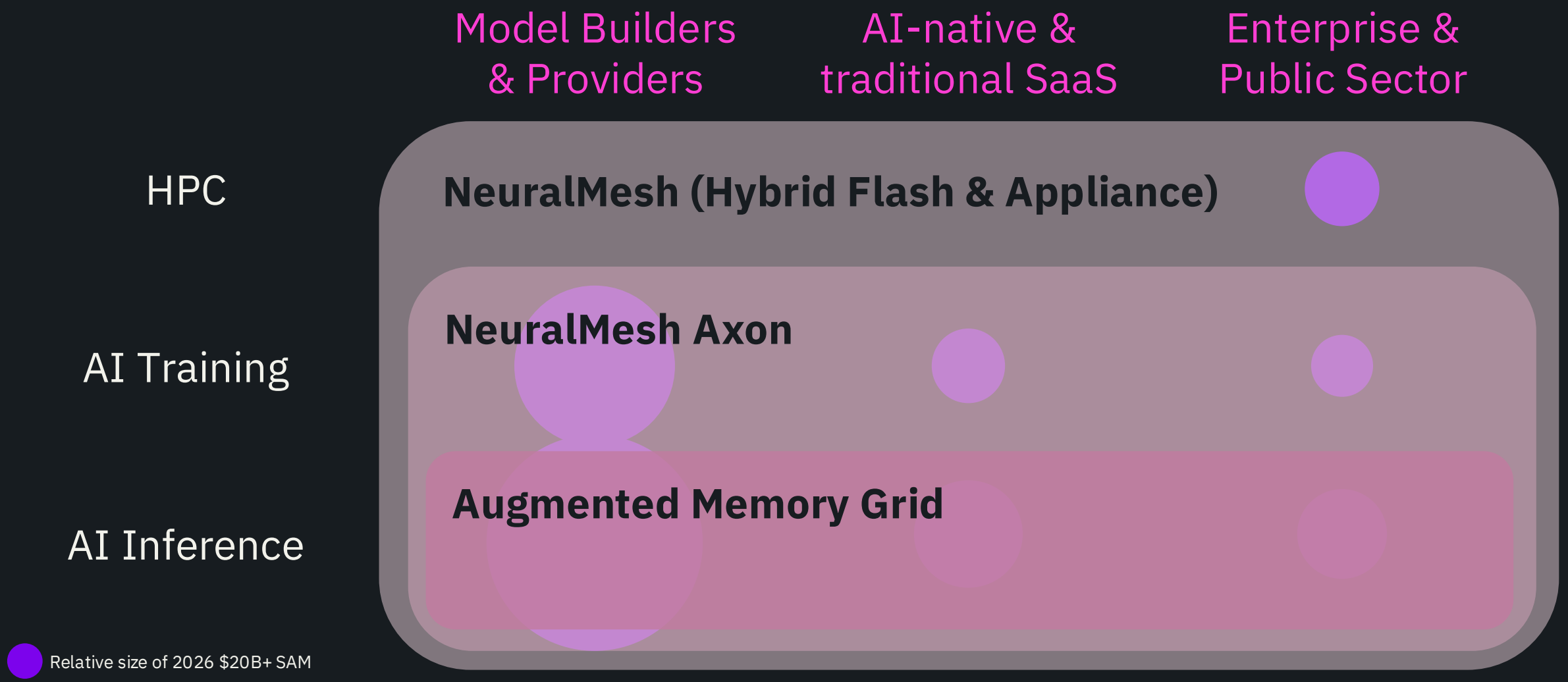
60–80% of AI pipeline time is spent on data preparation. Enterprises spend 15–24 months creating a single AI pipeline. Data preparation is where the journey begins — and where WEKA accelerates it.

WEKA®

Inference overtaking training in $20B market in 2026



WEKA product portfolio covers all market segments



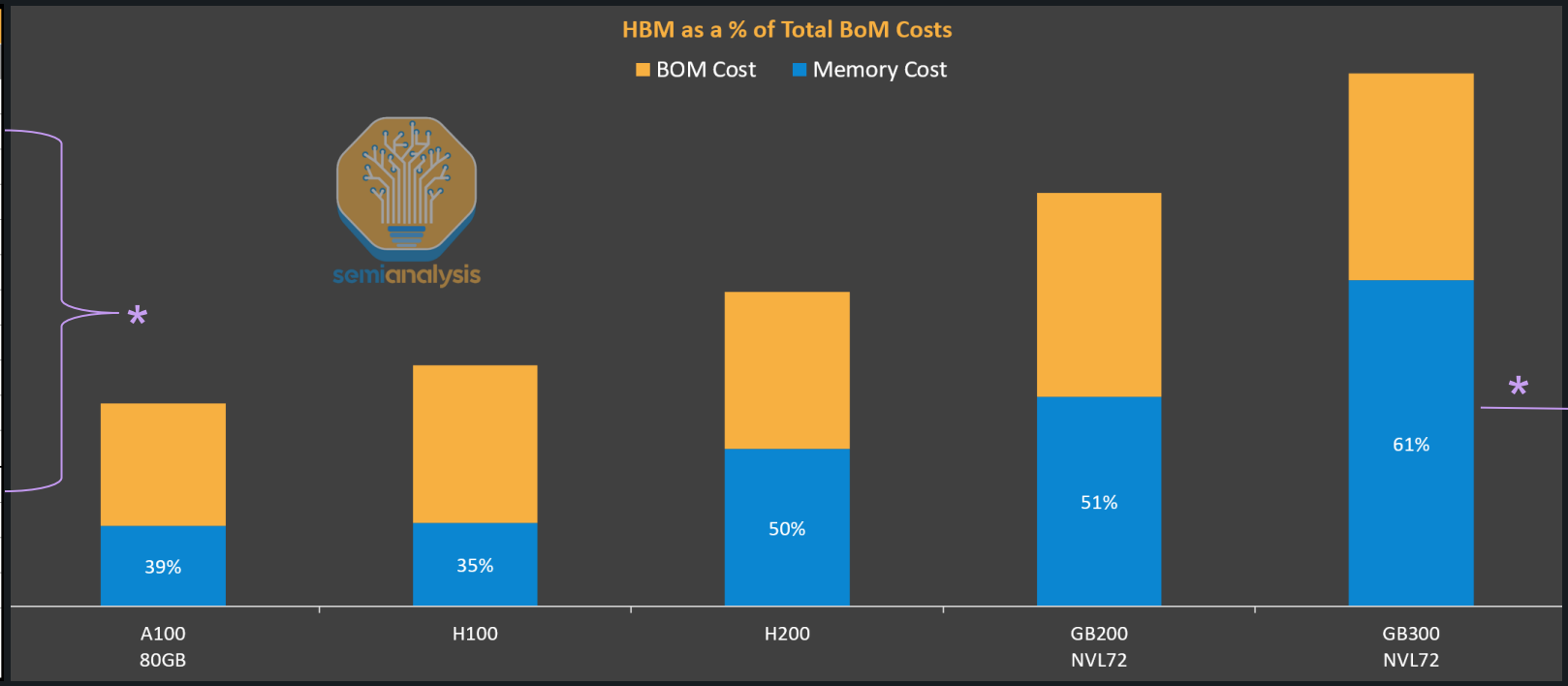
AI Infrastructure – Value Segmentation – WEKA GPU Value*

Report Link

Nvidia DGX H100	
Component	AI Server
CPU	$ 5,200
* 8 GPU + 4 NVSwitch Baseboard	$ 195,000
Memory	$ 7,860
Storage	$ 3,456
SmartNIC	$ 10,908
Chassis (Case, backplanes, cabling)	$ 563
Motherboard	$ 875
Cooling (Heatsinks+fans)	$ 463
Power Supply	$ 1,200
Assembly and Test	$ 1,485
Markup	$ 42,000
Total Cost	$ 269,010
DRAM BOM %*	2.9%
* NAND BOM %	1.3%
Memory BOM %*	4.2%

*HBM Costs are not included as they are part of the GPU. We have broken the Nvidia BOM below for subscribers.

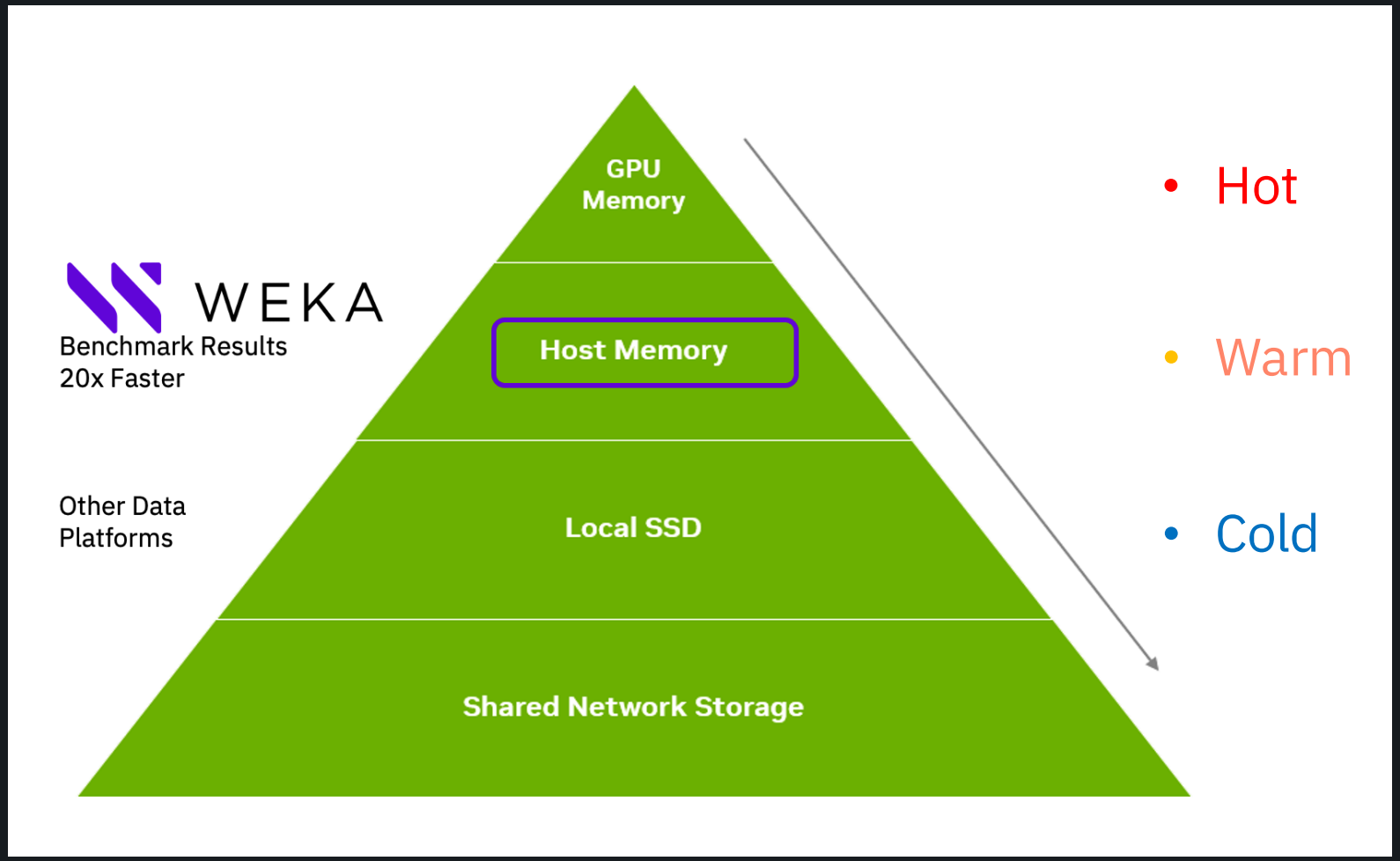
Report Link



WEKA®

WEKA®

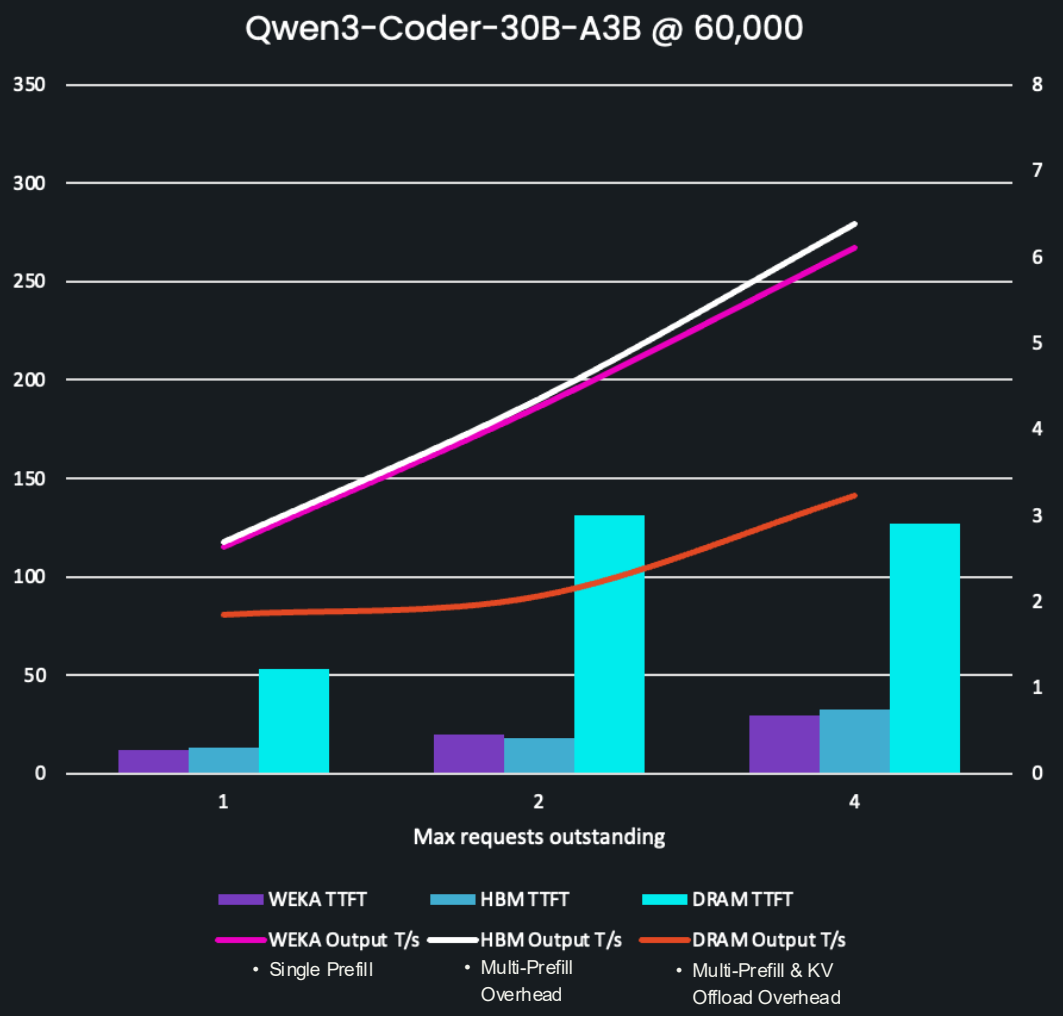
vLLM or Dynamo - KV Cache Hierarchy

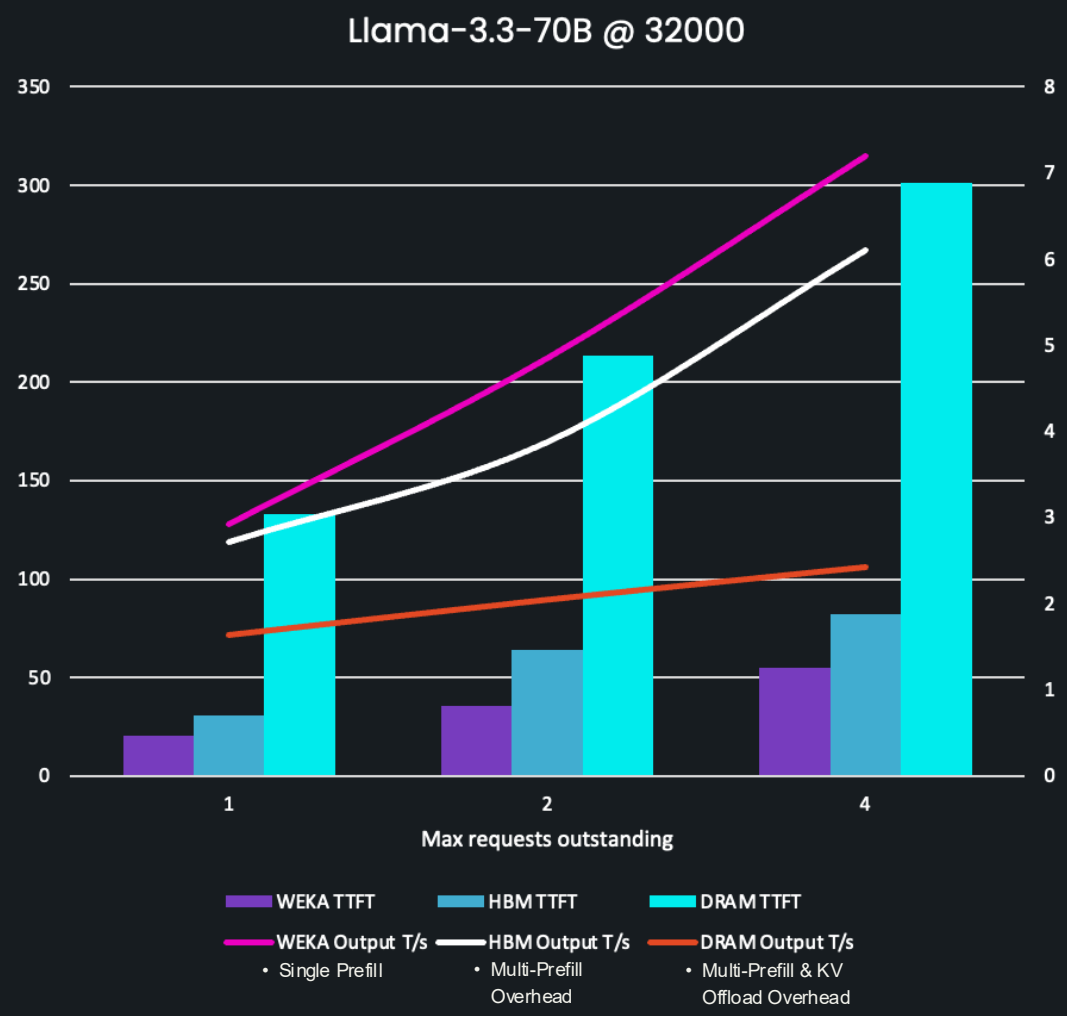


https://developer.nvidia.com/blog/introducing-nvidia-dynamo-a-low-latency-distributed-inference-framework-for-scaling-reasoning-ai-models/

AMG Scaling HBM performance for TTFT & tokens/sec

NCP Tests-Weights & Biases Inference (Multi-User, HBM+DRAM KV Cache Saturated)





WEKA®

5 AI Metrics that Matter

Token Throughput (Tokens per Second)

TTFT: Time to First Token

End to End Latency (TTFT + Token Throughput)

Cost per Token

Context Window

semianalysis

WEKA®

3 Token Prices

INPUT

75% discount

CACHED INPUT

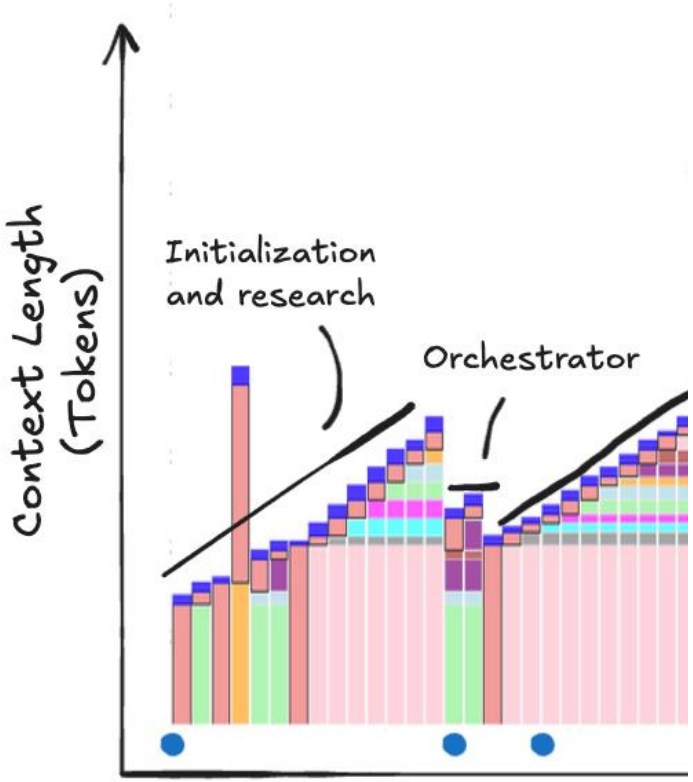
5-15 minutes

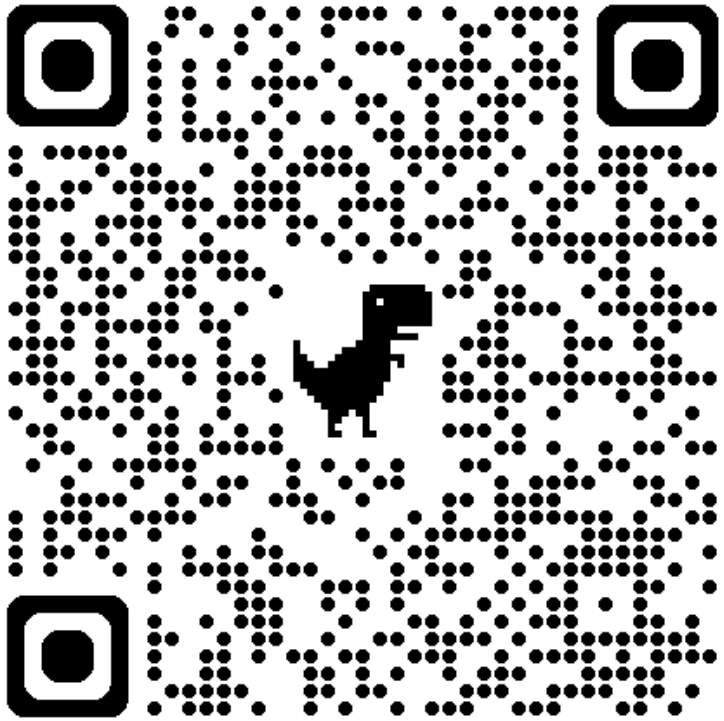
OUTPUT

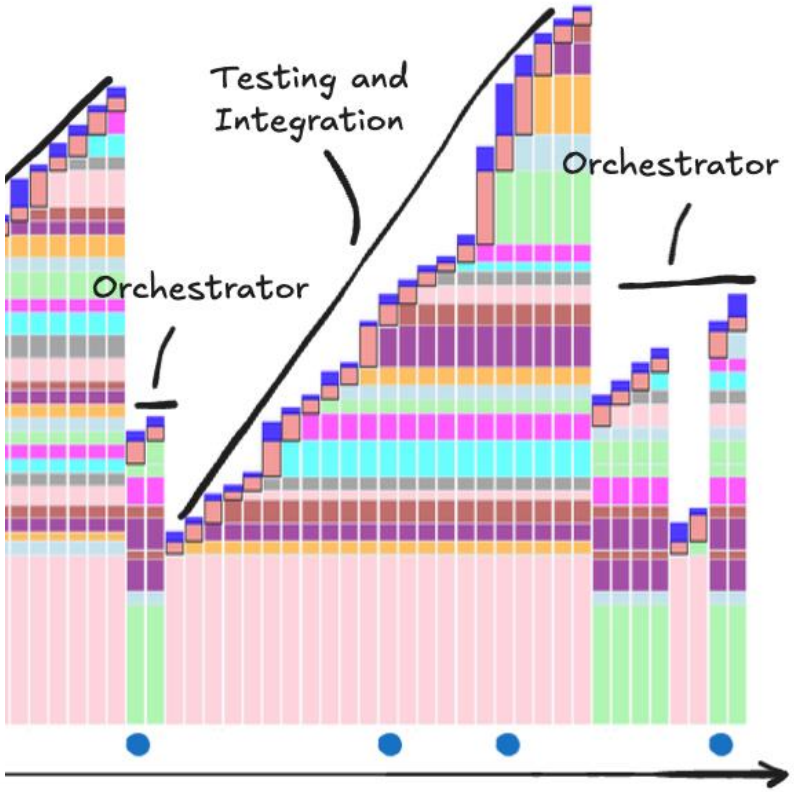
semianalysis

WEKA®

Agents & KV Cache





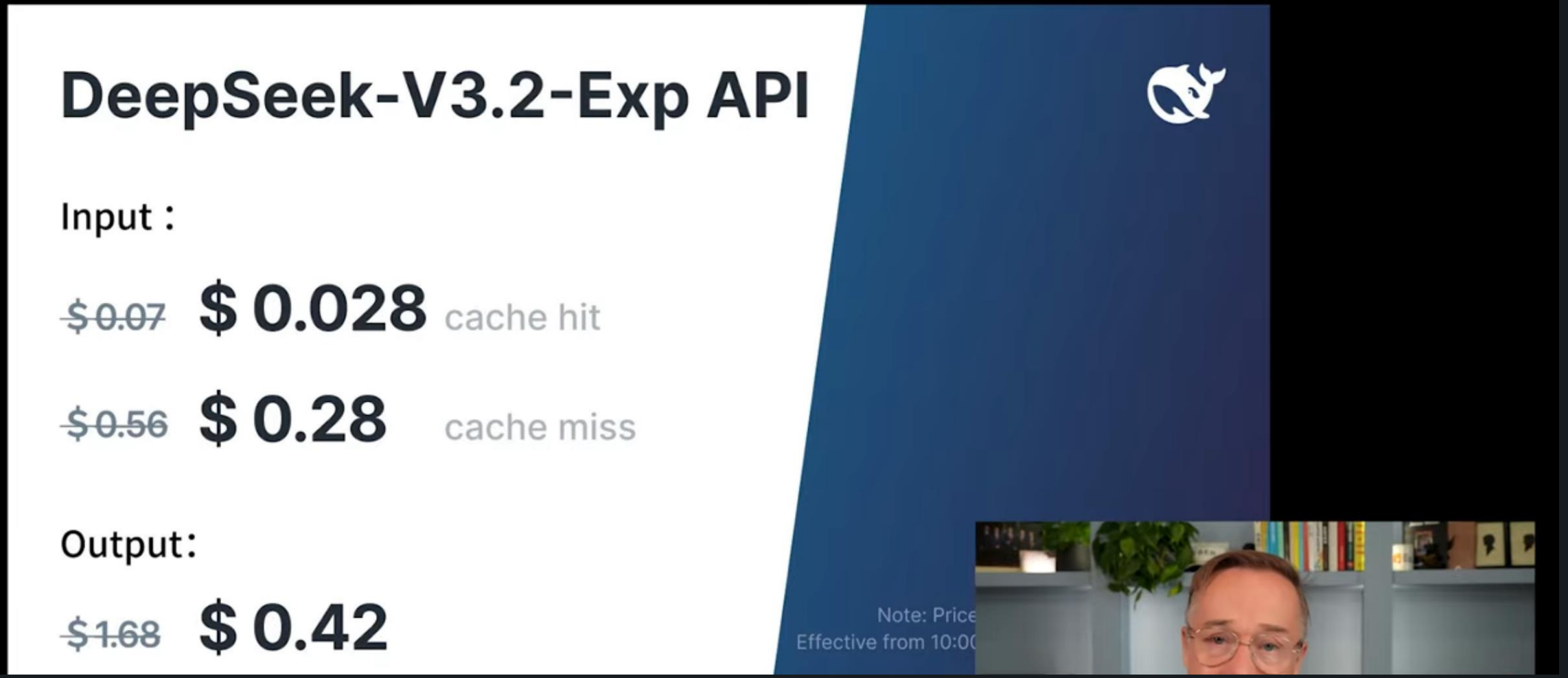


Prompts

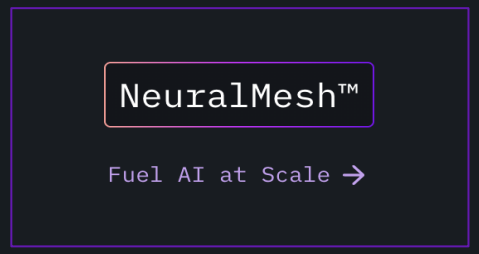
https://www.linkedin.com/pulse/visual-guide-how-ai-agents-use-inference-inside-llm-callan-fox-q9brc/

WEKA®

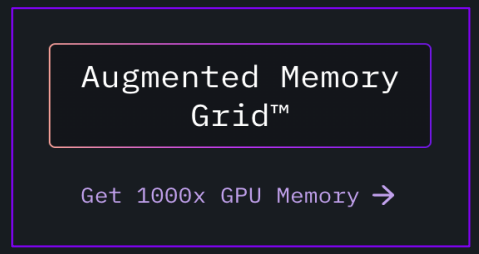
Token Pricing Example – DeepSeek Oct 2025



WEKA Product Portfolio

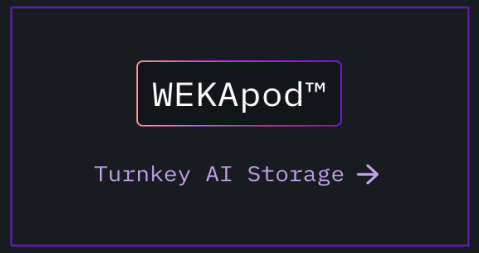


ExaScale, Multi-Protocol Storage

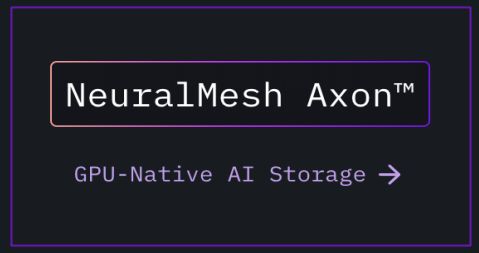


Persistent Inference Memory

WEKA AI Extensions



NeuralMesh Appliance



GPU-Native AI Storage

WEKA AI Nodes

WEKAPod.NEXT

Cohere Uses NeuralMesh Axon For Exascale AI Deployments

INDUSTRY AI/ML Platform & Infrastructure
USE CASE AI/ML, Cloud, Generative AI, Hybrid Cloud, GPU Acceleration
REGION Global

"

"Embedding WEKA's NeuralMesh Axon into our GPU servers enabled us to maximize utilization and accelerate every step of our AI pipelines."

Autumn Moulder
VP of Engineering, Cohere

Context

Cohere, a leading enterprise AI company, needed to accelerate AI model training and inference workloads while maximizing GPU utilization and reducing infrastructure costs across their public cloud deployments.

Challenge

Cohere faced high innovation costs, data transfer bottlenecks, and underutilized GPUs that slowed model development and increased time-to-market for new AI solutions like their North secure AI agents platform.

The Power of WEKA

By embedding WEKA's NeuralMesh Axon into their GPU servers on CoreWeave Cloud, Cohere unified their AI stack and achieved breakthrough performance improvements across their entire AI pipeline, enabling faster iteration and model deployment.

15 seconds
for inference deployments
(reduced from 5 minutes)

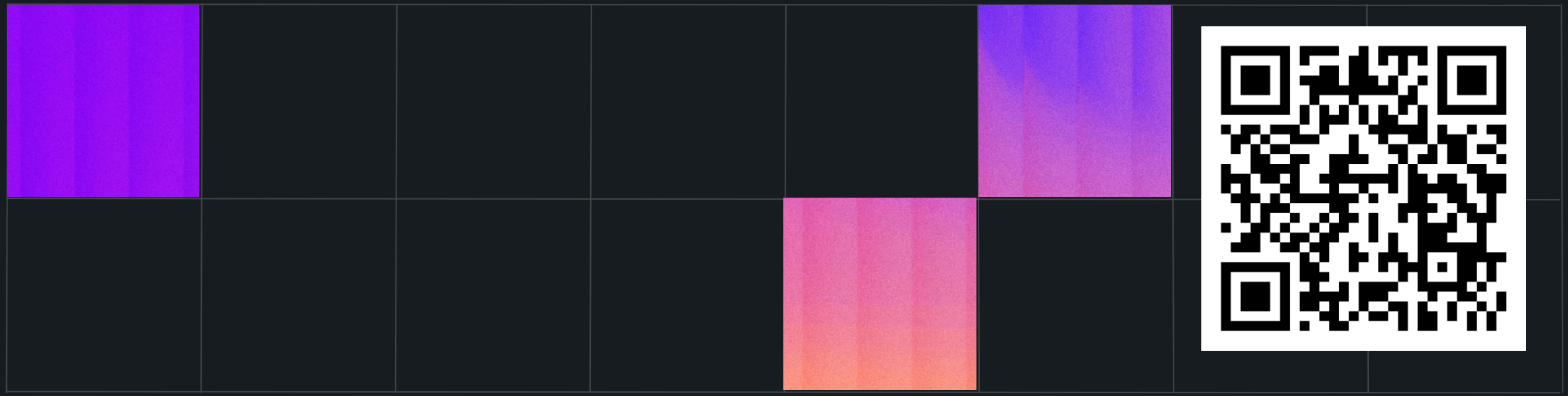
10x faster
checkpointing performance

THANK YOU

Come Meet Us At Our Stand

NEURALMESH BY WEKA

SCAN TO LEARN MORE



WEKA®