



Accelerate Your AI Projects With The WEKA Data Platform

Kevin Tubbs, PhD

Field CTO, HPC & AI

Modern AI Workloads

WHAT'S CHANGING?

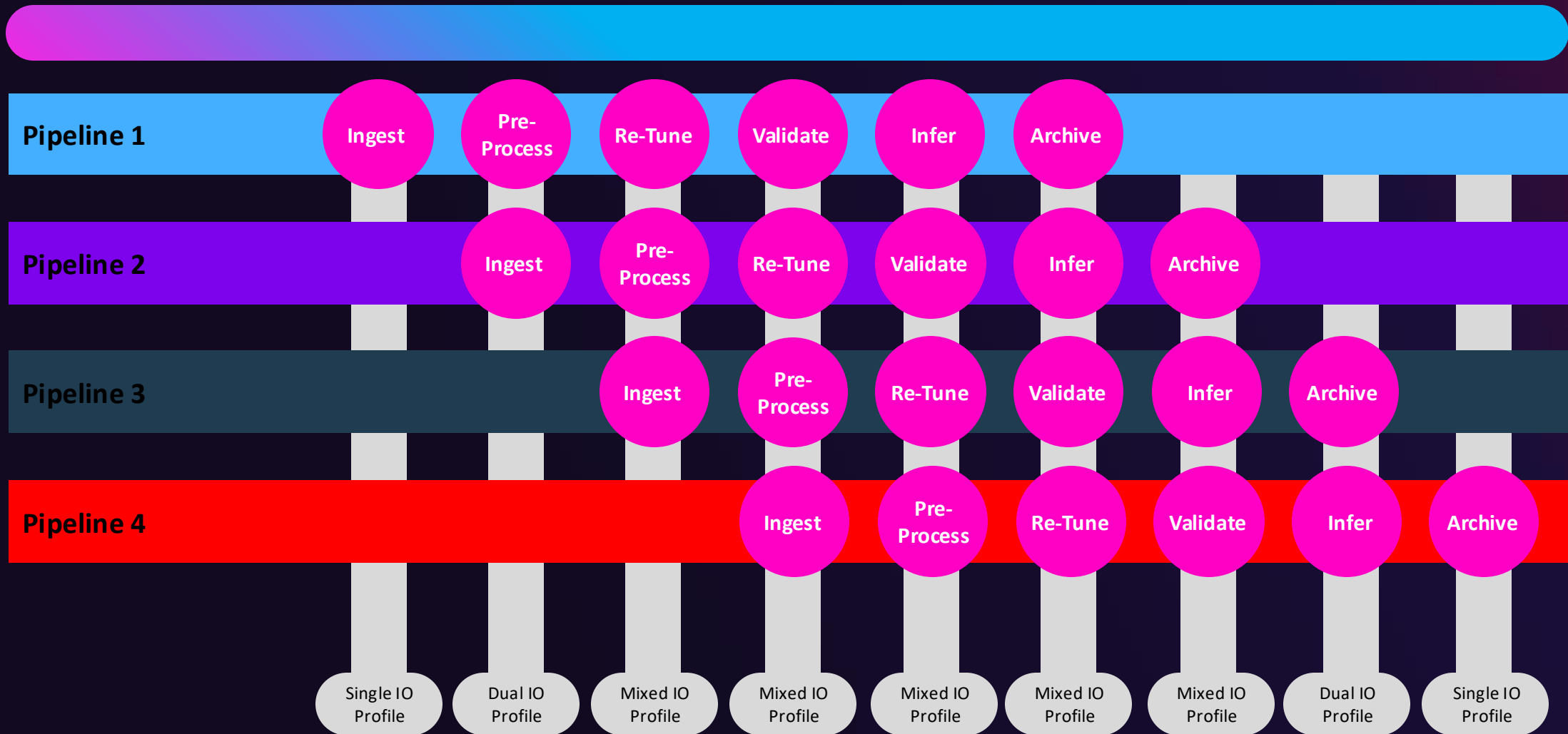
FROM

- AI Feasibility Projects
- Practitioner Grade AI
- Petabyte Scale
- Foundational Models

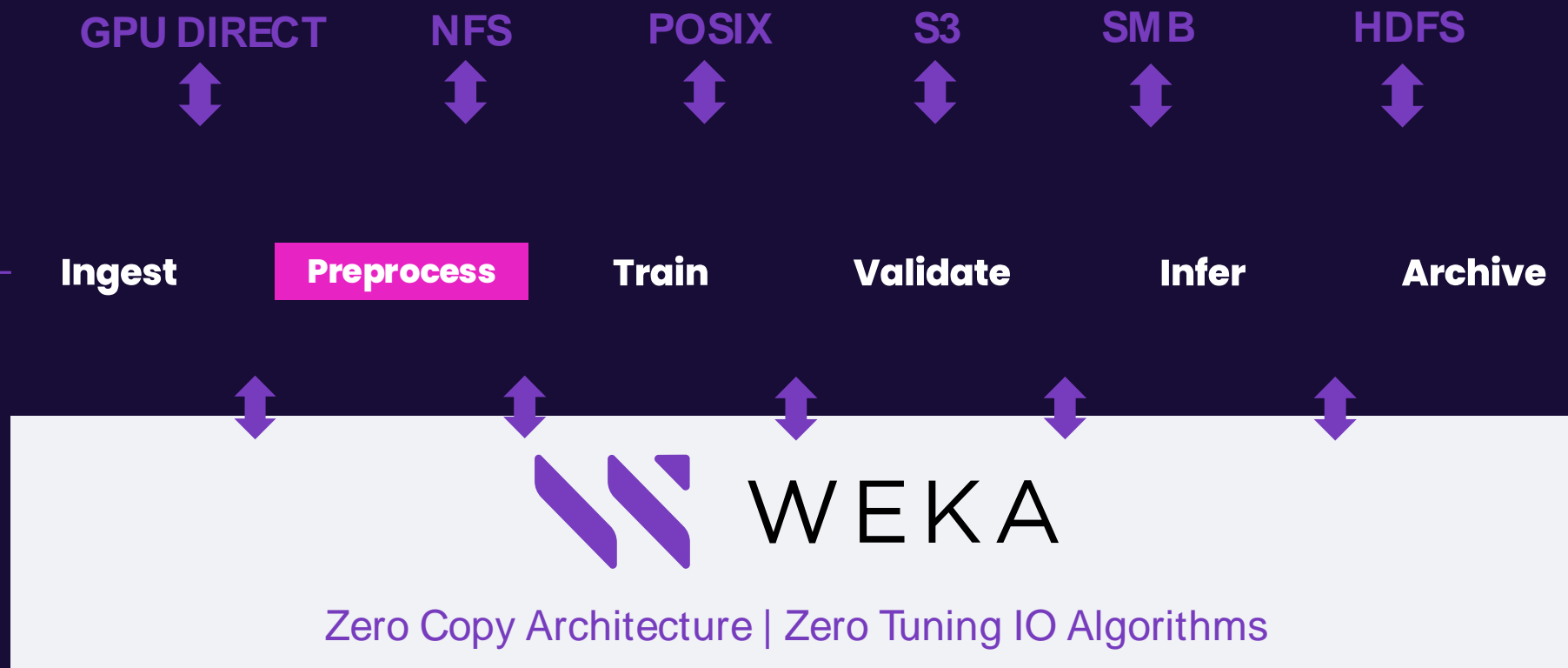
TO

- Core Operational Assets
- Enterprise Grade AI
- Exascale
- Pre-trained Models with Domain-specific tuning

The data challenge – multiple concurrent IO patterns to the same data




WEKA Data Platform: 10–100x faster business outcomes for next gen workloads



The importance of Latency

Every 100 milliseconds of latency is about 8% difference in engagement rates

 **Jonathan Ross** · Following
CEO @ Groq, the Most Popular API for Fast Inference | Creator of the TP...
[Visit my website](#)
2w · 🌐


One of the key reasons why Groq is so popular to developers:

Every 100 milliseconds of latency is about 8% difference in engagement rates.

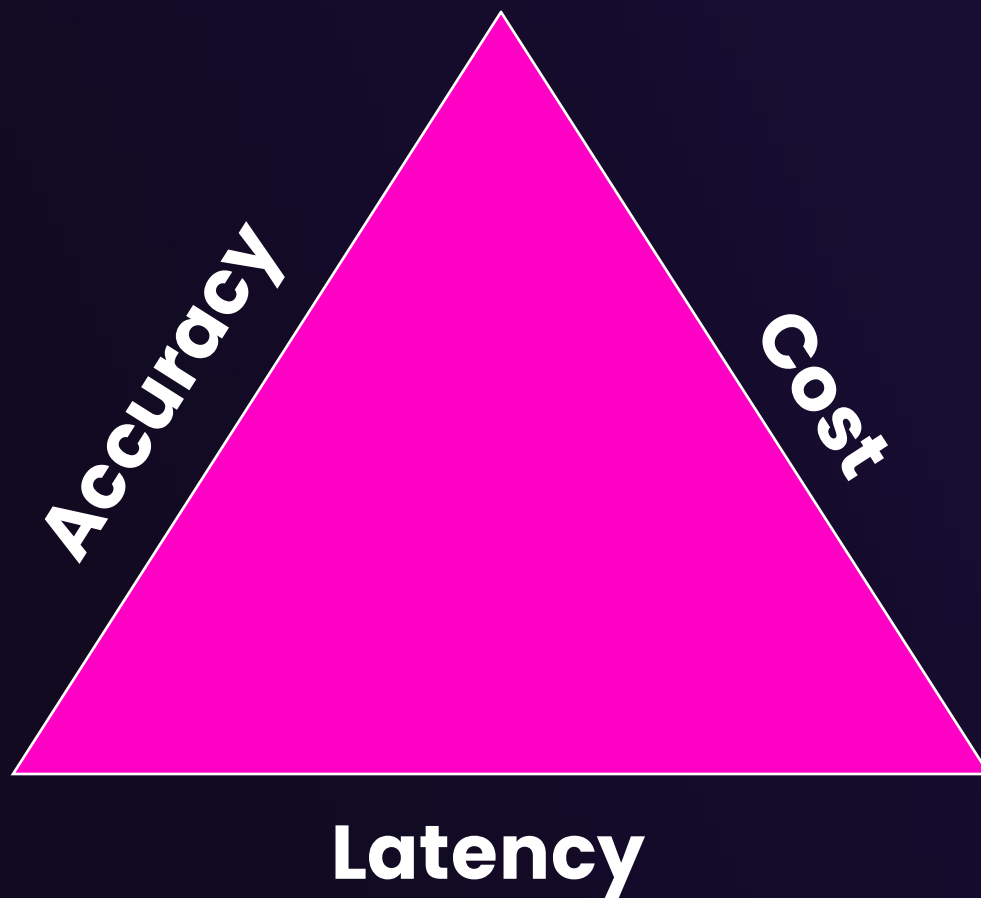
On mobile, where people really have no attention span, it's more like 30%.

We bring latency down by FULL SECONDS in time. And for agentic use cases, half a minute.

That is invaluable when you're trying to achieve growth.



The AI Triad



New SPEC-AI World Record – Image AI latency

WEKA = 10x better Latency per SPEC and other benchmarks

Business Metric	Requested Op Rate	Achieved Op Rate	Avg Lat (ms)	Total KBps	Read KBps	Write KBps	Run Sec	# Cl	Cl Proc	Avg File Size KB
500	217500.00	217528.609	0.395	48900397.671	48196471.598	703926.073	300	24	354	9952
1000	435000.00	435057.215	0.392	97807943.708	96402915.835	1405027.874	300	24	708	9952
1500	652500.00	652585.858	0.389	146700540.885	144590317.690	2110223.195	300	24	1062	9952
2000	870000.00	870114.585	0.471	195584089.353	192783905.372	2800183.981	300	24	1416	9952
2500	1087500.00	1087643.294	0.566	244499360.740	240993248.067	3506112.674	300	24	1770	9952
3000	1305000.00	1305171.746	0.648	293392665.839	289182750.616	4209915.223	300	24	2125	9952
3500	1522500.00	1522700.308	0.778	342271984.326	337368950.153	4903034.173	300	24	2479	9952
4000	1740000.00	1740228.830	1.015	391182097.847	385580483.639	5601614.208	300	24	2833	9952
4500	1957500.00	1957753.802	1.661	440080610.910	433773690.777	6306920.134	300	24	3187	9952
5000	2175000.00	2174983.735	6.803	488945723.200	481940565.813	7005157.387	300	24	3541	9952

Low Latency = More Cycles



Test-Time Compute (Reasoning)

VERY LATENCY-SENSITIVE & MEMORY-INTENSIVE



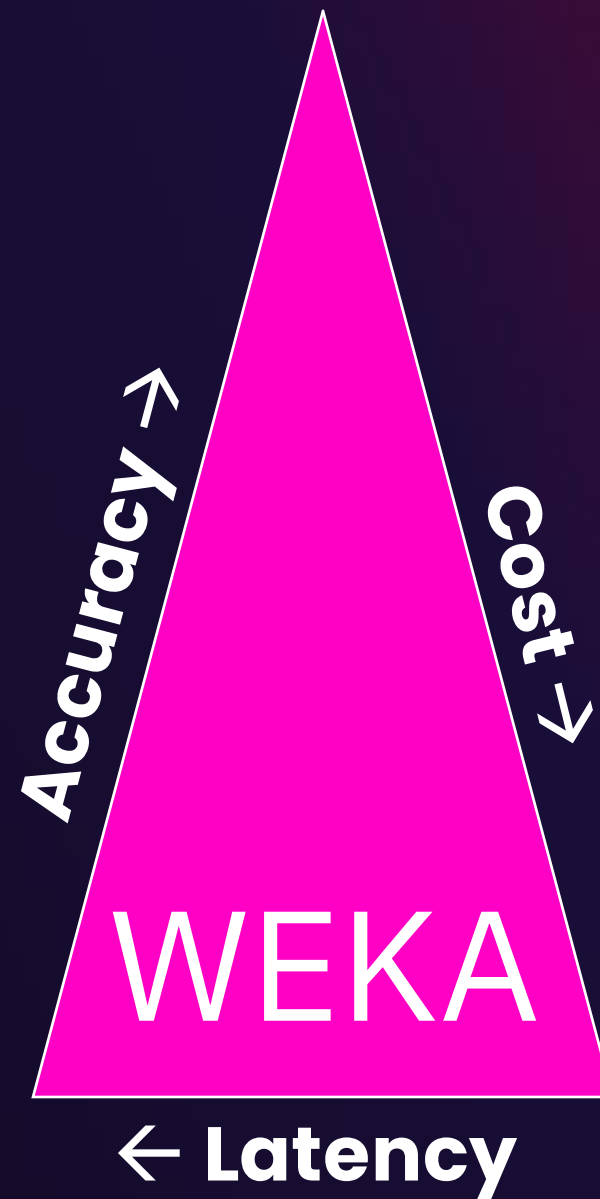
Multi-Pass Pre & Post-Training (RL)

More Checkpoints, Epochs, IOPS, Lower-Latency



Synthetic Data Processing

Generation, Processing, Deletion (Storage is not enough)



WARRP – WEKA AI Rag Reference Platform

What is WARRP?

Introduction

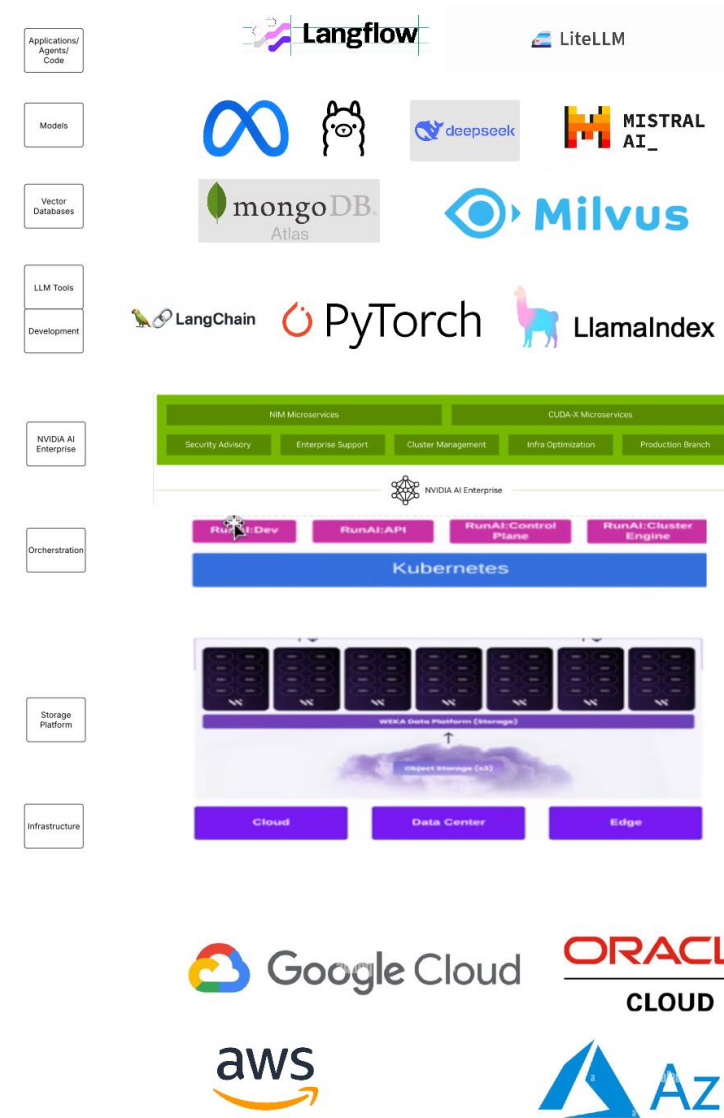
The WEKA AI RAG Reference Platform streamlines RAG pipelines for scalable operations.

Built on WEKA

Accelerates inferencing workloads and optimizes performance metrics.

Purpose

Designed for organizations navigating complex AI landscapes.



Key Benefits of WARRP

Optimized Performance



Accelerates model inferencing and reduces latency.

Scalable and Modular



Supports hybrid and multi-cloud environments with ease.

Cost Efficiency



Improves resource utilization and reduces operational expenses.

Technology Stack

**WEKA Data
Platform**



Delivers ultra-fast, low-latency data access and advanced features.

**NVIDIA AI
Enterprise**



Provides scalable AI applications and robust inferencing tools.

**Run:ai
Integration**



Maximizes GPU ROI with automated resource allocation.

Performance Metrics

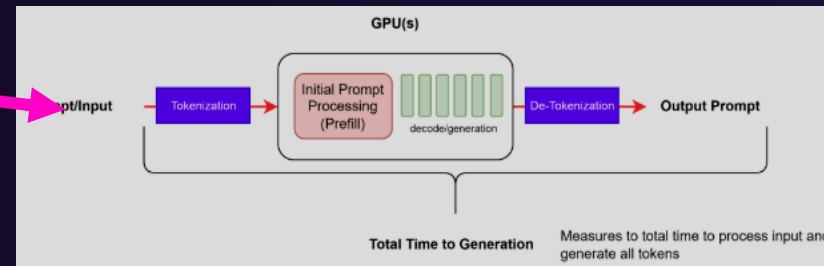
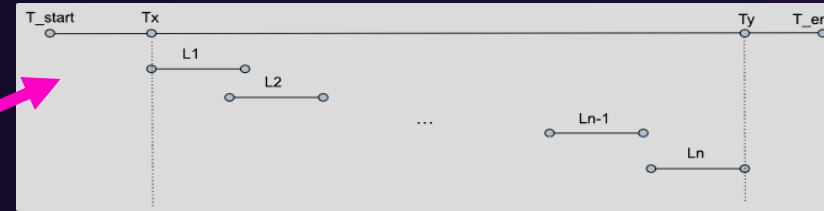
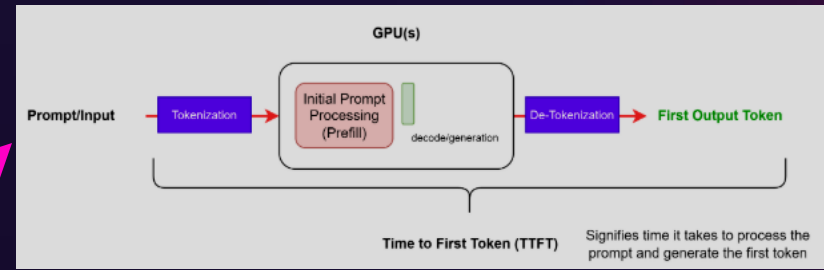
Optimizing AI Operations

What metrics define success in RAG pipelines?

Key metrics include:

- Time to First Token
- Token Throughput
- Token Output Latency

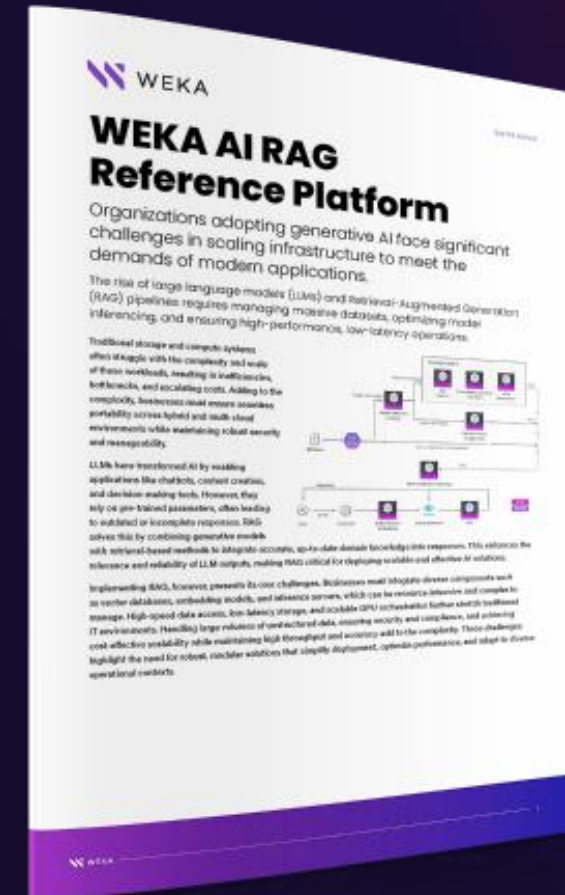
which ensure efficiency and scalability.



WARRP's Role in the AI Economy

Ensures that AI enterprises, data scientists, and IT teams can focus on innovation

- **Modular Design:** WARRP ensures adaptability and efficiency to meet AI business goals
- **Future-Ready:** Designed to address evolving technology challenges
- **Leading Performance:** WEKA delivers high-performance scale with sub-millisecond workload latency
- **Kubernetes-Native:** Seamless integration for modern AI workflows
- **Optimized for AI:** Provides a robust foundation for demanding AI workloads





WEKA Augmented Memory Grid

Distributed KV Cache Over Fabric

WEKA Enables the **NVIDIA AI Data Platform**

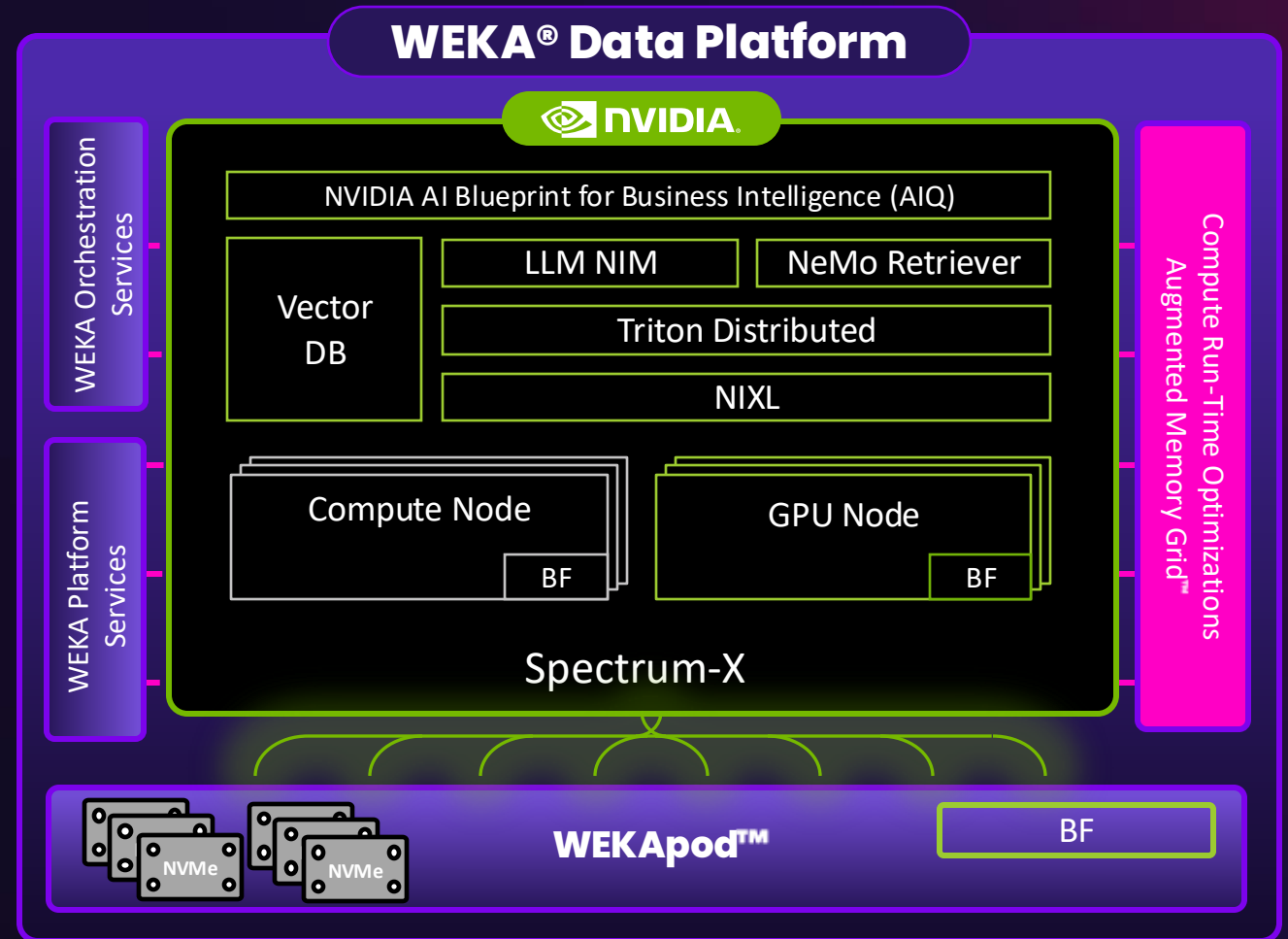
Scalable, efficient, high-performance AI infrastructure for agentic reasoning

**21x Faster Time to
First Token**

**Maximize AI
Infrastructure ROI**

Fastest model load

**Scalable RAG Inference
for Multi-Agent AI**



What is KV-Cache?

Refresher on the LLM KV Cache

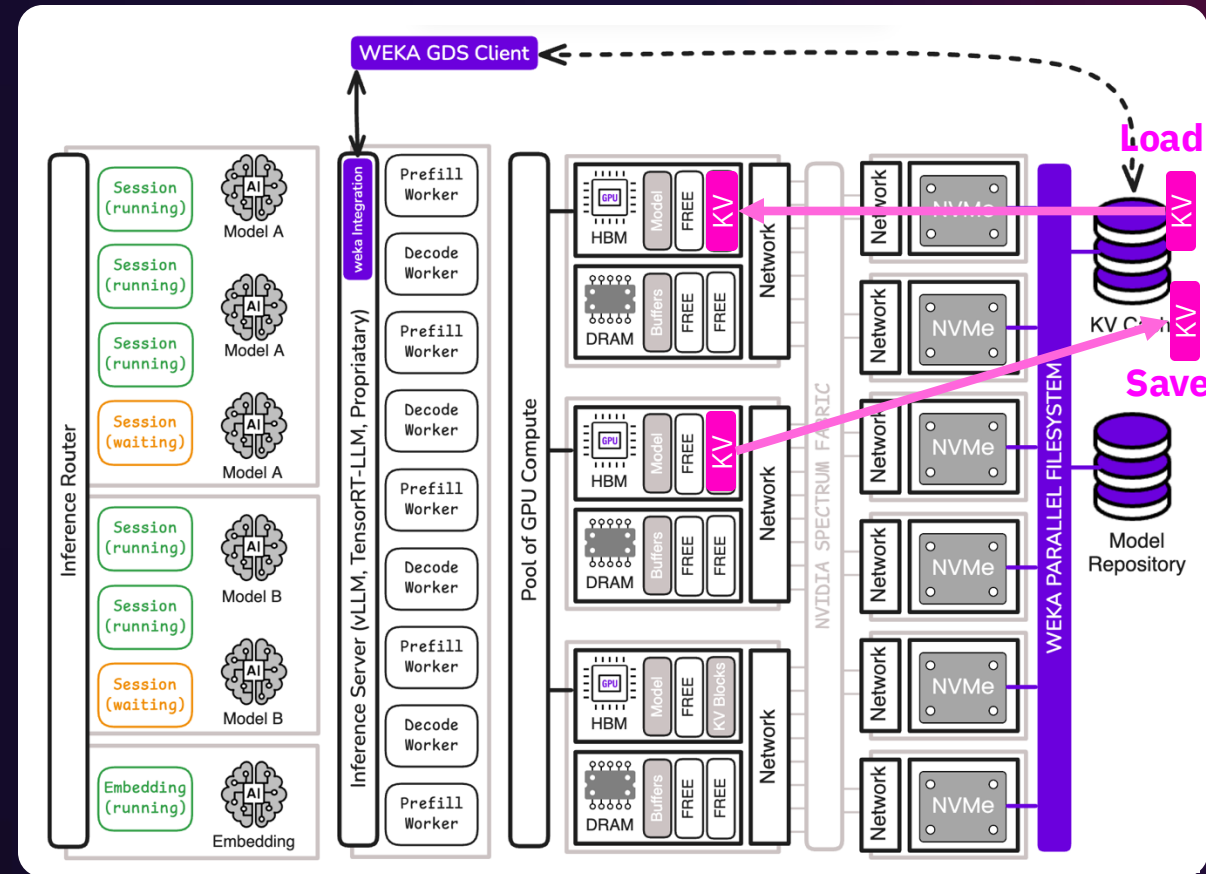
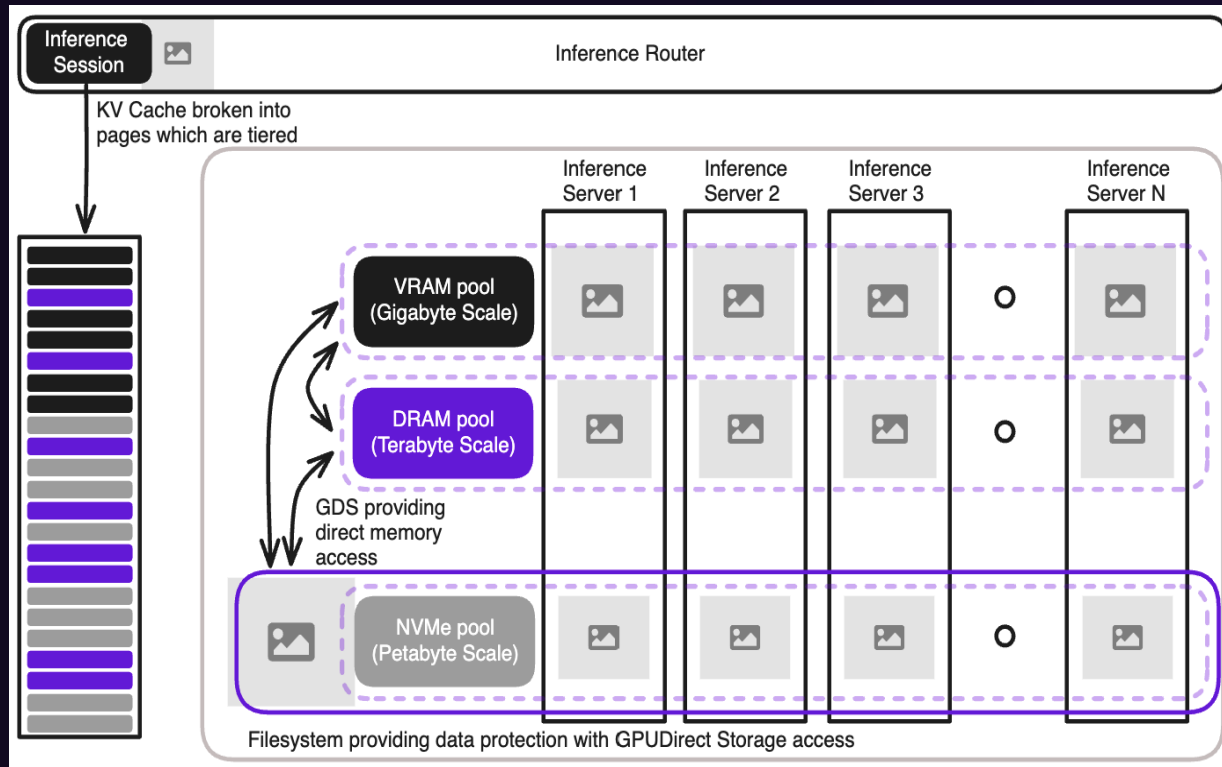
- LLM Key-Value (KV) cache is critical optimization for reducing impact of token lookups - two phases of inference:
 - **Prefill**, where the KV cache is populated for input tokens
 - **Decode**, where output tokens are generated

What's the problem?

- Generating KV Cache is a GPU intensive operation and takes GPU memory capacity
- As context length grows the KV cache size increases linearly
- HBM and even DRAM resources are under constant pressure and need to constantly evict cached data quickly

Distributed KV-Cache over Fabric

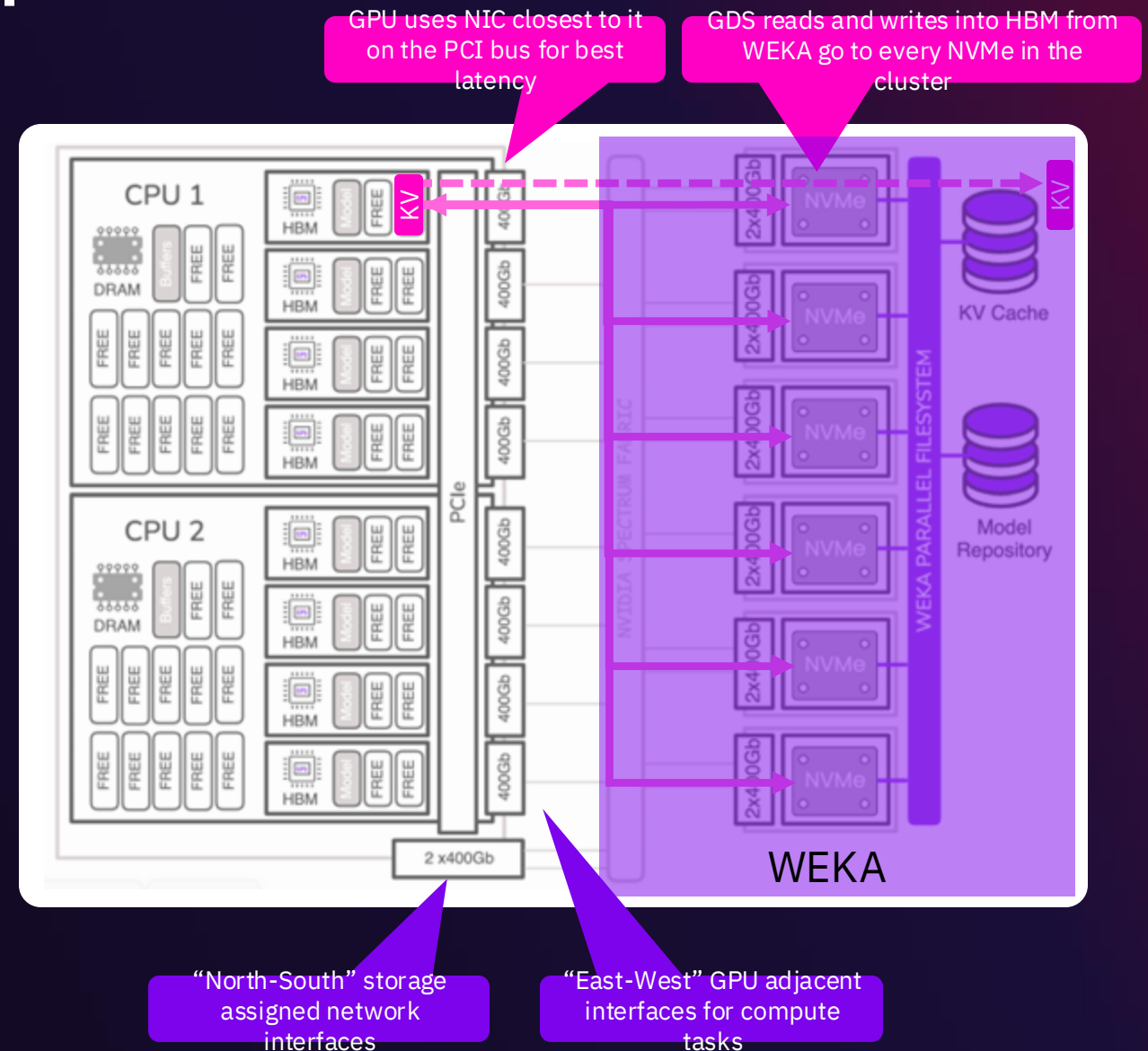
Use WEKA microsecond latency as a shared GPU memory layer



Lab Setup – GPU host network setup

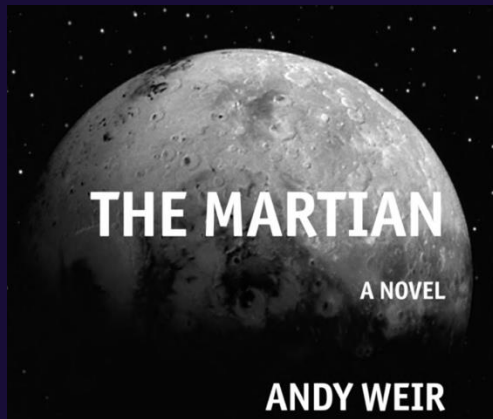
WEKA integration for minimum latency

- DGX H100 has all network ports cabled at full speed (non-blocking with no oversubscribe)
- Our switching in the WEKA lab is Quantum-2 NDR 400Gb switching (host ports in IB mode)
- While “East-West” ports are utilized during training, they are very under-utilized during inference
- WEKA uses “East-West” ports to provide compute time optimizations (effective memory extension)

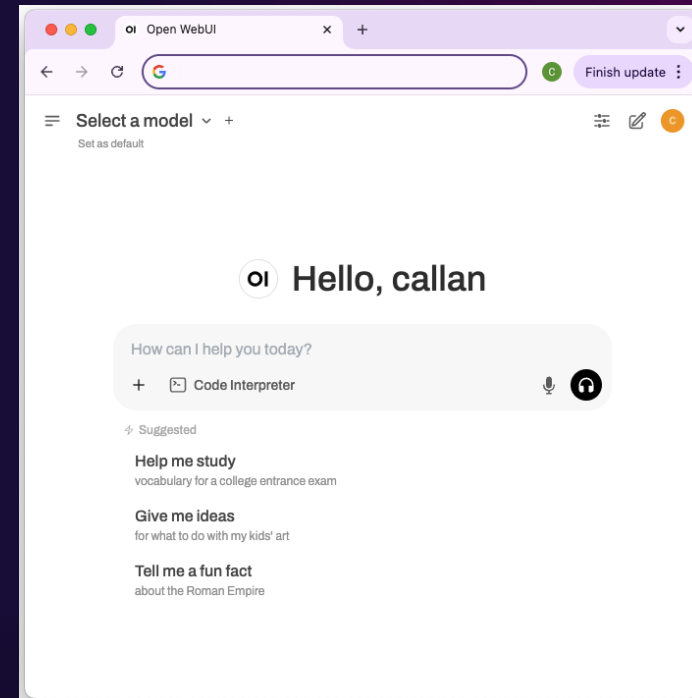


**Set the scene
for demo**





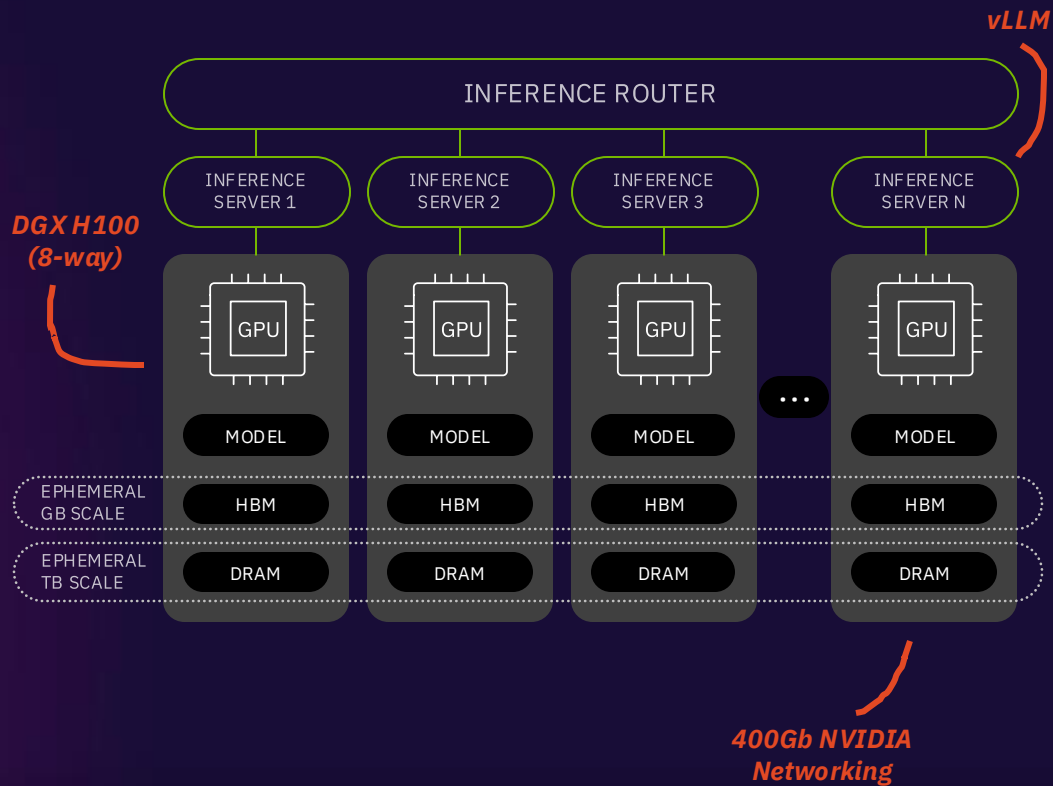
Downloadable book



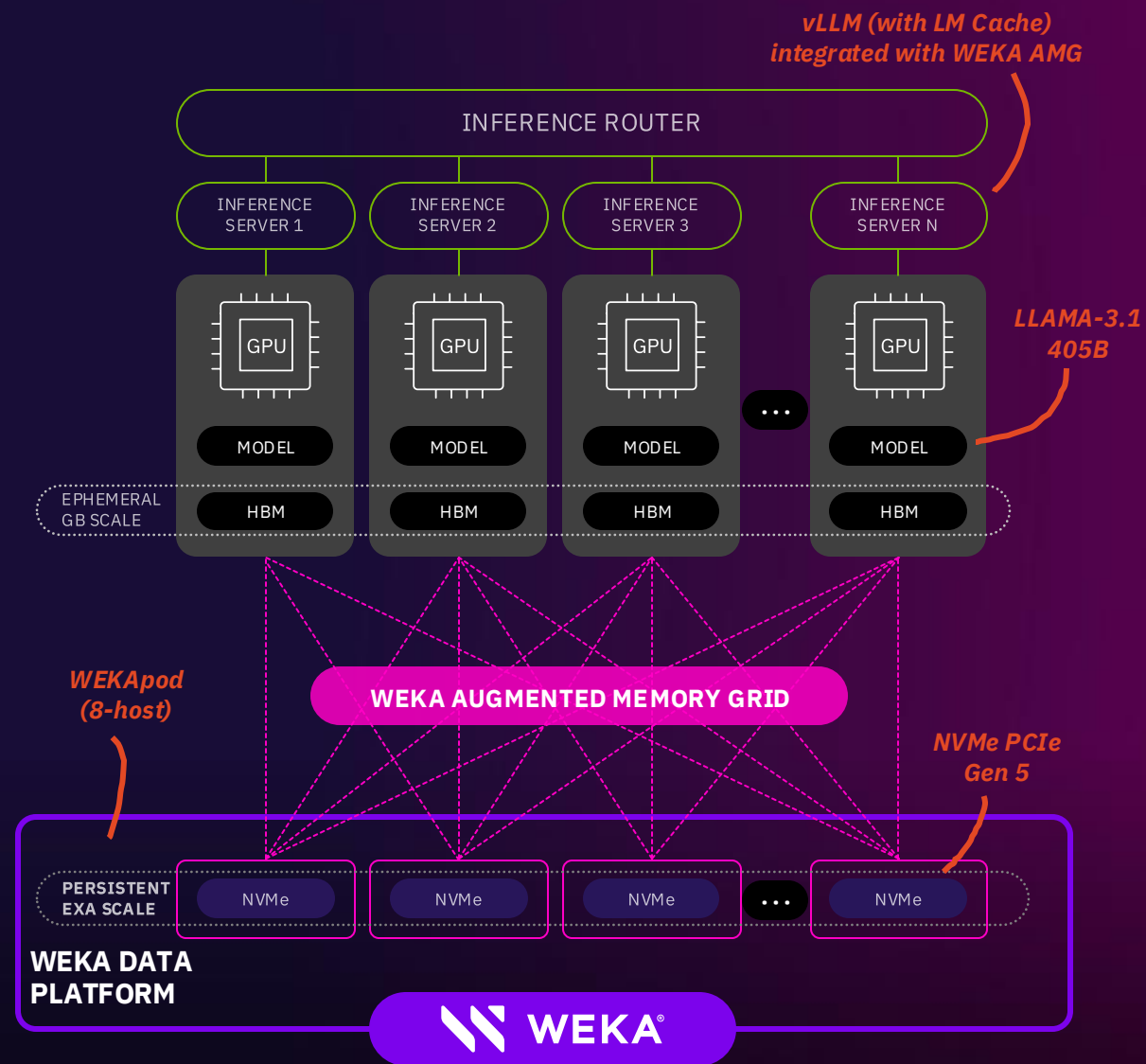
Introducing **WEKA Augmented Memory Grid**

Distributed KV Cache Over Fabric

Without WEKA AMG



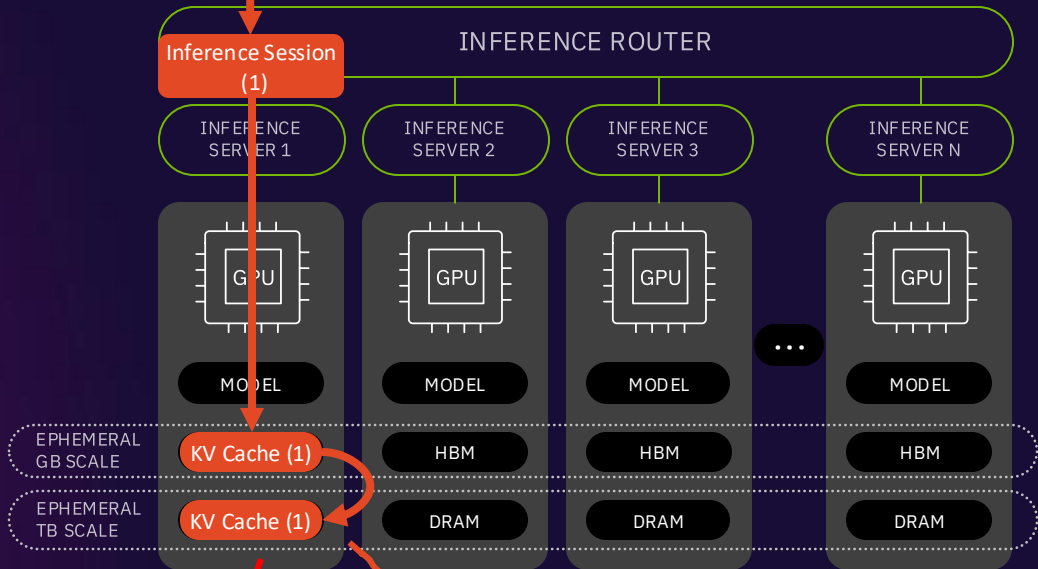
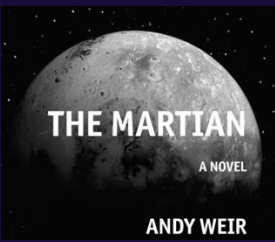
With WEKA AMG



User 1

Without WEKA AMG

Initial prefill of "The Martian" (37.3s)



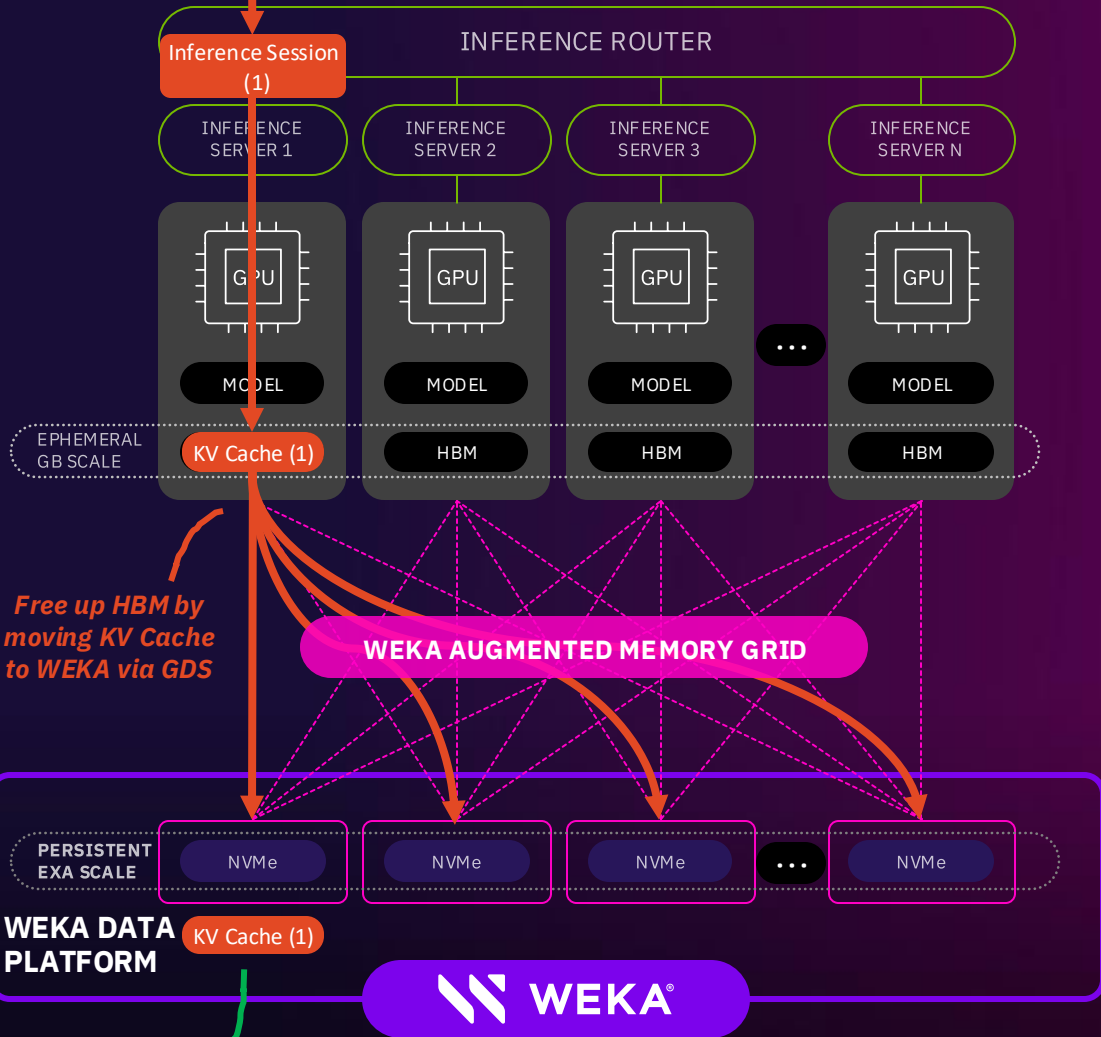
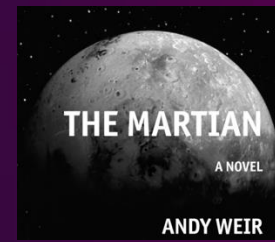
X KV Cache available to local host only

Free up HBM by moving KV Cache to DRAM

User 1

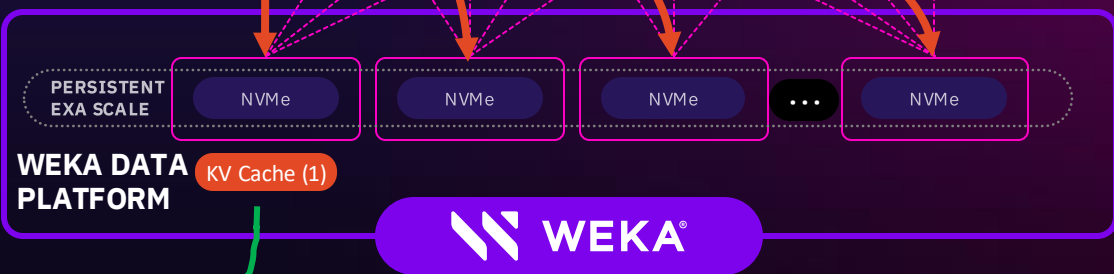
With WEKA AMG

Initial prefill of "The Martian" (37.3s)

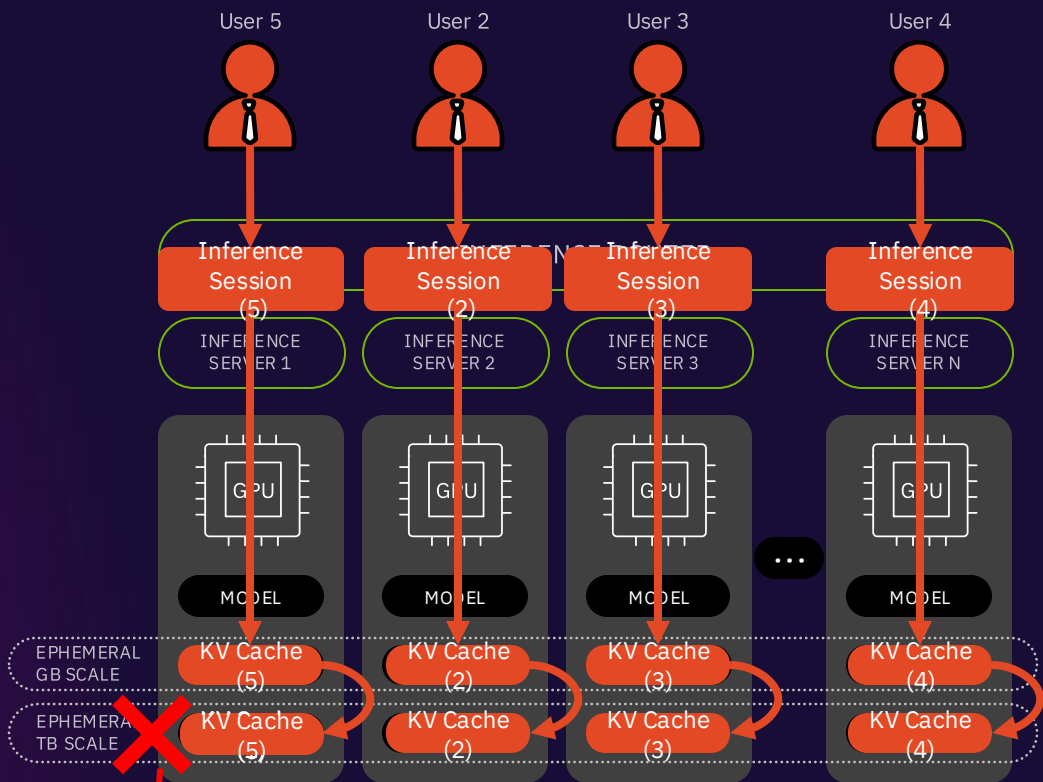


Free up HBM by moving KV Cache to WEKA via GDS

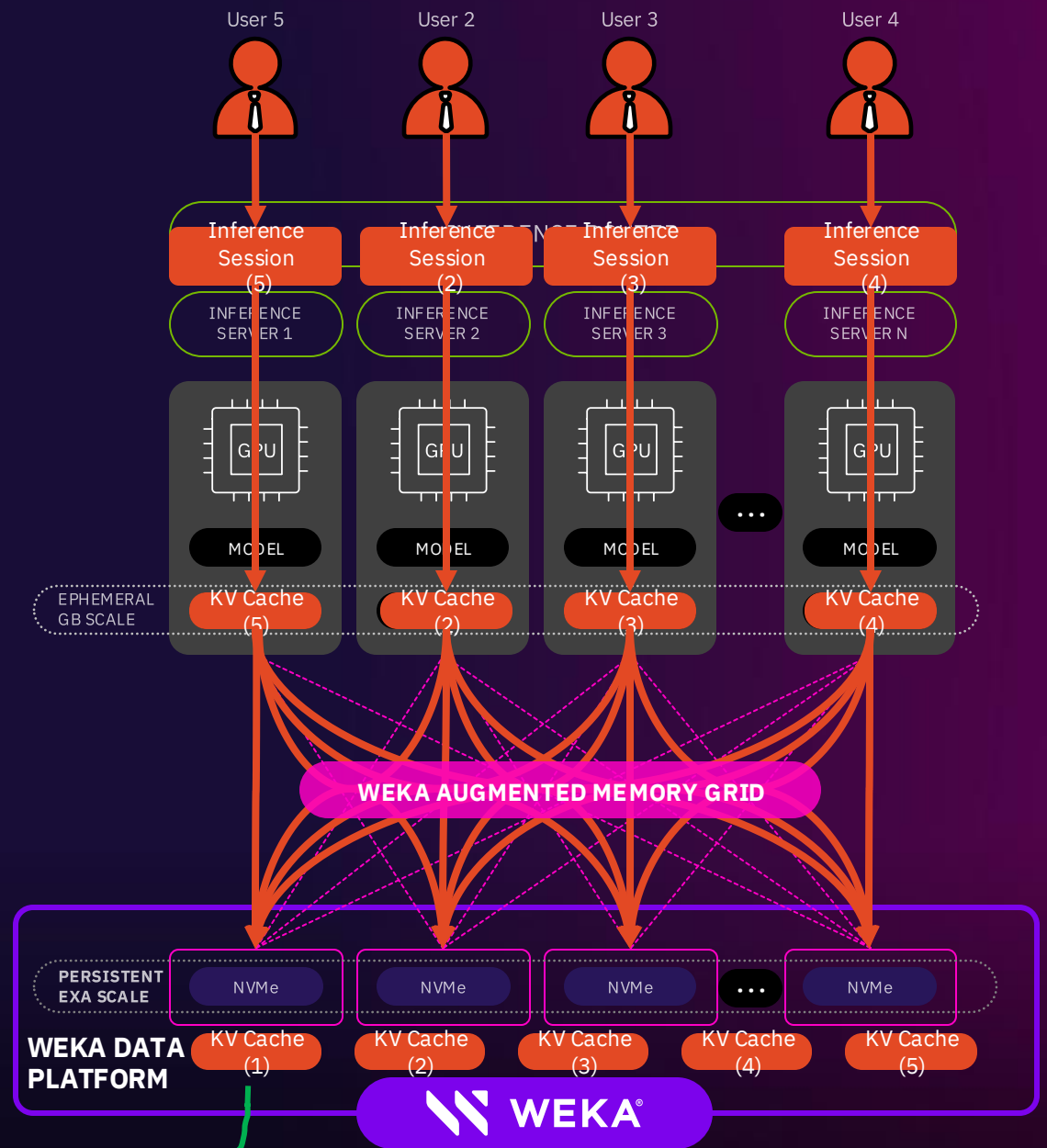
WEKA AUGMENTED MEMORY GRID



✓ KV Cache is available to all hosts



Lack of sizable centralized cache causes KV Cache to be purged for user 1

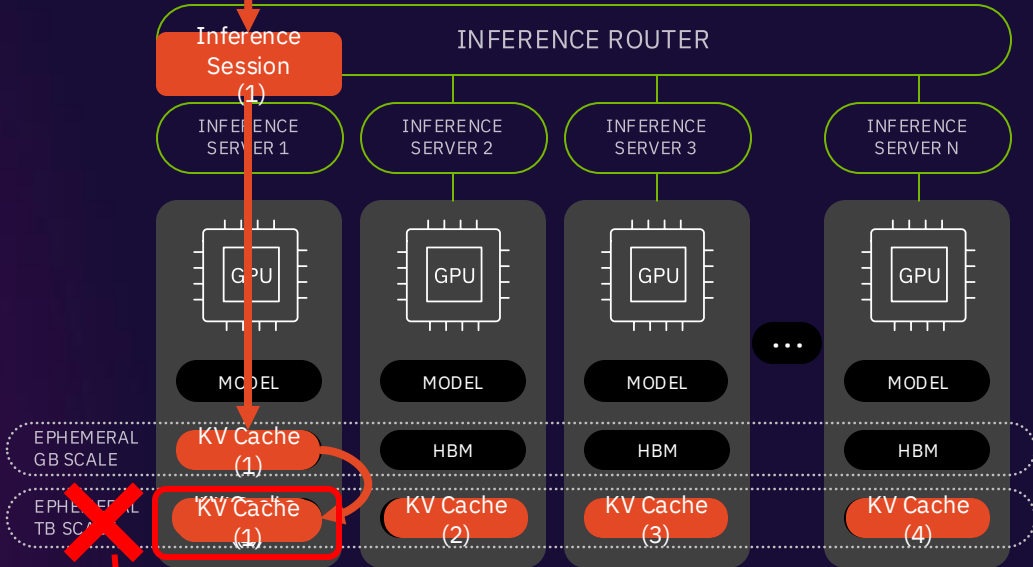


Significantly larger cache allows for significantly higher cache hit rates

Without WEKA AMG



TTFT 37.3s



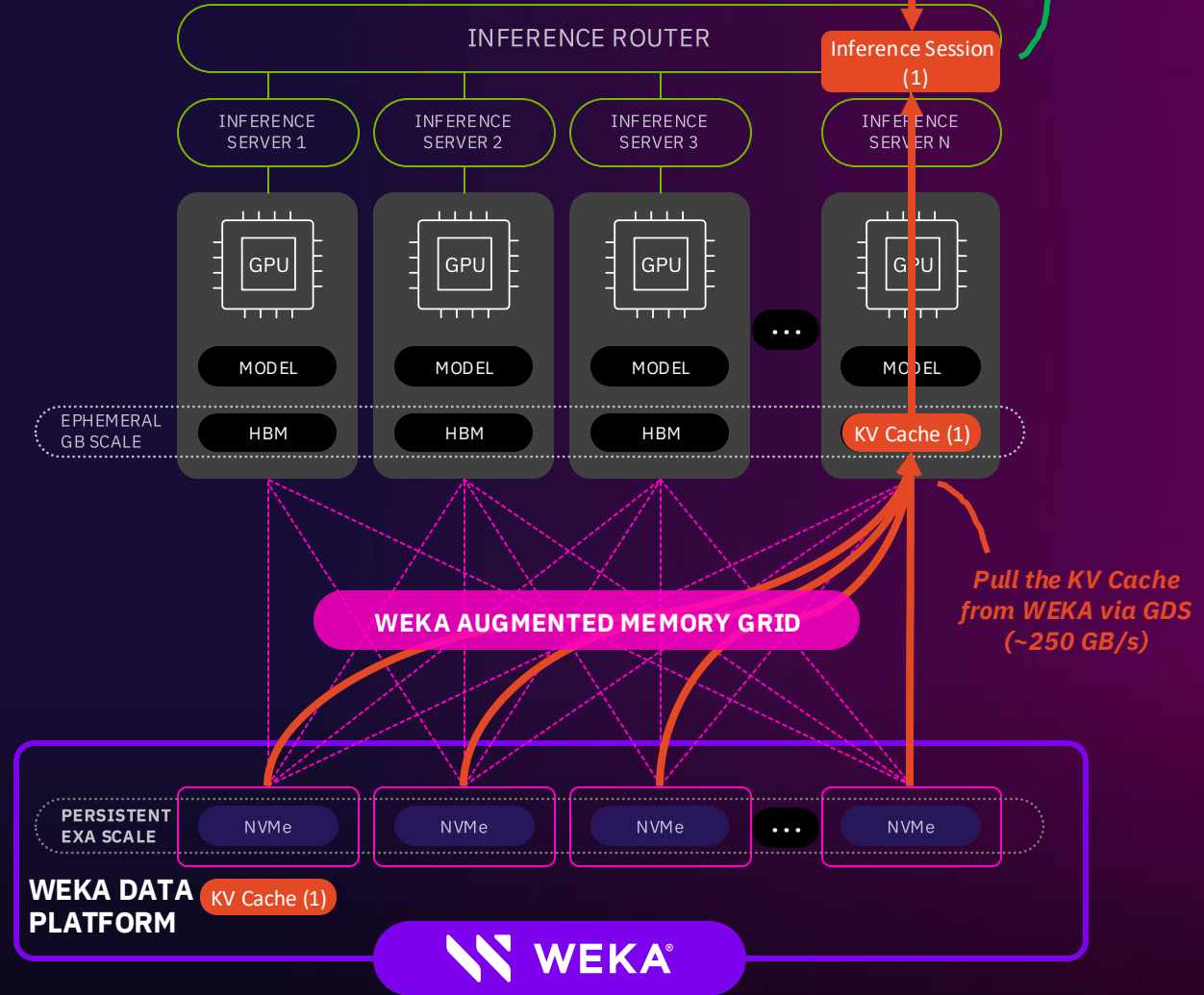
User 1 has no KV Cache for its session, needs to prefill again and purge cache for session 5 (cache thrashing)

With WEKA AMG

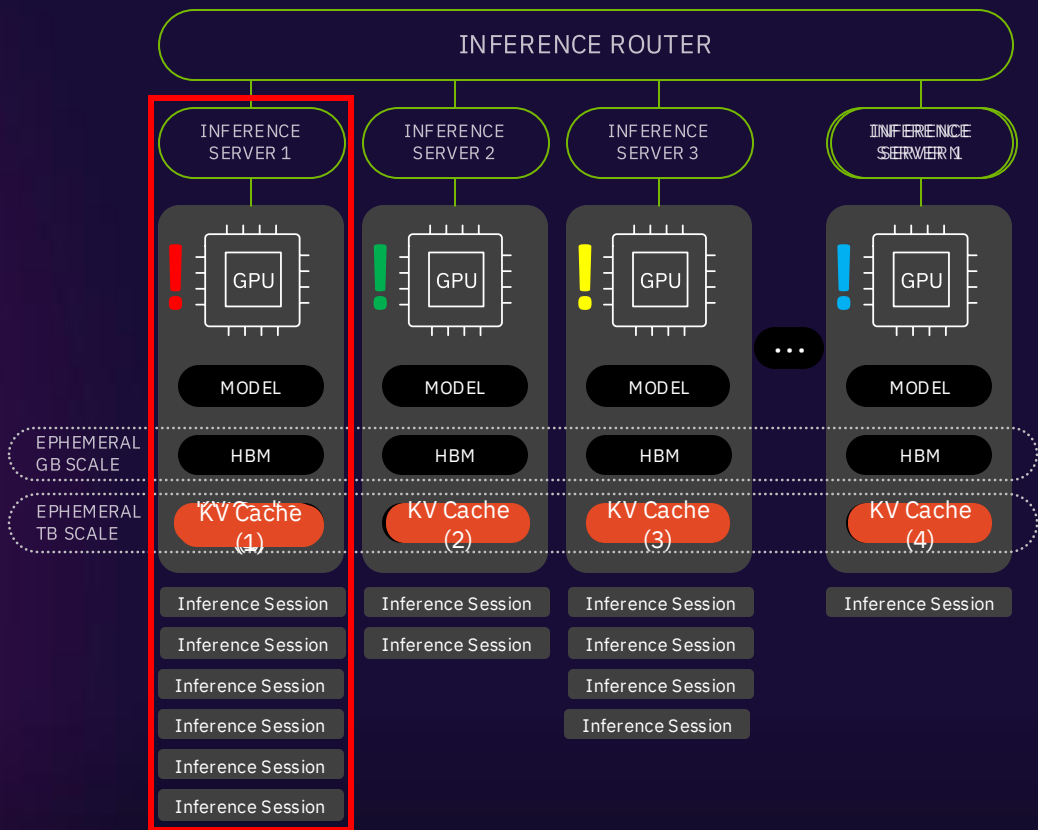


TTFT ~1.5s

Session can run against any host without penalty

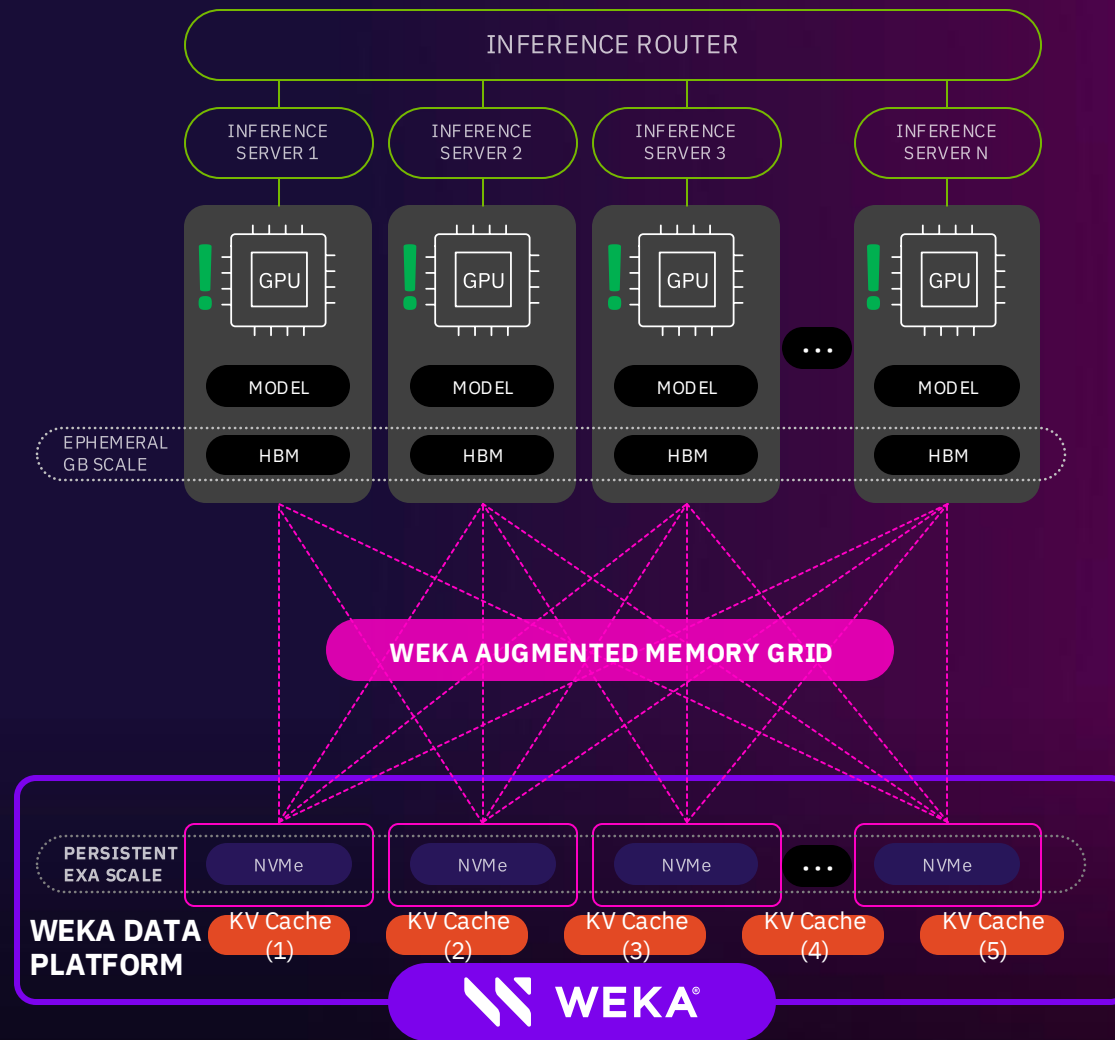


Without WEKA AMG



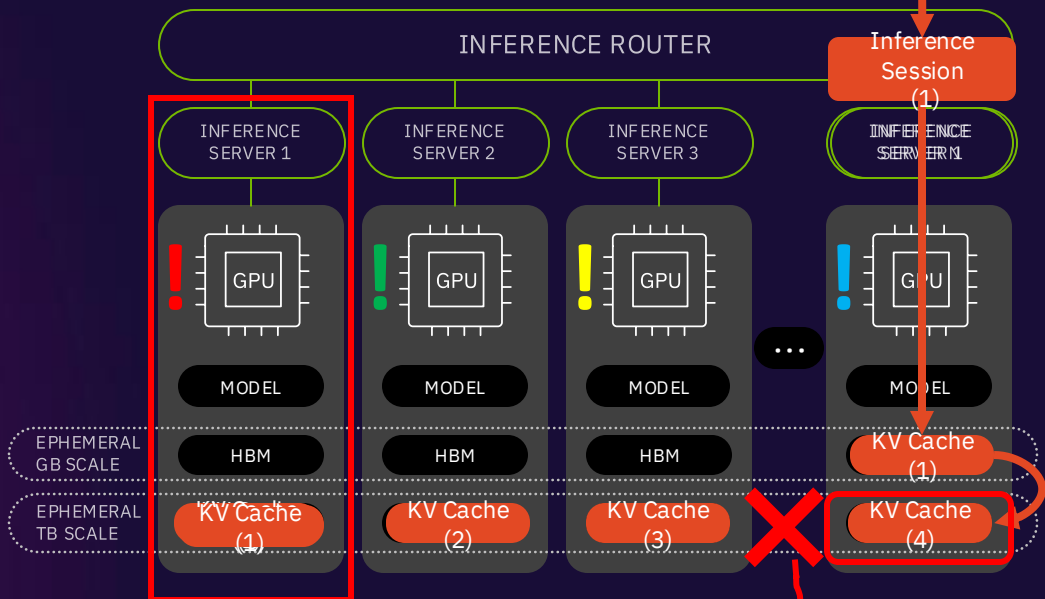
X Creates hotspots due to host to cache affinity

With WEKA AMG



Without WEKA AMG

TTFT 37.3s

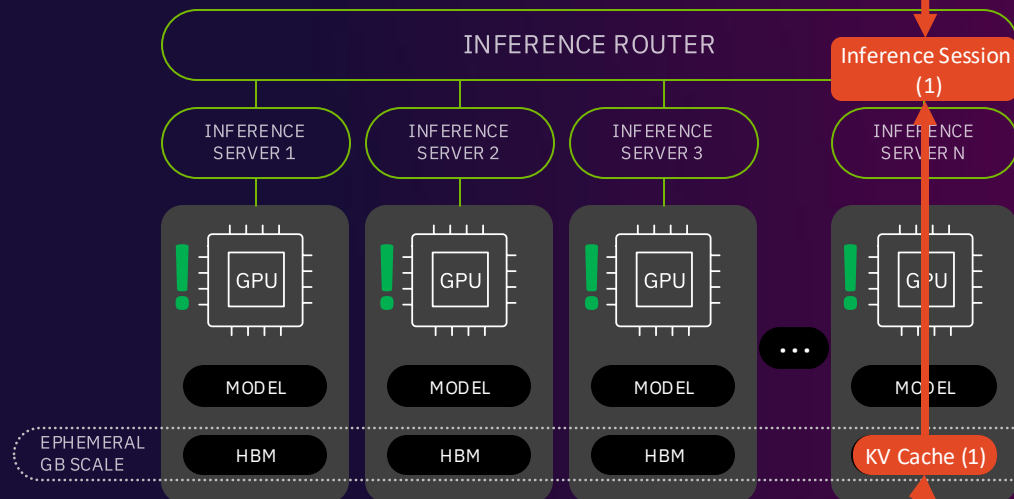


X Host is too busy for user 1, router decides to prefill

User 1 has no KV Cache for its session, needs to prefill again and purge cache for session 4 (cache thrashing)

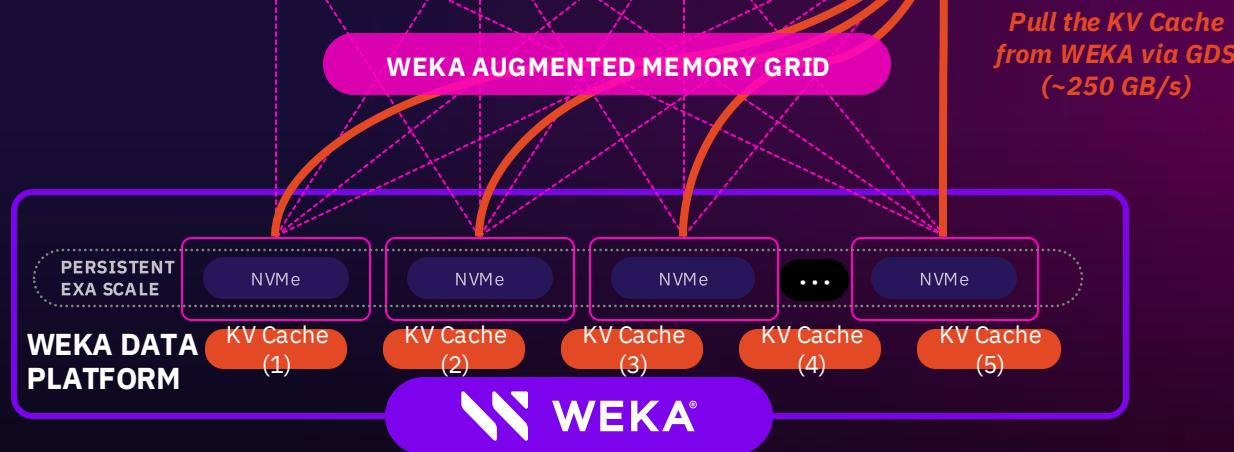
With WEKA AMG

TTFT ~1.5s



Session can run against any host without penalty

WEKA AUGMENTED MEMORY GRID



Pull the KV Cache from WEKA via GDS (~250 GB/s)

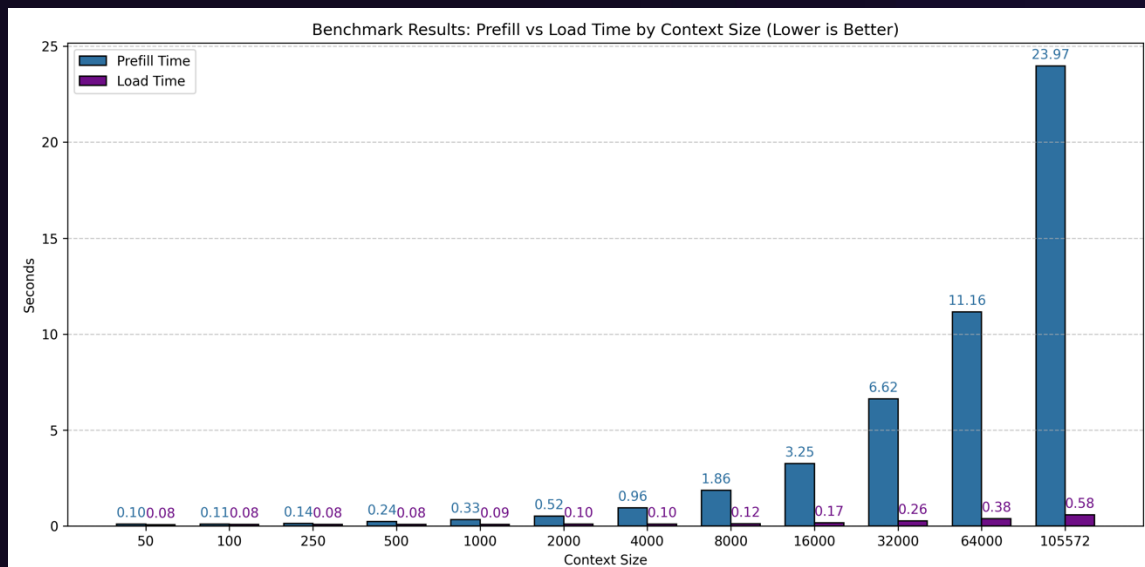


Results so far

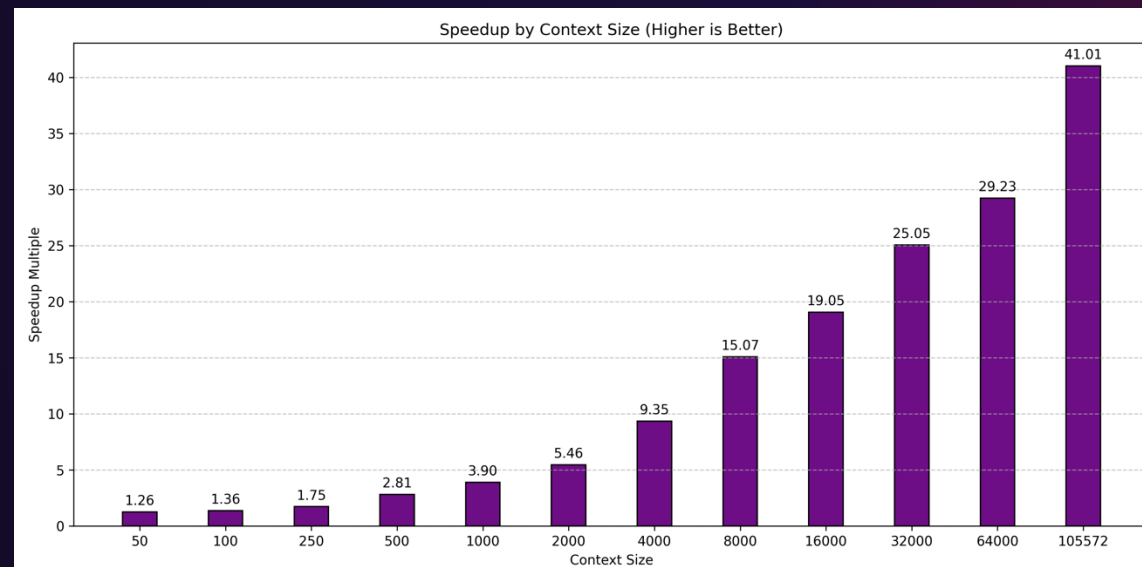
WEKA LLM KV Cache optimization massively increases **GPU effective utilization**

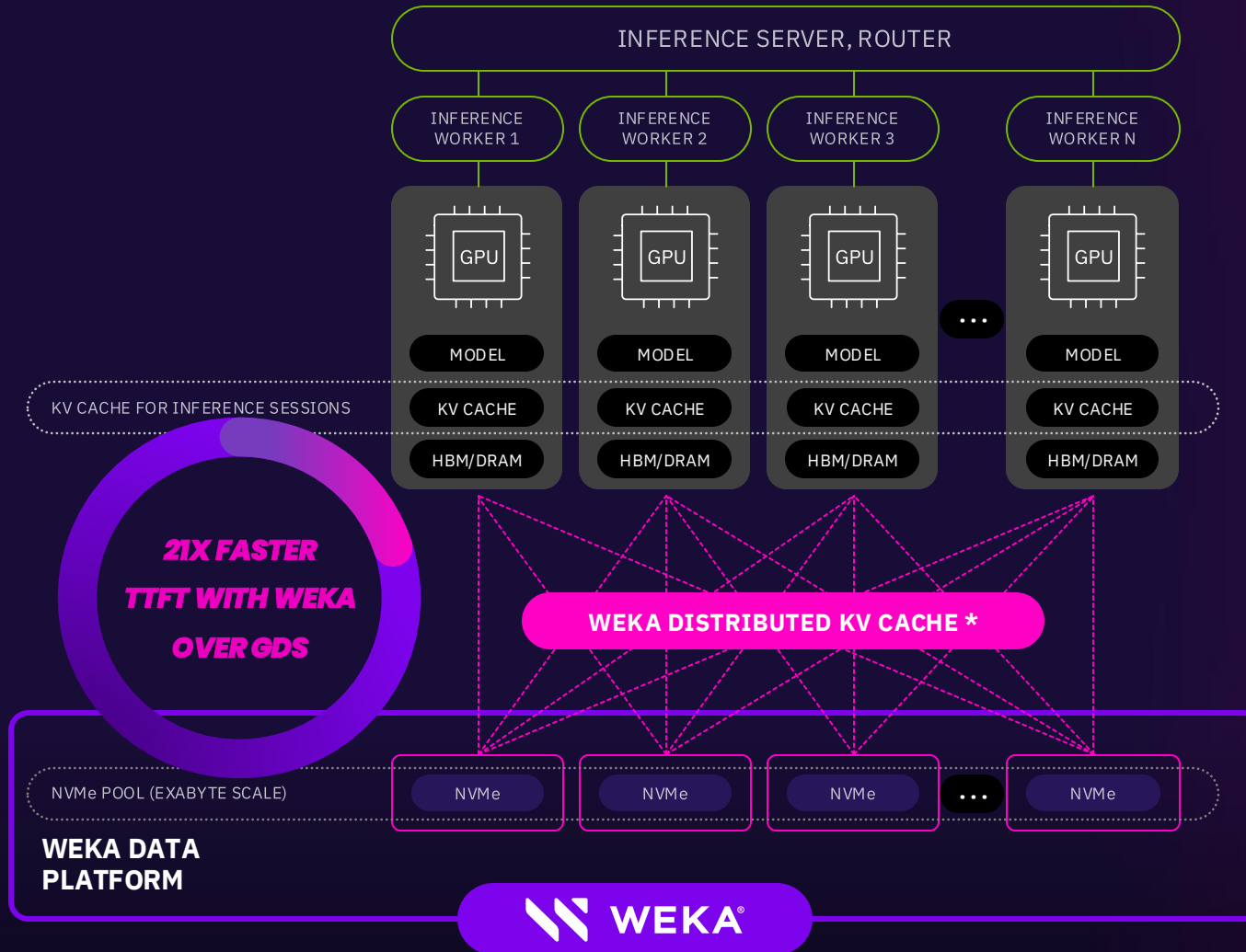
New vLLM version improves TTFT and reduce this by 50% (which is still 20x improvement over prefil)

Prefill versus Load time (seconds)



Speedup by context size (multiples)





**21X FASTER
TTFT WITH WEKA
OVER GDS**

Simplified integration

Slashes Average TTFT

Free-up GPU Cycles

WEKA AMG enables **Elastic Training & Inference**

Utilize Your Most Expensive Assets in a Dynamic Fashion

