Session Abstract

Scaling AI: Leverage Caching Algorithms to Maximize Agentic ROI

Agentic AI innovation is creating cost pressures that are crushing profit margins across industries. Today's AI architects face a critical challenge: balancing cost efficiency with rapid development cycles and competitive time-to-market advantages.

Key-value caching (KV Cache) offers a powerful solution. When implemented effectively, KV Cache can transform your AI economics by dramatically reducing time to first token and slashing cost per token—all without sacrificing performance or time to market.

WEKA's Val Bercovici and Betsy Chernoff will demonstrate how to architect and deploy caching algorithms that optimize your token economics, sharing practical strategies to achieve cost-effective agentic AI innovation at scale.

WEKA®

Scaling AI: Leveraging Caching Algorithms to Maximize Agentic ROI

Val Bercovici | Chief AI Officer

Betsy Chernoff | AI Product Marketing Lead

Agenda

Workshop Goal

Leave with concrete understanding of design patterns around KV cache optimizations.

1 Challenges

2 Key Metrics

3 Impact of Caches in LLMs

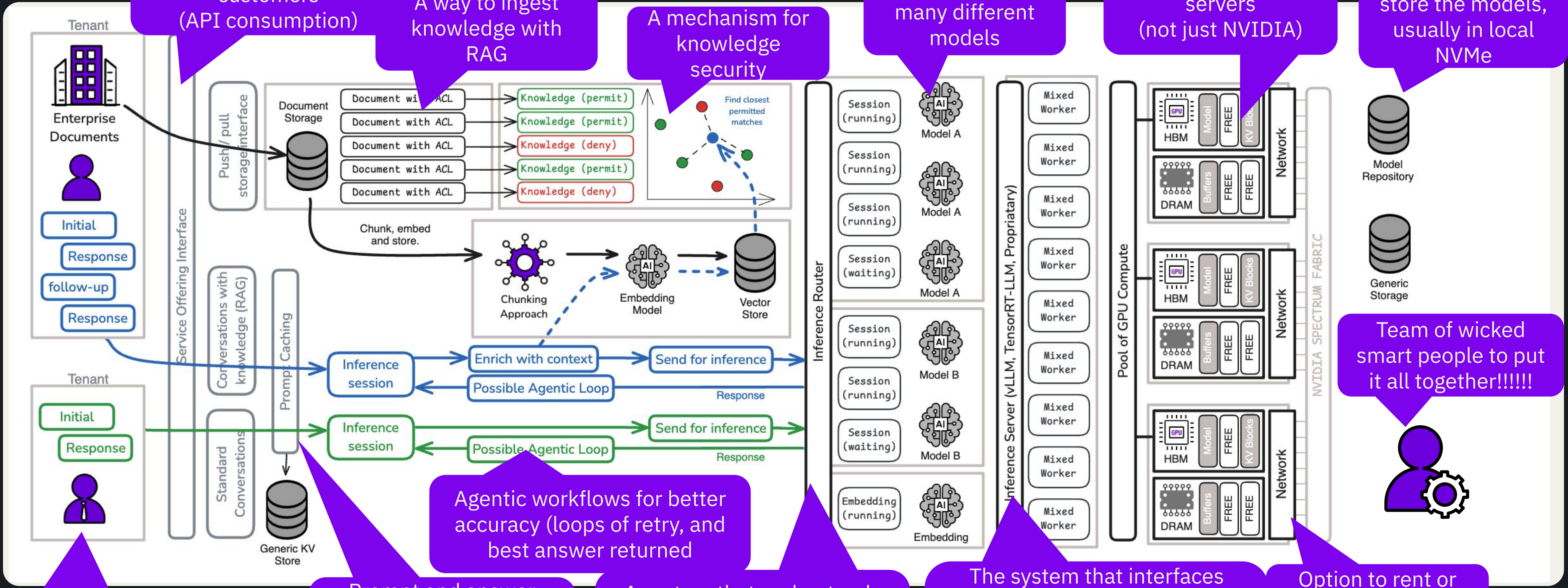
4 Implementation

5 Insights from Our Labs

6 Best Practices

Achieving Profitable Inference is Extremely Difficult

Common pattern of a modern inference system



List of services the system offers to customers (API consumption)

A way to ingest knowledge with RAG

A mechanism for knowledge security

Ability to run many different models

Choice of GPU servers (not just NVIDIA)

Somewhere to store the models, usually in local NVMe

Team of wicked smart people to put it all together!!!!!!

Multiple tenants with the huge numbers of users

Prompt and answer caching to prevent the need to do inference at all

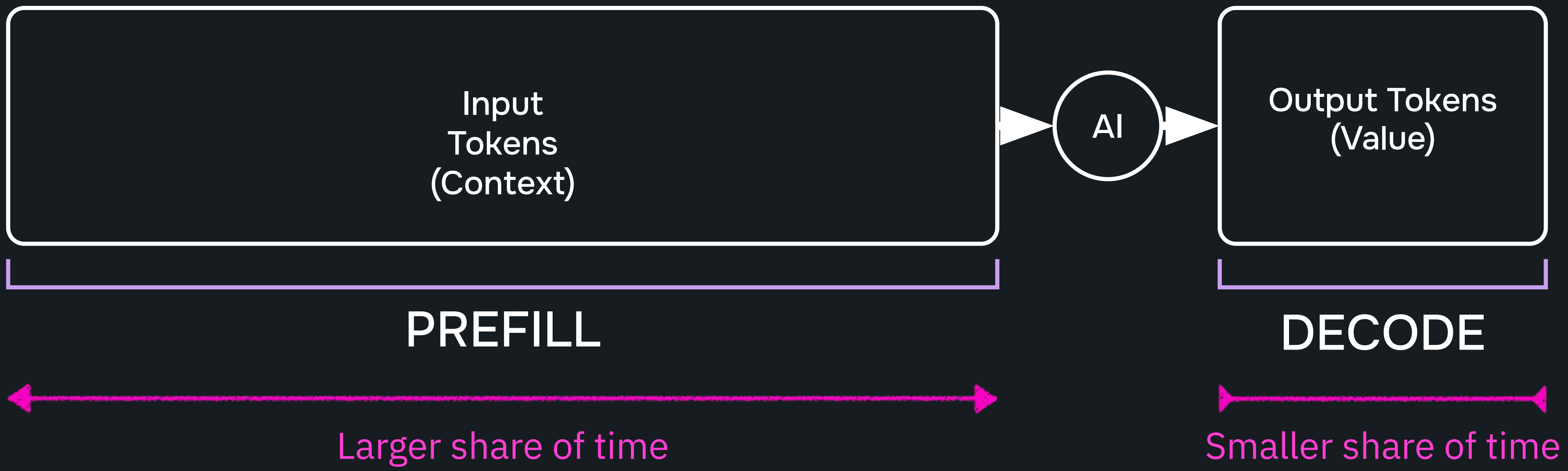
Agentic workflows for better accuracy (loops of retry, and best answer returned

A system that understands inference sessions and can route them to pools of resources

The system that interfaces directly with the GPU servers and schedules the work – DeepSeek showed that working at this level enables efficiencies

Option to rent or buy the equipment in COLO or Cloud

WEKA®



Input Tokens (Context)

AI

Output Tokens (Value)

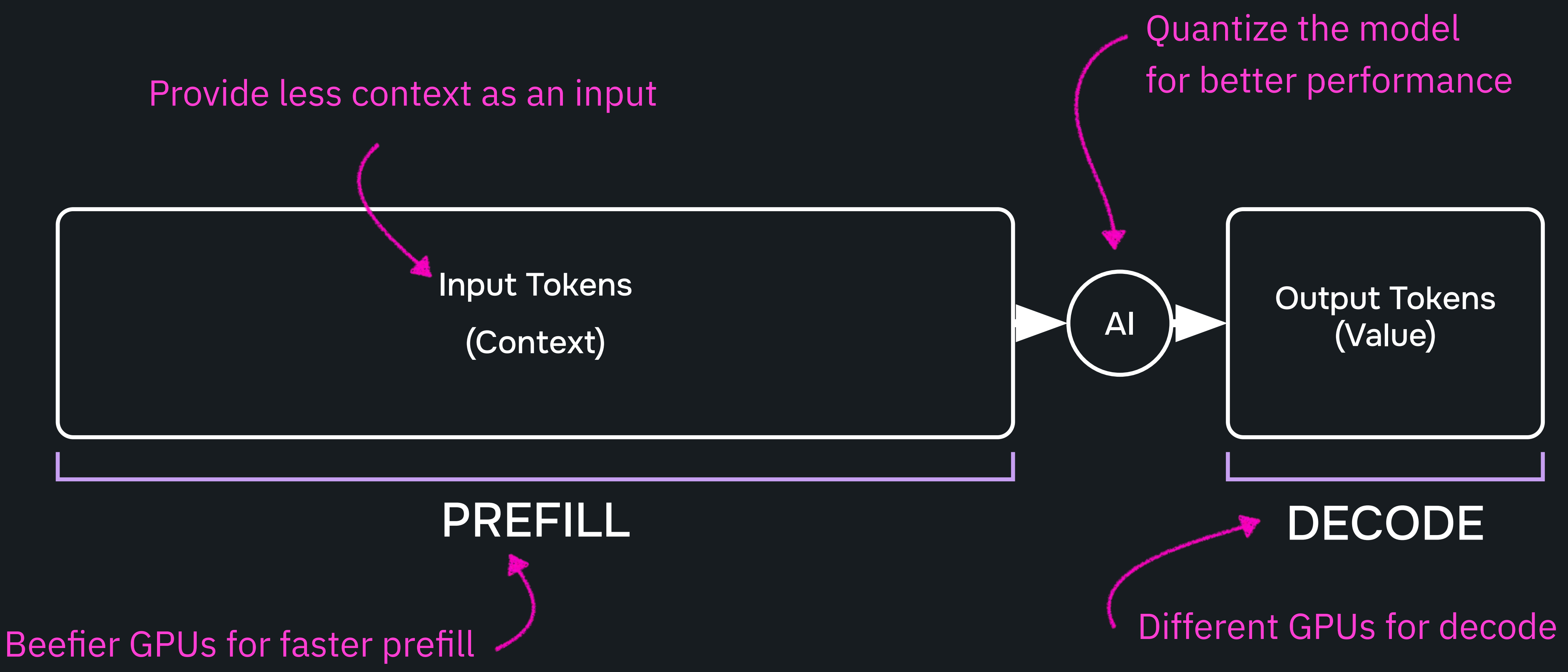
PREFILL

DECODE

Larger share of time

Smaller share of time

WEKA®



Provide less context as an input

Quantize the model for better performance

Input Tokens (Context)

AI

Output Tokens (Value)

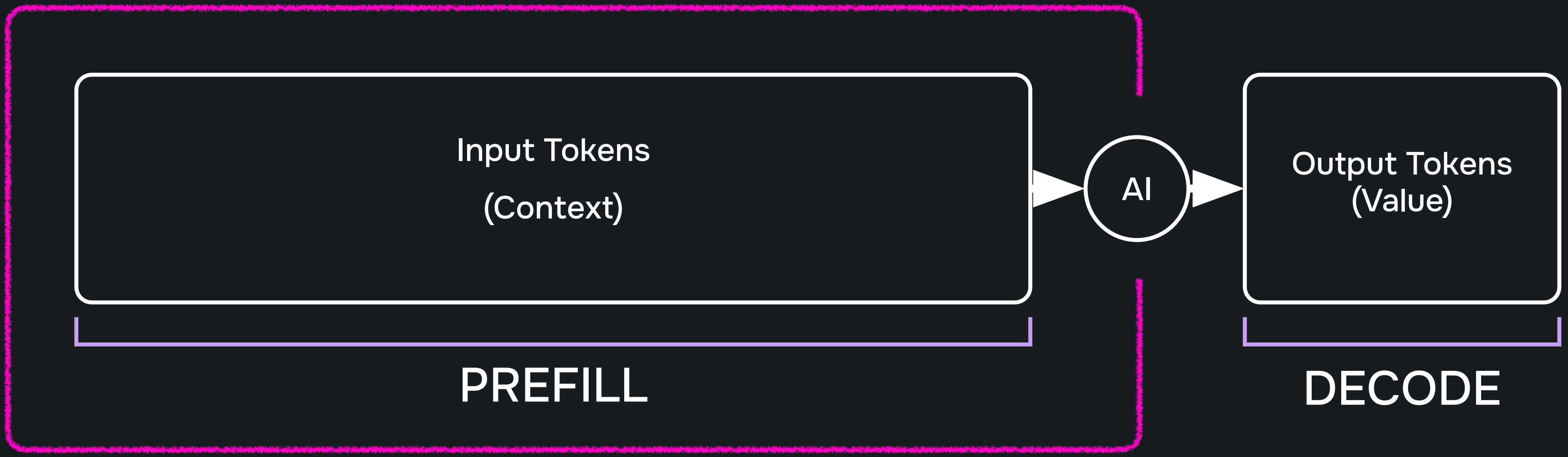
PREFILL

DECODE

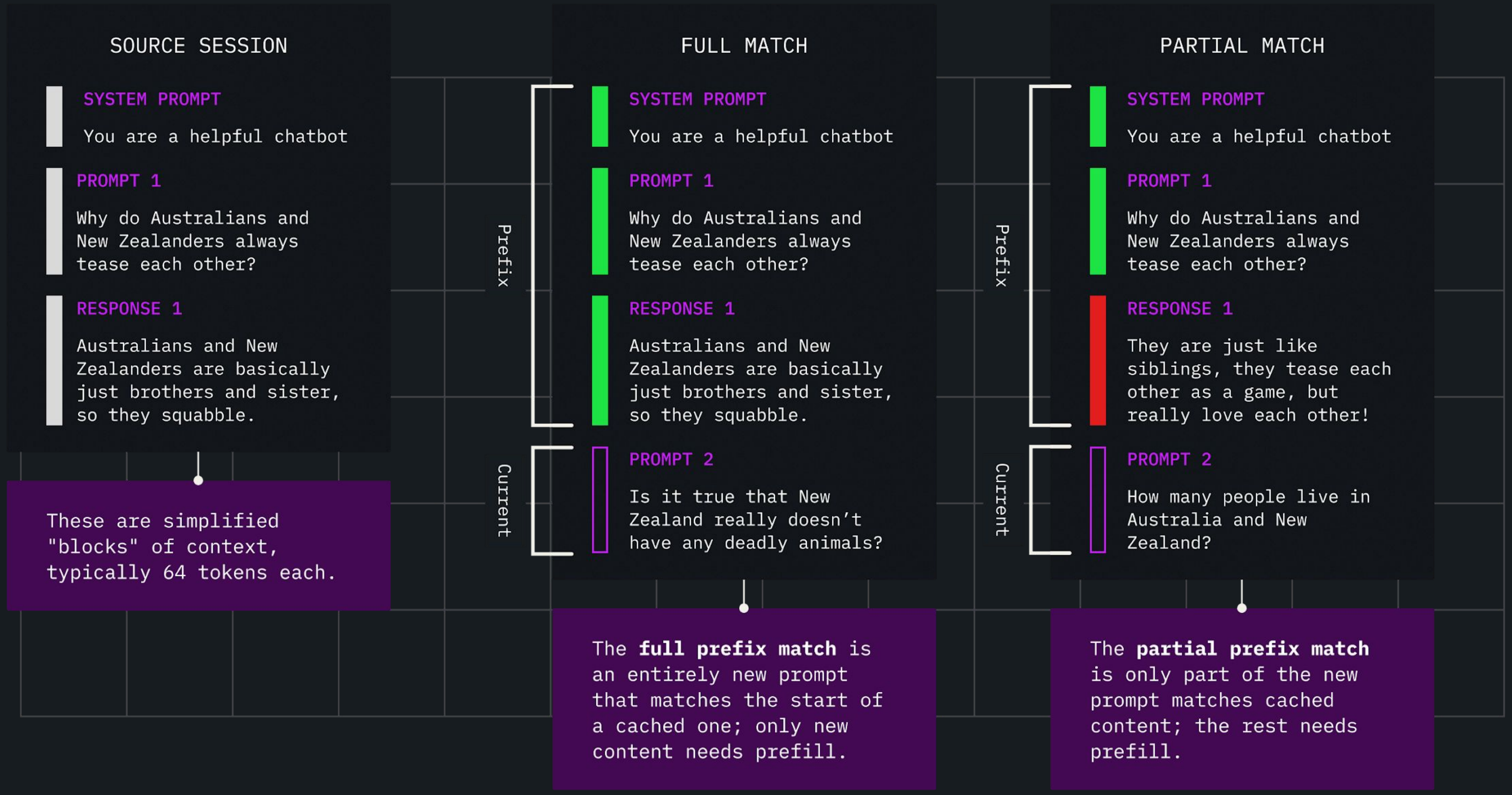
Beefier GPUs for faster prefill

Different GPUs for decode

DeepSeek showed 56% Cache Hit



Cache Hit Rate & Prefix Matching



SOURCE SESSION

⬜ SYSTEM PROMPT

You are a helpful chatbot

⬜ PROMPT 1

Why do Australians and New Zealanders always tease each other?

⬜ RESPONSE 1

Australians and New Zealanders are basically just brothers and sister, so they squabble.

These are simplified "blocks" of context, typically 64 tokens each.

FULL MATCH

Prefix

🟩 SYSTEM PROMPT

You are a helpful chatbot

🟩 PROMPT 1

Why do Australians and New Zealanders always tease each other?

🟩 RESPONSE 1

Australians and New Zealanders are basically just brothers and sister, so they squabble.

Current

▯ PROMPT 2

Is it true that New Zealand really doesn’t have any deadly animals?

The **full prefix match** is an entirely new prompt that matches the start of a cached one; only new content needs prefill.

PARTIAL MATCH

Prefix

🟩 SYSTEM PROMPT

You are a helpful chatbot

🟩 PROMPT 1

Why do Australians and New Zealanders always tease each other?

🟥 RESPONSE 1

They are just like siblings, they tease each other as a game, but really love each other!

Current

▯ PROMPT 2

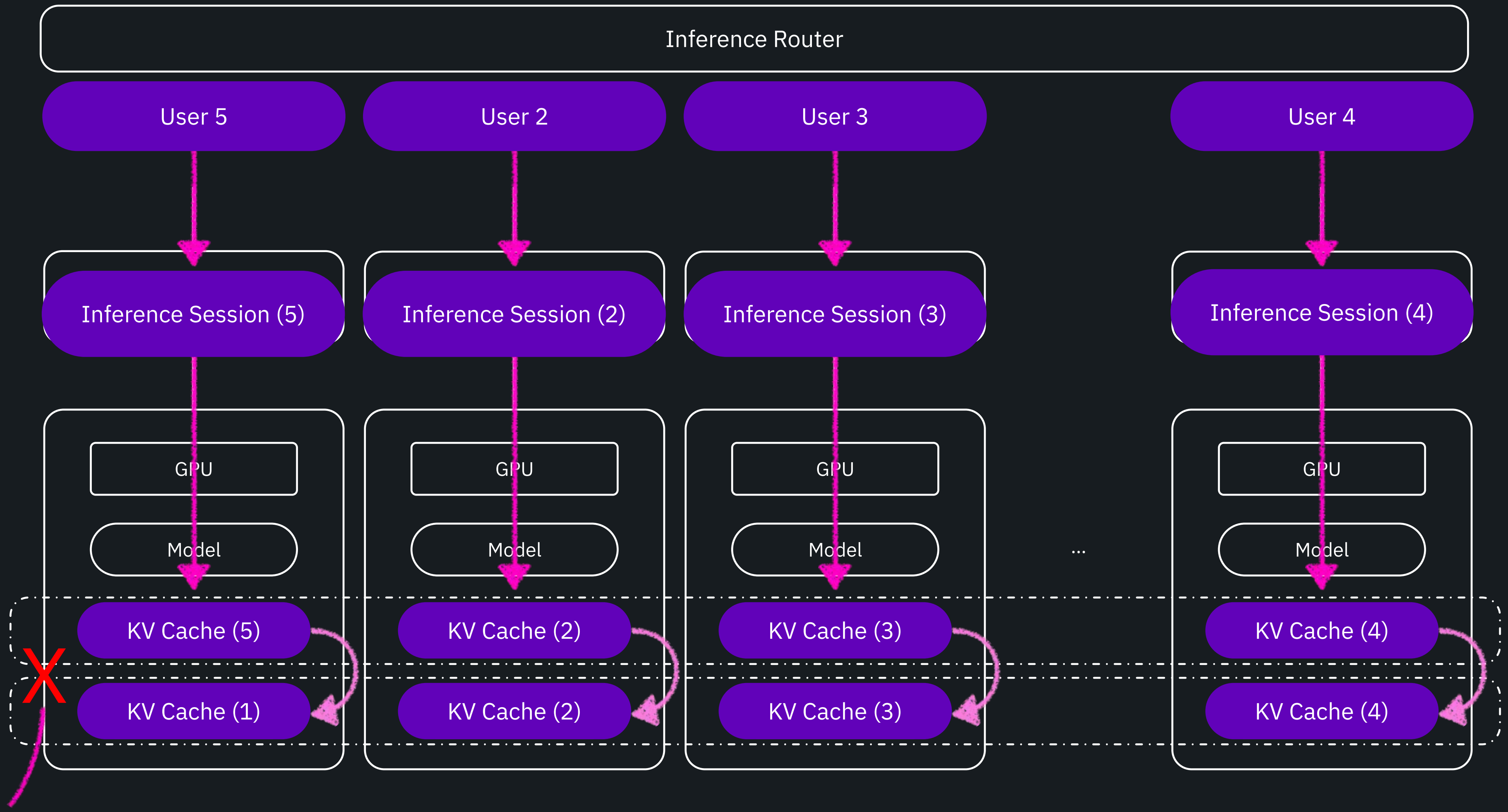
How many people live in Australia and New Zealand?

The **partial prefix match** is only part of the new prompt matches cached content; the rest needs prefill.

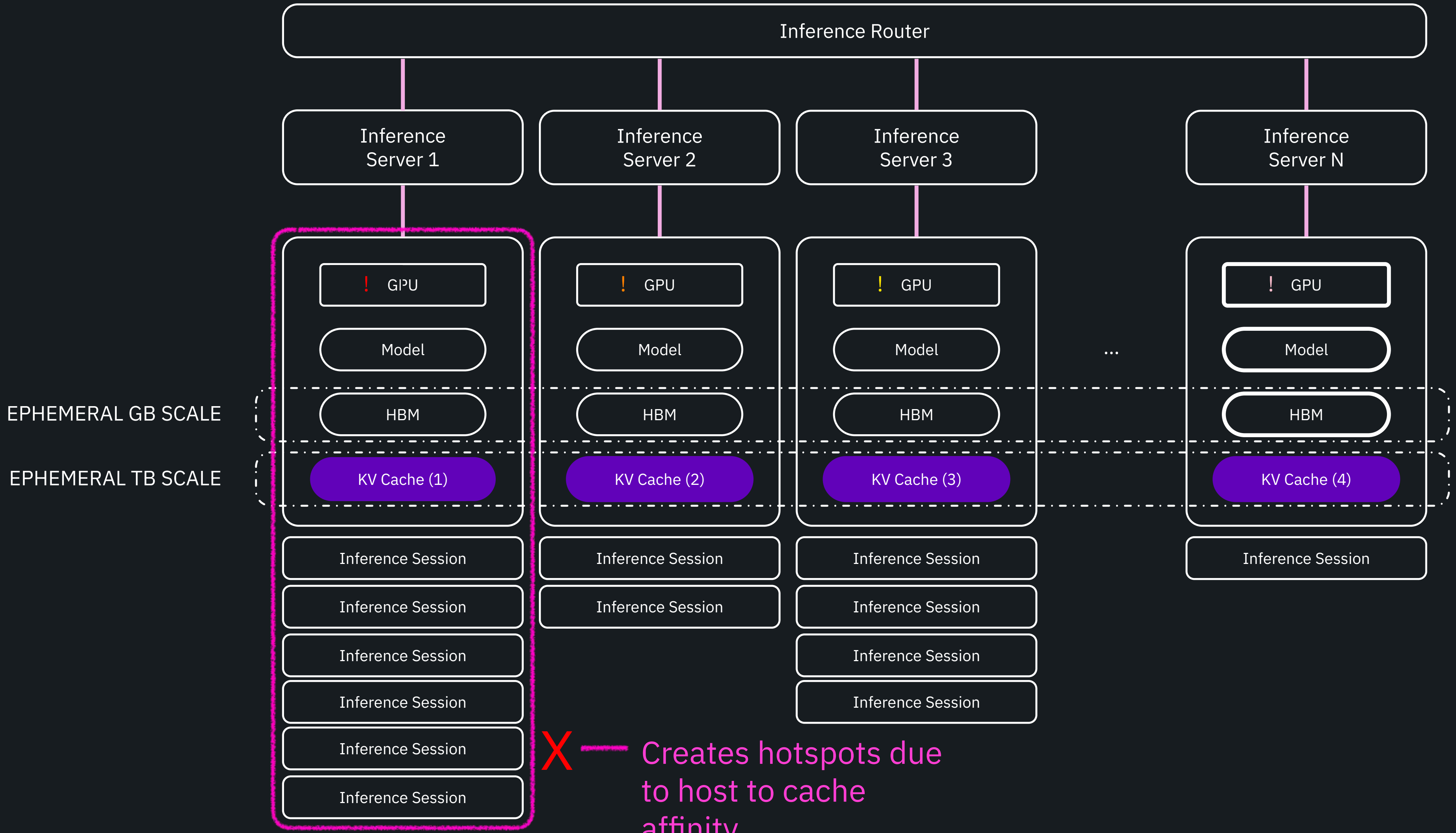
THE GOAL

Faster AI at Lower Costs

Implementing Caching Algorithms



Purged due to lack of sizable centralized cache



Creates hotspots due to host to cache affinity

WEKA®

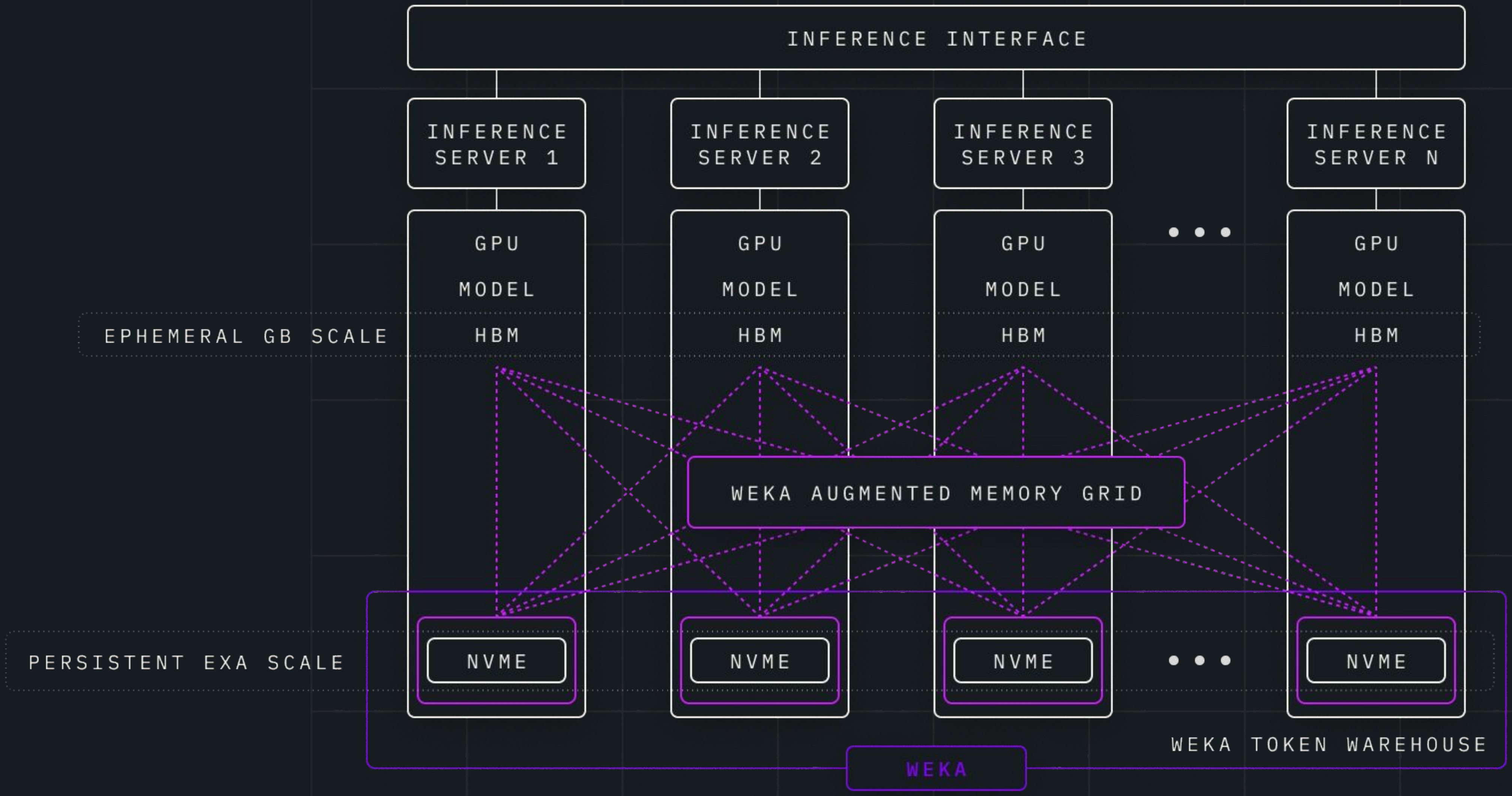
Challenges in Production Systems

Slow Time To First Token (TTFT) for complex workloads and long context cache.

Significant periods of under-utilization.

Cache hotspots often leave other inference systems under-utilized.

How It Works


INFERENCE INTERFACE
INFERENCE SERVER 1
INFERENCE SERVER 2
INFERENCE SERVER 3
INFERENCE SERVER N
GPU
MODEL
HBM
EPHEMERAL GB SCALE
WEKA AUGMENTED MEMORY GRID
PERSISTENT EXA SCALE
NVME
NVME
NVME
NVME
WEKA TOKEN WAREHOUSE
WEKA

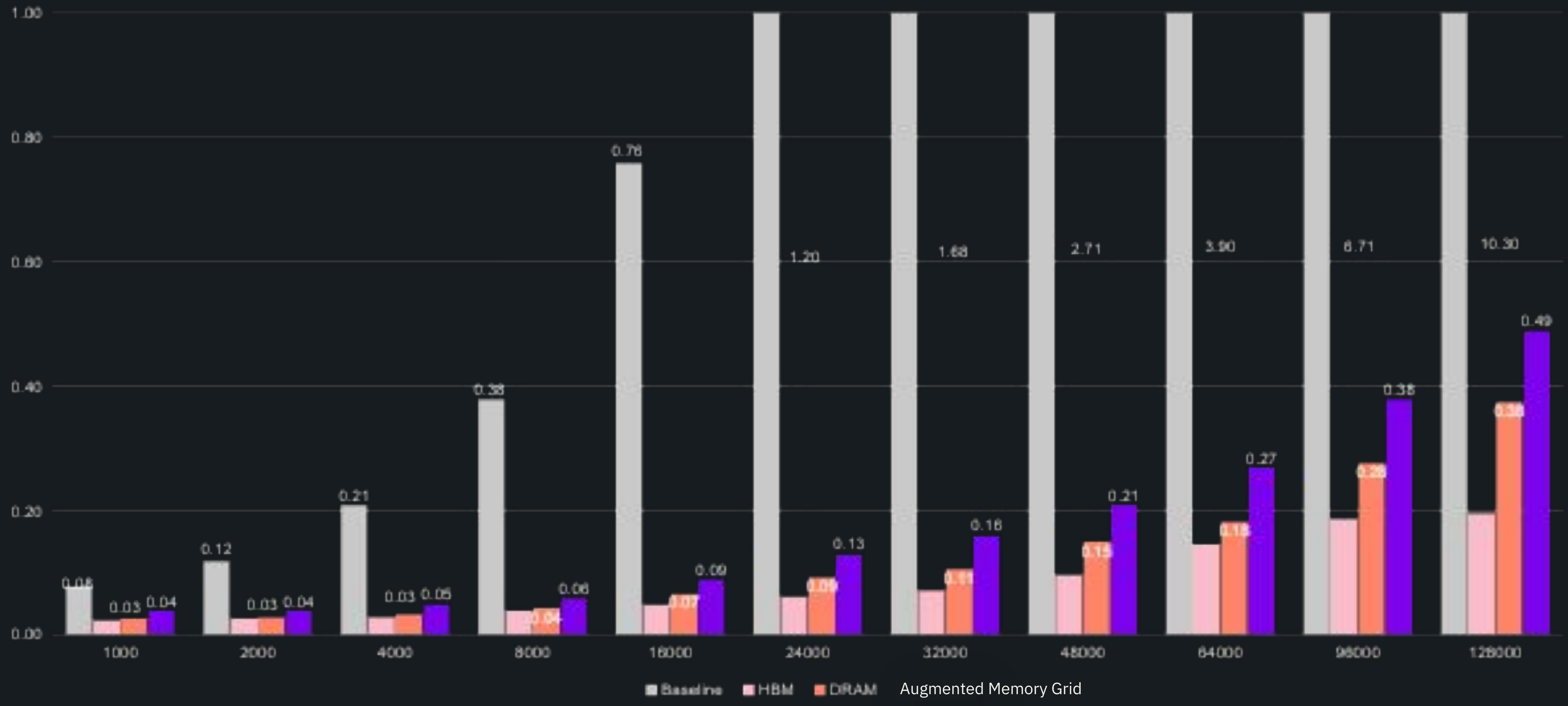
WEKA®

Impact of KV Cache in LLMs

- Faster Time To First Token (TTFT)
- Better token throughput cluster-wide
- Fewer GPUs needed to achieve overall volume of inference for current and future Service Level Agreements (SLAs)
- More consistent Quality of Service (QoS)

Insights from Our Labs

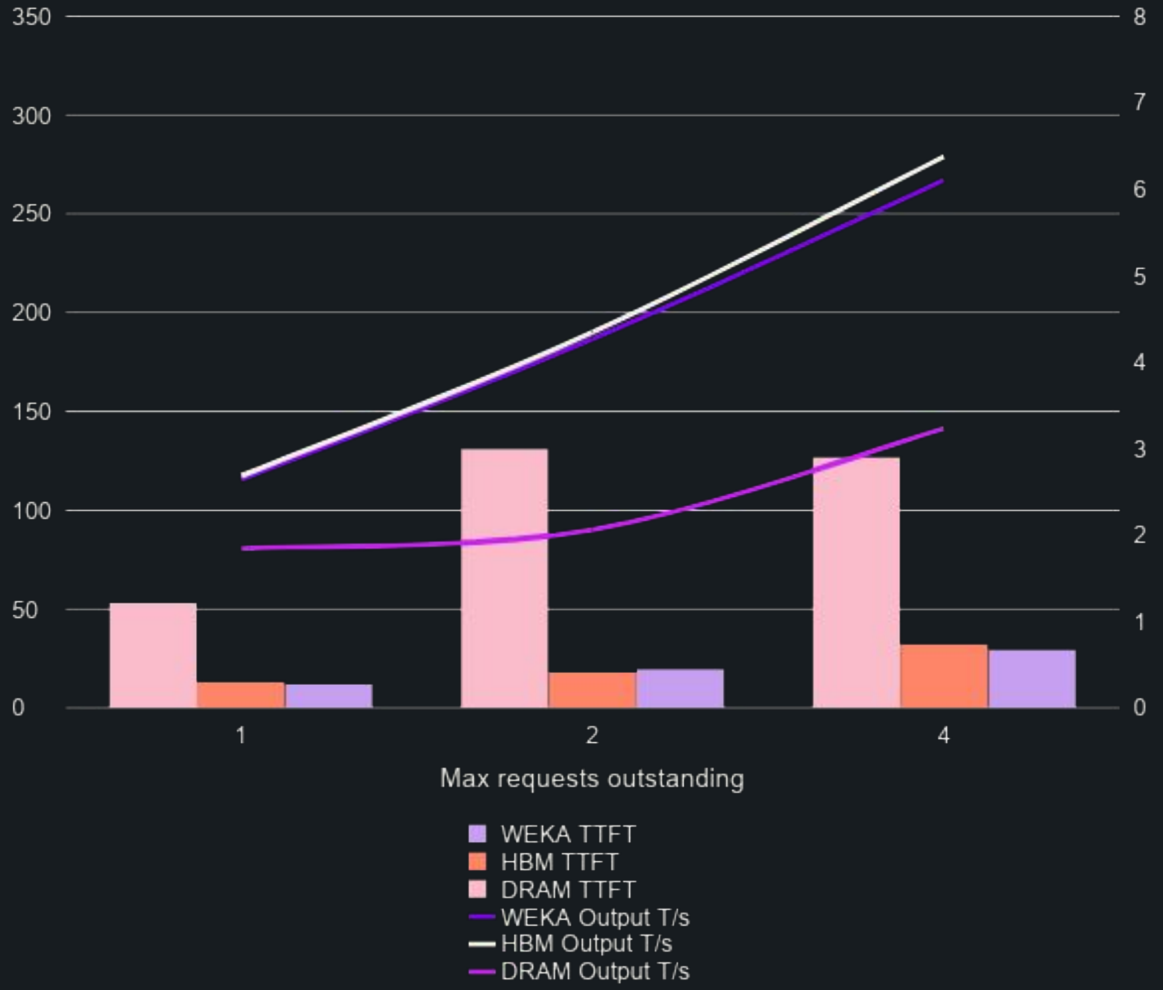
Inference Performance with Single Shot Prompts



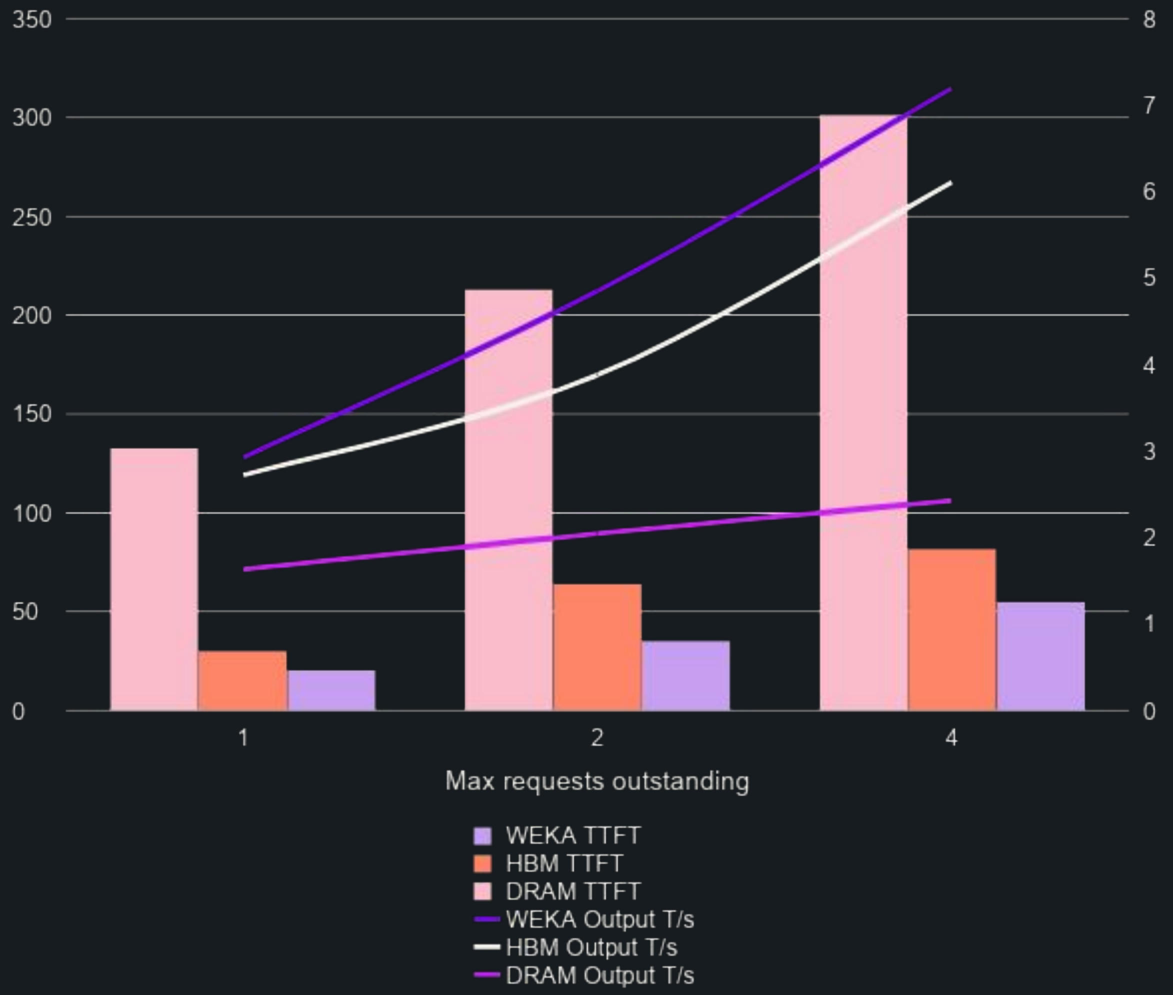
*5 runs, we took the average

Inference Performance with Concurrent Users

Qwen3-Coder-30B-A3B @ 60,000

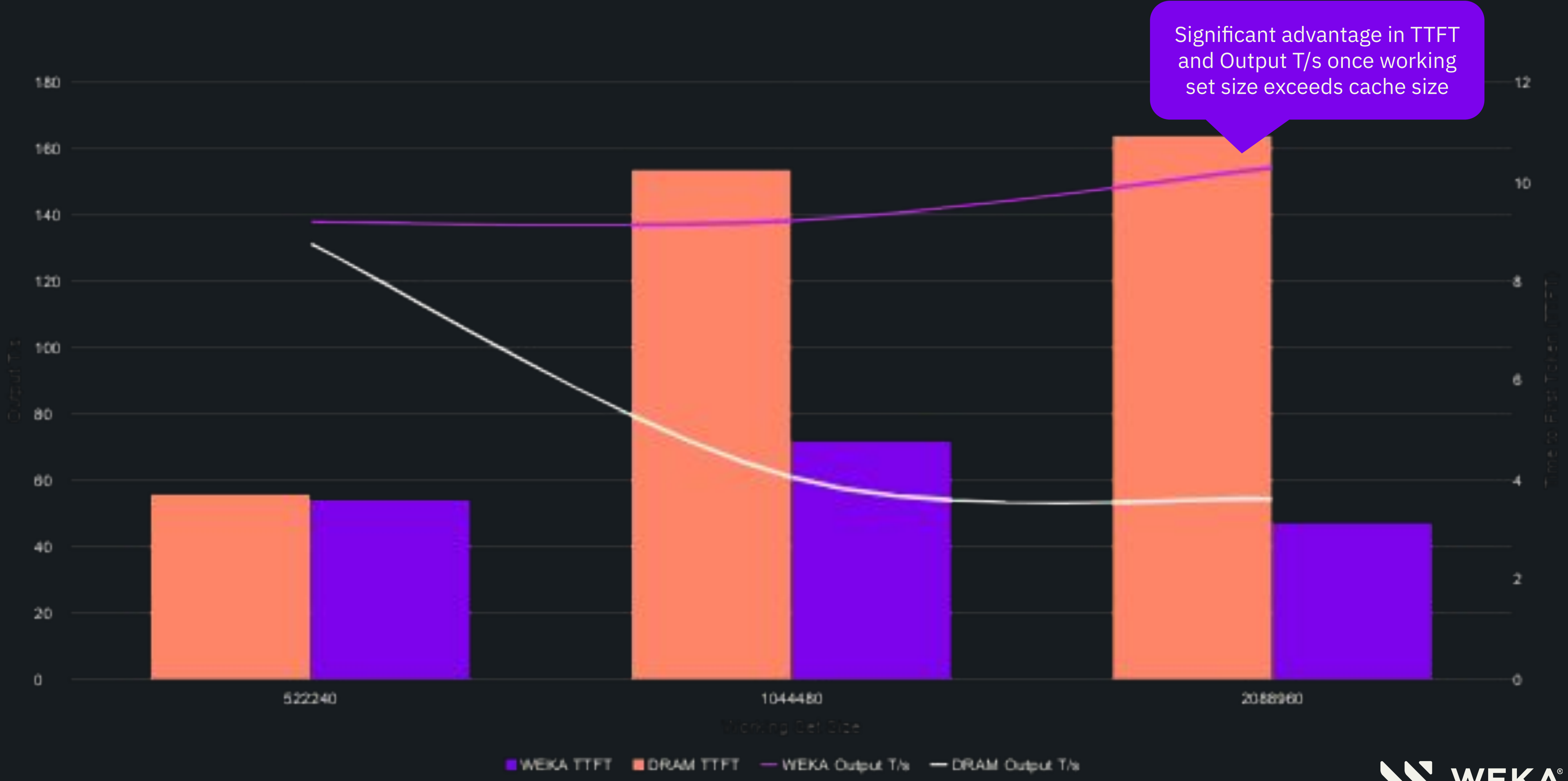


Llama-3.3-70B @ 32000

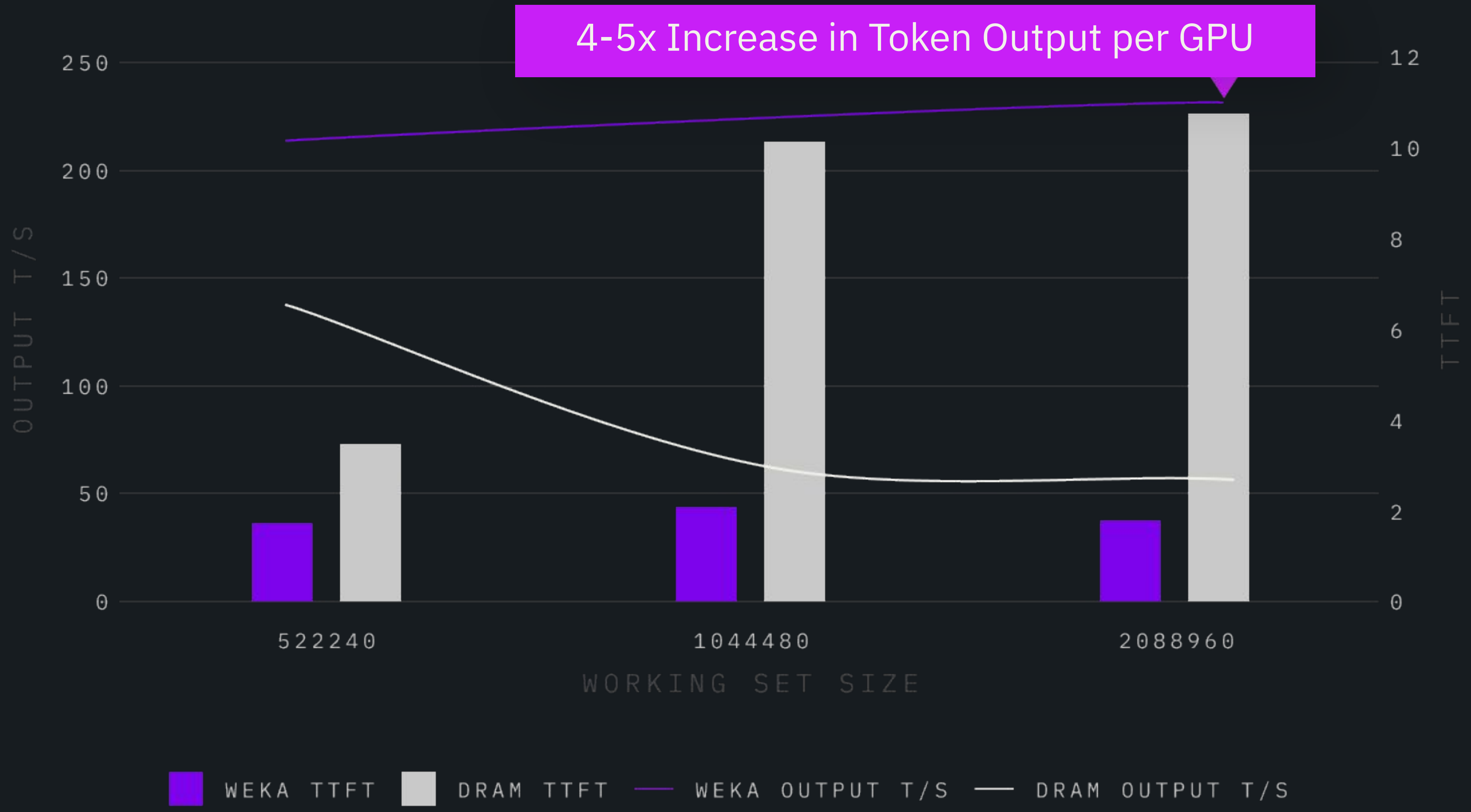


WEKA®

Inference Performance with Large Cache



Significant advantage in TTFT and Output T/s once working set size exceeds cache size



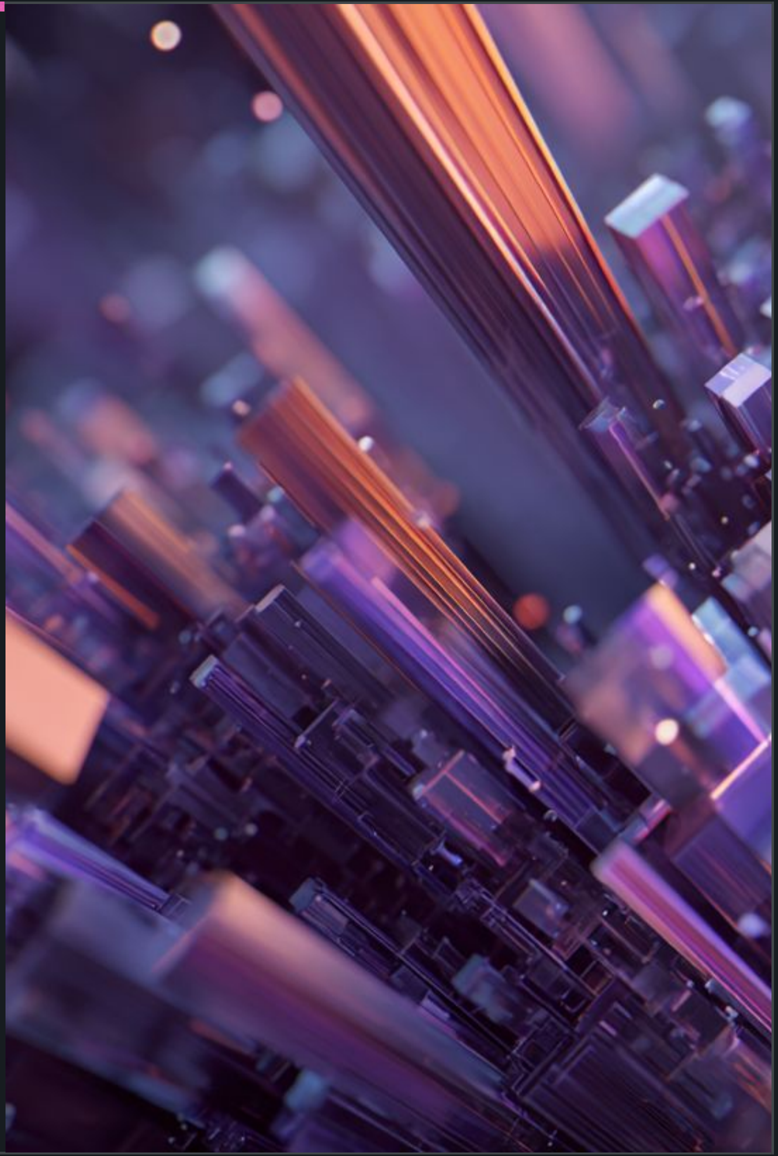
4-5x Increase in Token Output per GPU
OUTPUT T/S
250
200
150
100
50
0
TTFT
12
10
8
6
4
2
0
522240
1044480
2088960
WORKING SET SIZE
WEKA TTFT
DRAM TTFT
WEKA OUTPUT T/S
DRAM OUTPUT T/S

Key Takeaways

AI algorithms are designed for isolated environments, not those that operate at scale.

AI Inference at Scale is challenging but needs of practitioners are easy to address with the right techniques.

TCO/ROI objectives can't be solved with throwing more compute at the problem. You can solve inference SLAs with orders of magnitude less expensive infrastructure.



THANKS FOR YOUR TIME

Learn How to Maximize Your AI Token Production