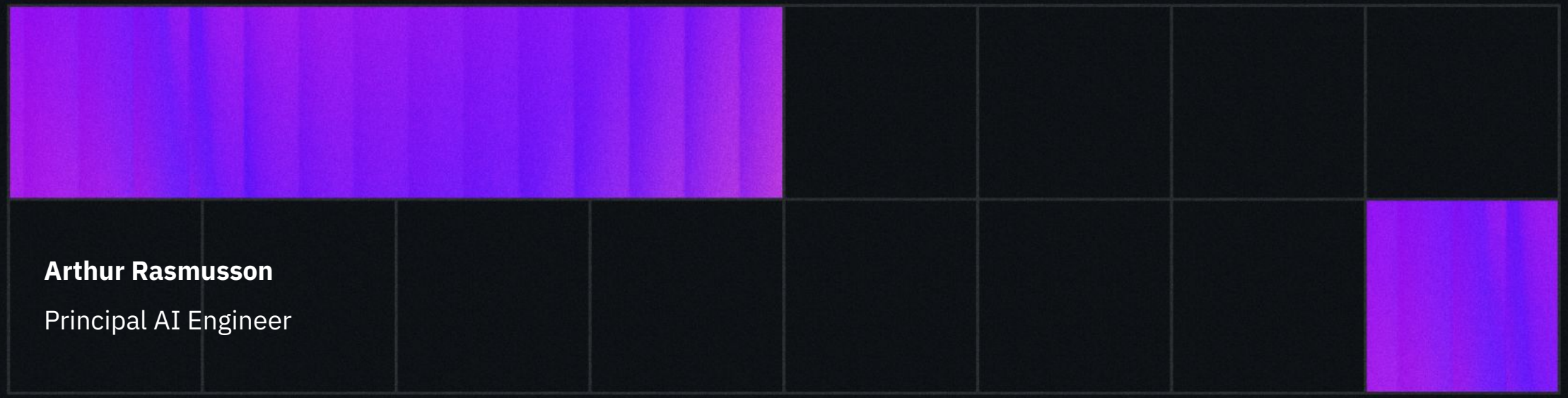
WEKA®

Leverage Caching Algorithms

Improve Your Token Economics



Arthur Rasmusson

Principal AI Engineer

THE GOAL

Faster AI at Lower Costs

Why It Matters

AI Inference is majority of lifecycle cost for production LLMs.

KV-cache efficiency directly affects $/tokens.

At scale, AI cache ≈ GPU utilization; gains translate to millions of dollars in annual savings for cloud and infrastructure bills.

Workshop Goal: leave with concrete patterns you can implement **this quarter**.

Workshop Agenda

1. **Inference Practitioner Perspective**
2. **Impact of Caches in LLMs**
3. **Challenges**
4. **Implementation**

AI Inference Practitioner Perspective

Arthur Rasmusson

- **Principal AI Engineer, WEKA**
- Model Efficiency & Infrastructure Teams, Cohere
- Founding Contributor, Open-IOV.org
- Co-Founder & Chief Operations Officer (COO), Arc Compute

WEKA

Observations from Working with LLM Production Systems

- AI Inference systems often do not incorporate a caching mechanism.
- Some practitioners have opted to disable caching due to routing complexity.
- Perceived challenge our routing prompts round robin and moving to cache aware routing. "This is too complicated"

Impact of Caches 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)

Top Challenges

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

Implementing Caching Algorithms

Implementation

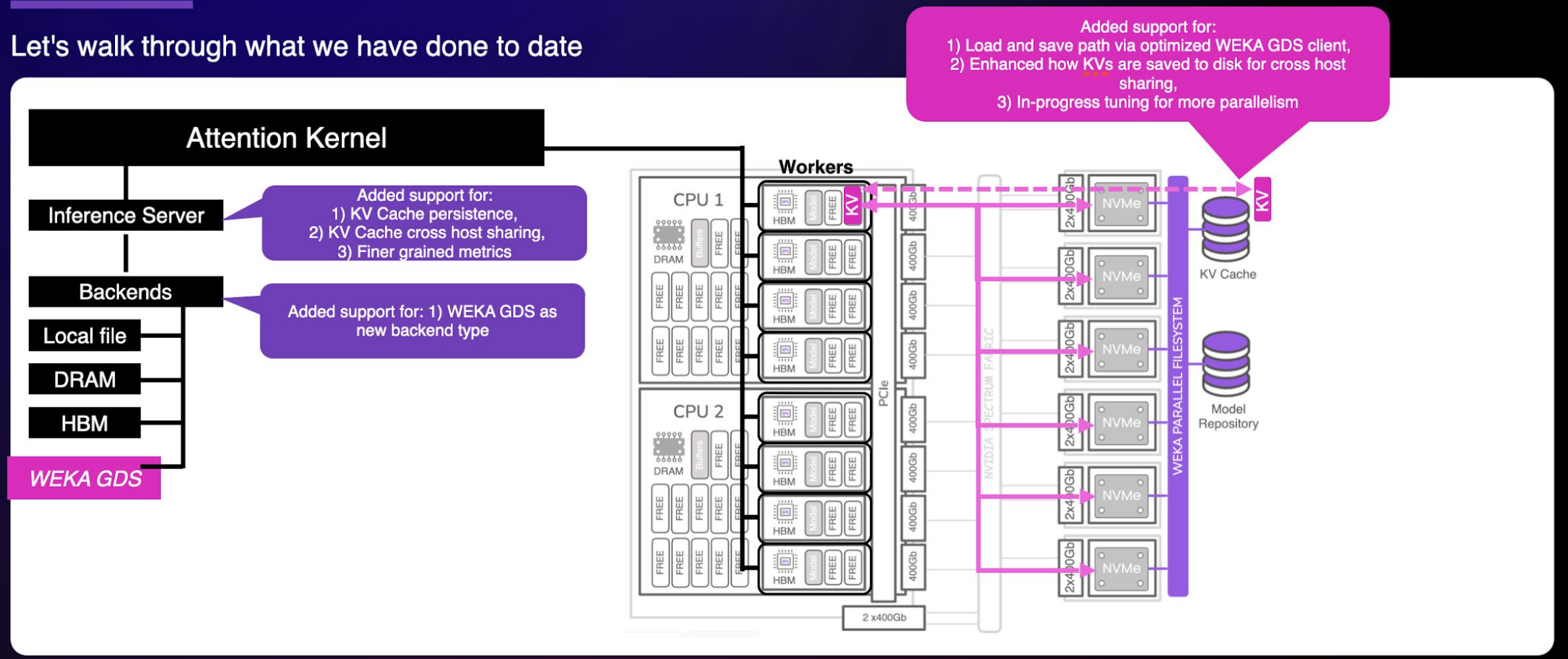
- Providers often think implementing cache aware routing is difficult, and they're right! (PagedAttention over RDMA eliminates challenges by distributing KV cache when and where it's needed so you don't have to think about routing).

- Even with cache aware routing there are challenges.

How It Works

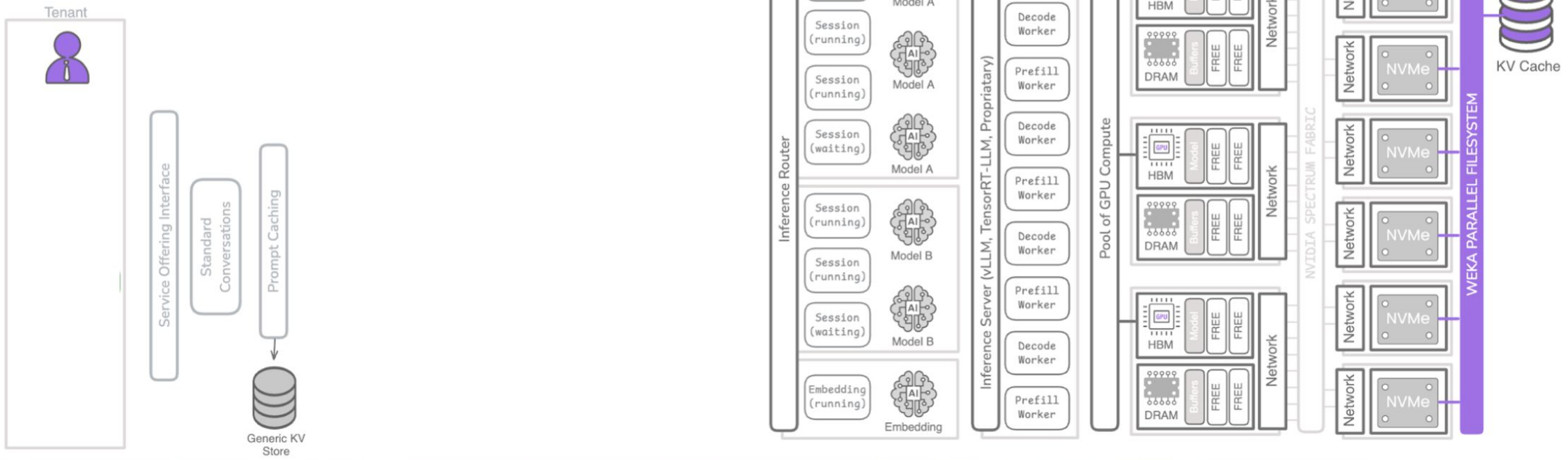
Architectural Changes

Let's walk through what we have done to date



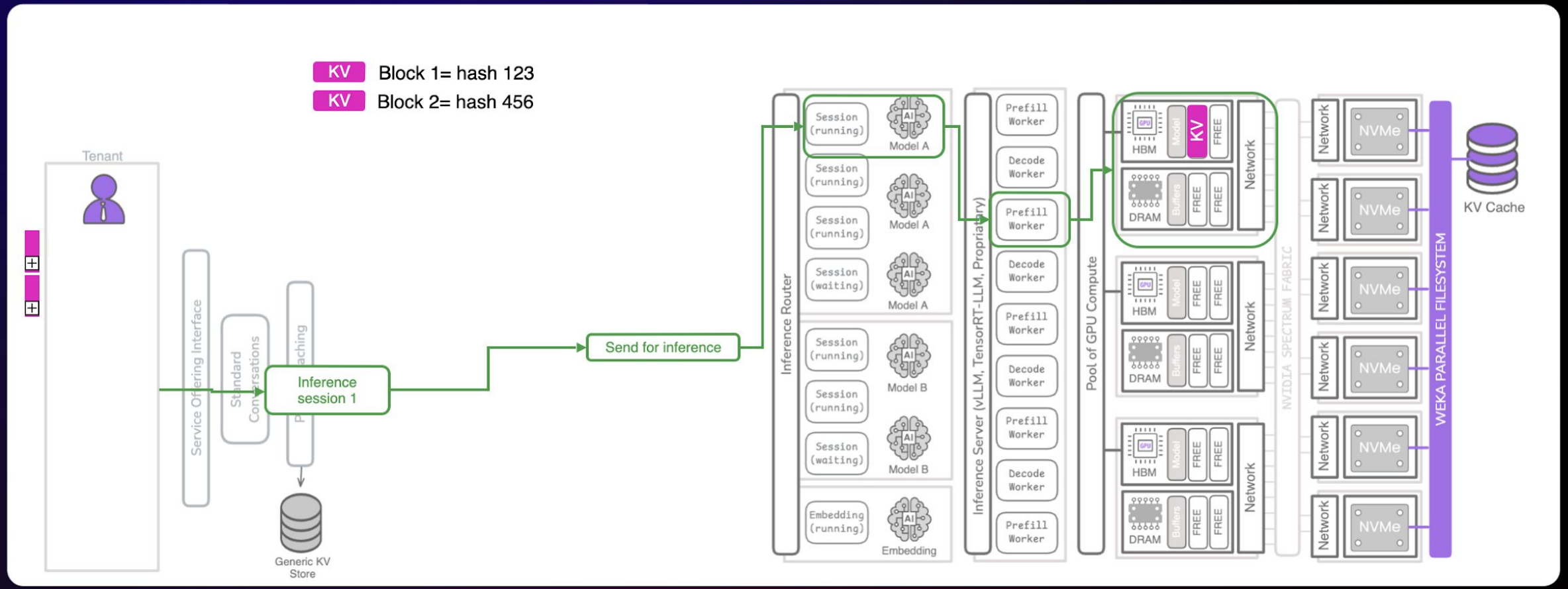
Walkthrough: Life of a Prompt (1)

Let's walkthrough how **“prefix caching”** will help the inference workflows (slides have buildout)



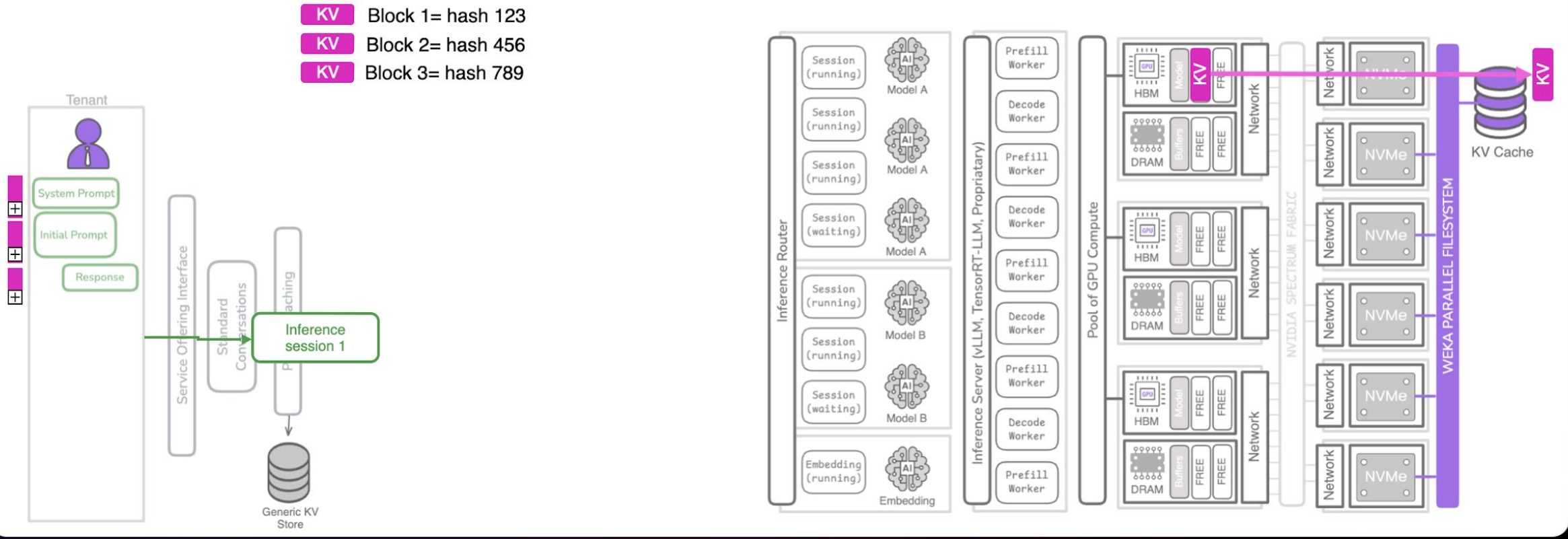
Walkthrough: Life of a Prompt (2)

Let's walkthrough how **“prefix caching”** will help the inference workflows (slides have buildout)



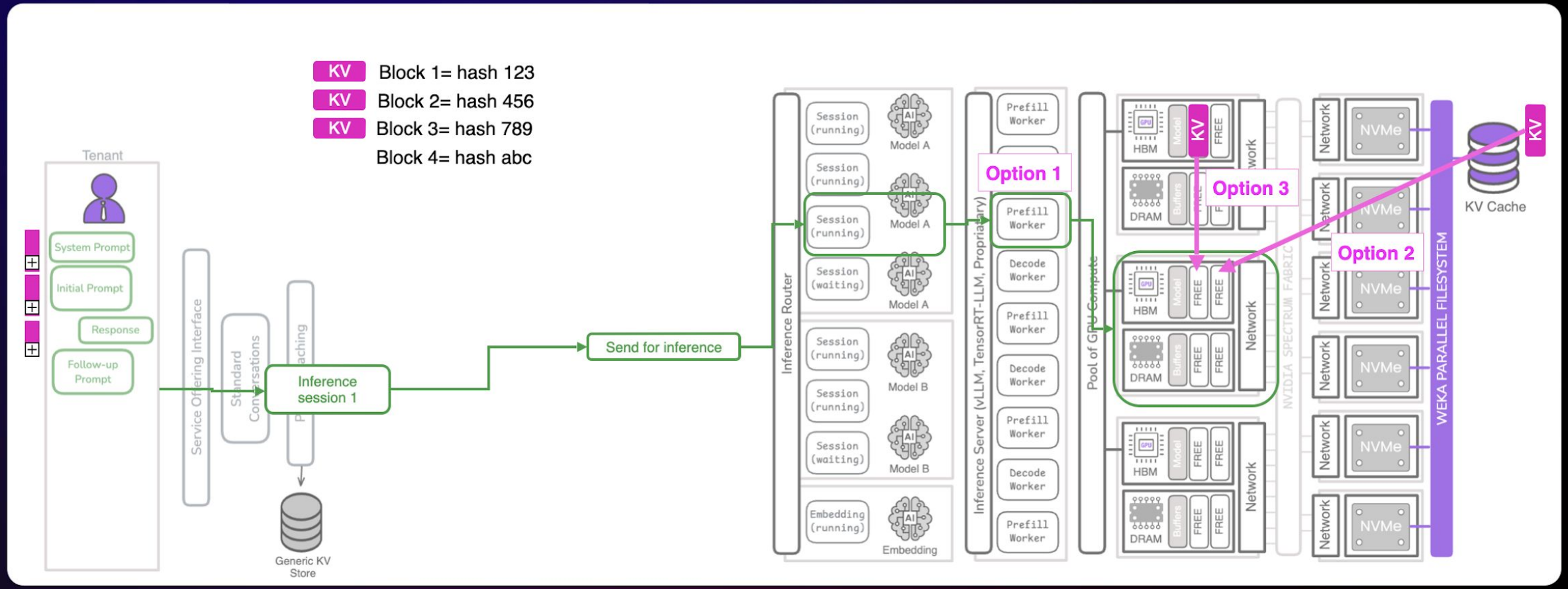
Walkthrough: Life of a Prompt (4)

Let's walkthrough how **“prefix caching”** will help the inference workflows (slides have buildout)



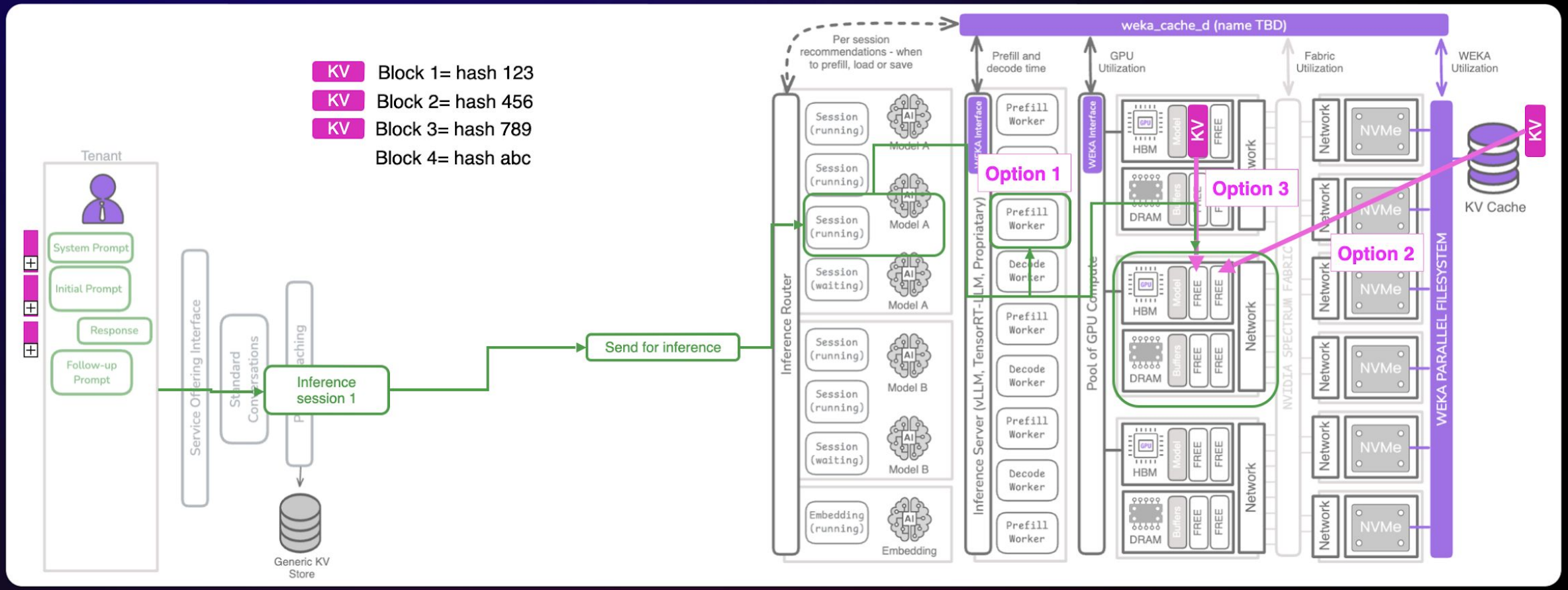
Walkthrough: Life of a Prompt (5)

Let's walkthrough how **“prefix caching”** will help the inference workflows (slides have buildout)



Walkthrough: Life of a Prompt (6)

Let's walkthrough how **“prefix caching”** will help the inference workflows (slides have buildout)

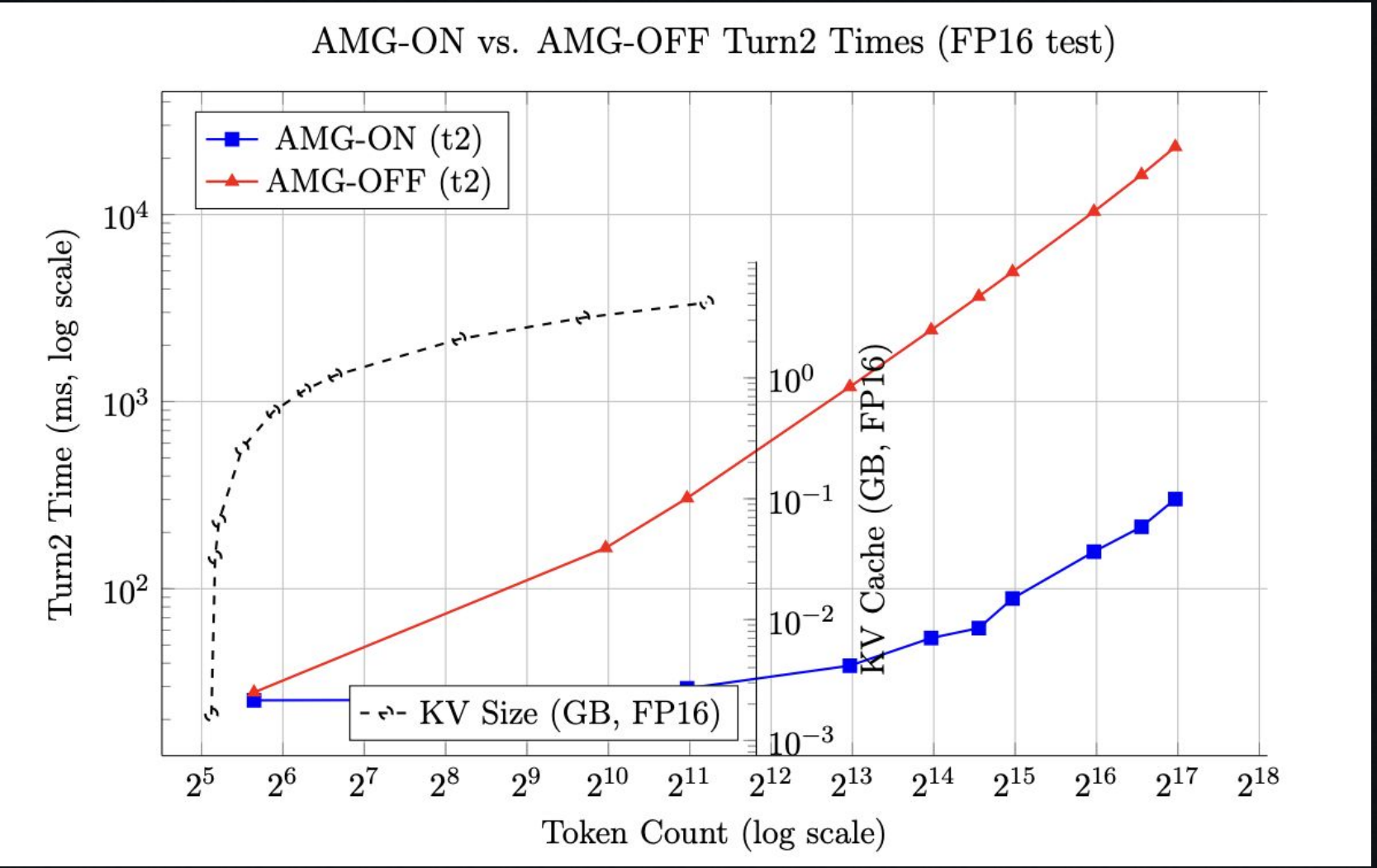


What To Expect

Insights from Our Labs

WEKA®

7,528.53% faster Time to First Token (TTFT)
in multi-round QA with Llama-3.1-70B at FP16

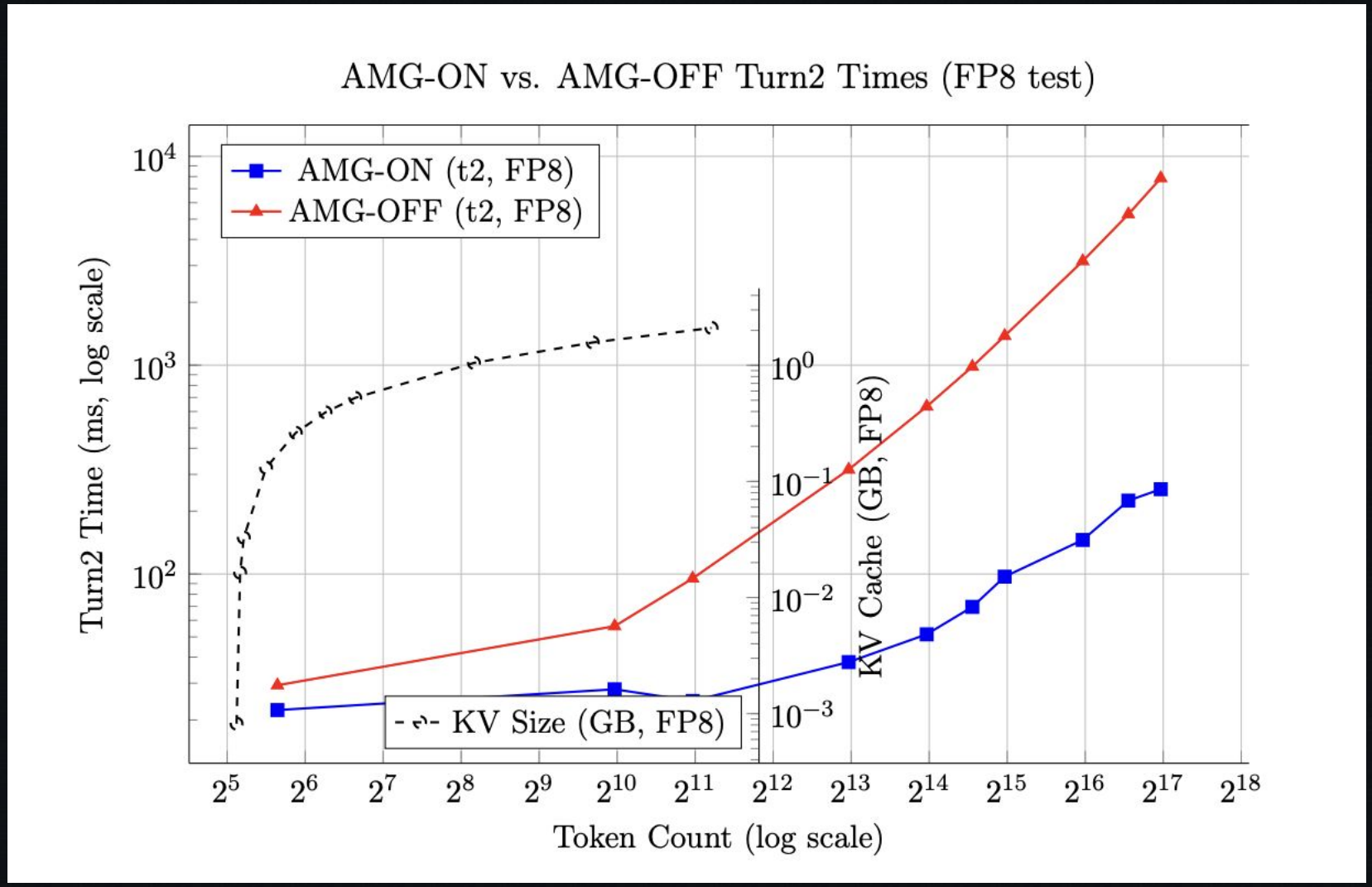


7,528.53% faster Time to First Token (TTFT)
in multi-round QA with Llama-3.1-70B at FP16

Tokens	AMG-ON t1	AMG-ON t2	AMG-OFF t1	AMG-OFF t2	% diff (t2)
50	36.894	25.328	29.081	27.925	10.25
1000	165.612	25.538	166.805	165.260	547.12
2000	305.454	29.350	304.692	304.856	938.69
8000	1189.103	38.787	1199.765	1199.311	2992.03
16000	2394.289	54.519	2410.260	2412.058	4324.28
24000	3612.110	61.598	3652.862	3648.305	5822.81
32000	4893.061	88.661	4931.505	4934.670	5465.75
64000	10288.430	157.749	10360.978	10369.098	6473.17
96000	16256.742	214.237	16304.672	16312.039	7514.03
128000	22981.840	301.481	22994.476	22998.583	7528.53

Table 1: **FP16 test, AMG-ON vs. AMG-OFF Turn2 Times.** Gains up to 7528.53% at 128k tokens.

2984.65% faster Time to First Token (TTFT)
in multi-round QA with Llama-3.1-70B at FP8



2984.65% faster Time to First Token (TTFT) in multi-round QA with Llama-3.1-70B at FP8

Tokens	AMG-ON t1	AMG-ON t2	AMG-OFF t1	AMG-OFF t2	% diff (t2)
50	39.439	22.308	43.725	29.327	31.47
1000	58.850	28.046	56.905	56.246	100.55
2000	97.152	24.662	95.072	95.206	286.05
8000	317.714	37.862	316.145	316.258	735.30
16000	641.439	51.424	633.103	634.030	1132.96
24000	1007.591	69.586	984.308	983.485	1313.34
32000	1367.059	97.098	1371.541	1381.308	1322.60
64000	3156.745	145.569	3143.190	3149.996	2063.92
96000	5290.473	224.575	5277.509	5279.433	2250.86
128000	7860.608	254.781	7846.349	7859.110	2984.65

Table 2: **FP8 test, AMG-ON vs. AMG-OFF Turn2 Times.** Gains up to 2984.65% at 128k tokens.

Best Practices

- Test TTFT before implementing caching software.
- Add caching software into inference stack.
- Measure TTFT performance after implementation.

Factors to Consider

- Which inference server are you using (vLLM, TensorRT-LLM)?
- What kind of fabric is available (RDMA, or non-RDMA)?
- Which kinds of accelerators (NVIDIA GPUs, AMD GPUs, Tenstorrent Blackhole, Intel Gaudi?)
- Use of NVIDIA NIM?
- Which routers are involved if any (Dynamo+NIXL, Triton)?

How to Start

- Start out with a test cluster.
- Measure results with POC.
- Plan staging cluster rollout.
- Schedule rollout during off-hours.
- Measure stability of staging cluster over a pre-determined stability window.

How to Start

- Plan production rollout.
- Train on-call production support staff (infrastructure team) on inference stack.
- Roll out during off-peak hours.
- Measure production efficiency gains.
- Publish a case study.

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.

WEKA®

THANKS FOR YOUR TIME

Learn How to Maximize Your AI Token Production

WEKA's Latest Open-Source Contributions

Tokens	AMG-ON t1	AMG-ON t2	AMG-OFF t1	AMG-OFF t2	% diff (t2)
50	39.439	22.308	43.725	29.327	31.47
1000	58.850	28.046	56.905	56.246	100.55
2000	97.152	24.662	95.072	95.206	286.05
8000	317.714	37.862	316.145	316.258	735.30
16000	641.439	51.424	633.103	634.030	1132.96
24000	1007.591	69.586	984.308	983.485	1313.34
32000	1367.059	97.098	1371.541	1381.308	1322.60
64000	3156.745	145.569	3143.190	3149.996	2063.92
96000	5290.473	224.575	5277.509	5279.433	2250.86
128000	7860.608	254.781	7846.349	7859.110	2984.65

Table 2: **FP8 test, AMG-ON vs. AMG-OFF Turn2 Times.** Gains up to 2984.65% at 128k tokens.

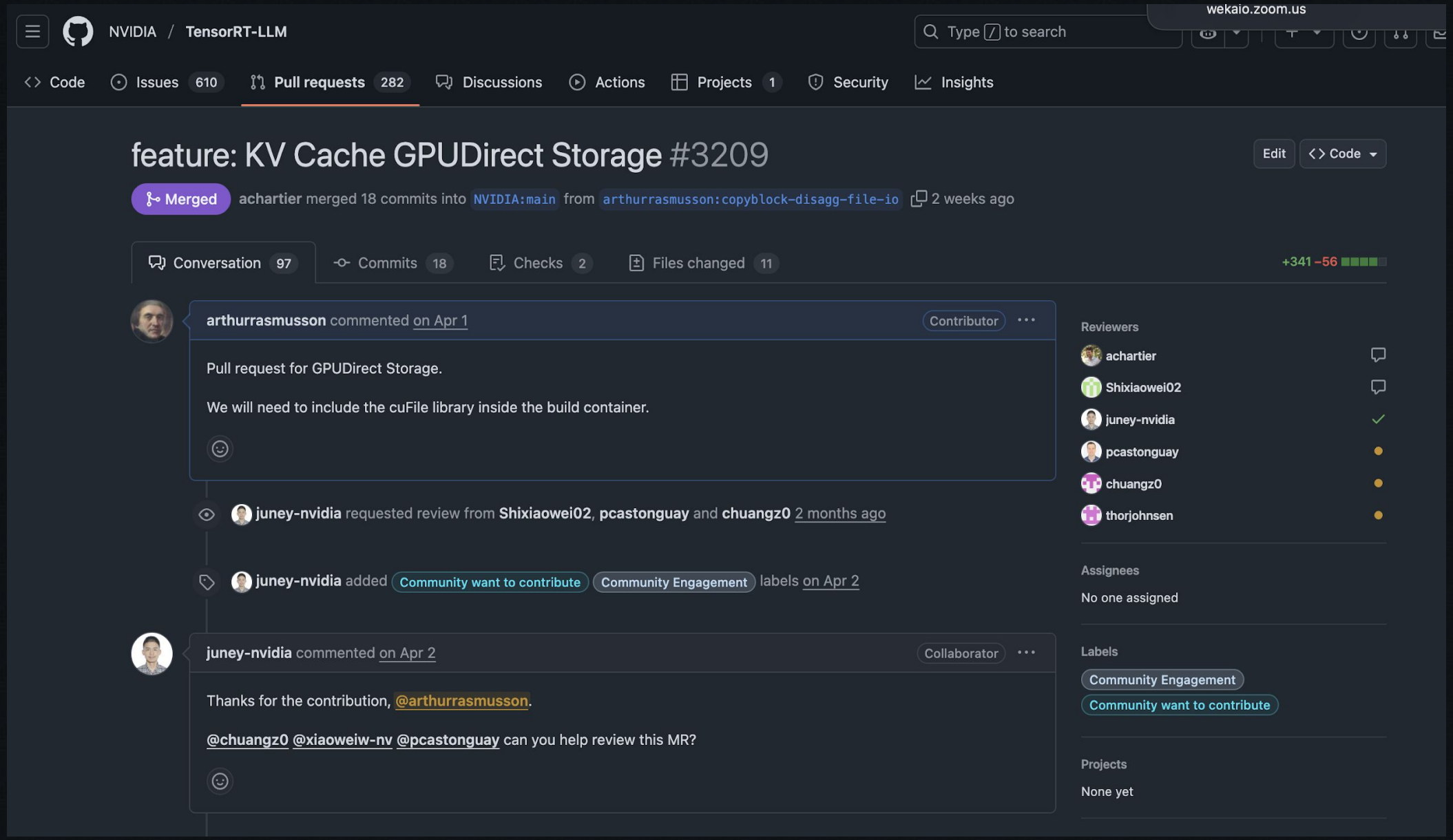
2984.65% faster Time to First Token (TTFT)
in multi-round QA with Llama-3.1-70B at FP8

WEKA's Latest Open-Source Contributions

Tokens	AMG-ON t1	AMG-ON t2	AMG-OFF t1	AMG-OFF t2	% diff (t2)
50	36.894	25.328	29.081	27.925	10.25
1000	165.612	25.538	166.805	165.260	547.12
2000	305.454	29.350	304.692	304.856	938.69
8000	1189.103	38.787	1199.765	1199.311	2992.03
16000	2394.289	54.519	2410.260	2412.058	4324.28
24000	3612.110	61.598	3652.862	3648.305	5822.81
32000	4893.061	88.661	4931.505	4934.670	5465.75
64000	10288.430	157.749	10360.978	10369.098	6473.17
96000	16256.742	214.237	16304.672	16312.039	7514.03
128000	22981.840	301.481	22994.476	22998.583	7528.53

Table 1: **FP16 test, AMG-ON vs. AMG-OFF Turn2 Times.** Gains up to 7528.53% at 128k tokens.

7,528.53% faster Time to First Token (TTFT)
in multi-round QA with Llama-3.1-70B at FP16


wekaio.zoom.us
NVIDIA / TensorRT-LLM
Type / to search
Code
Issues 610
Pull requests 282
Discussions
Actions
Projects 1
Security
Insights
feature: KV Cache GPUDirect Storage #3209
Edit
Code
Merged
achartier merged 18 commits into NVIDIA:main from arthurrasmusson:copyblock-disagg-file-io 2 weeks ago
Conversation 97
Commits 18
Checks 2
Files changed 11
+341 −56
arthurrasmusson commented on Apr 1
Contributor
Pull request for GPUDirect Storage.
We will need to include the cuFile library inside the build container.
juney-nvidia requested review from Shixiaowei02, pcastonguay and chuangz0 2 months ago
juney-nvidia added Community want to contribute Community Engagement labels on Apr 2
juney-nvidia commented on Apr 2
Collaborator
Thanks for the contribution, @arthurrasmusson.
@chuangz0 @xiaoweiw-nv @pcastonguay can you help review this MR?
Reviewers
achartier
Shixiaowei02
juney-nvidia
pcastonguay
chuangz0
thorjohnsen
Assignees
No one assigned
Labels
Community Engagement
Community want to contribute
Projects
None yet