inferact

vLLM

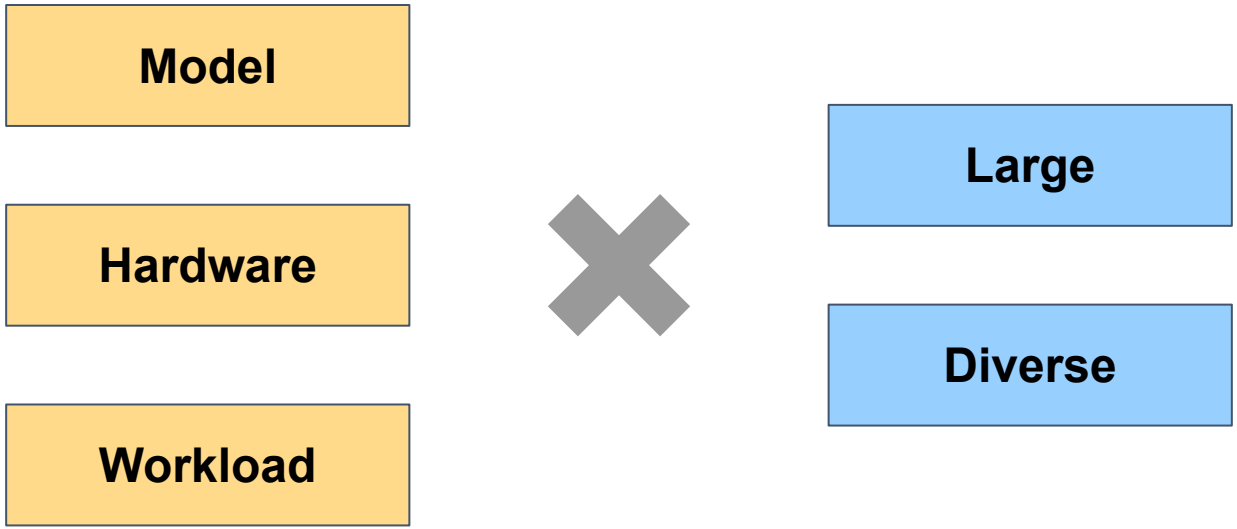
MLSys'26 Lightning Talk

Rethink LLM Inference Abstractions

New Trends and Challenges in LLM Serving

Yifan Qiao
Inferact

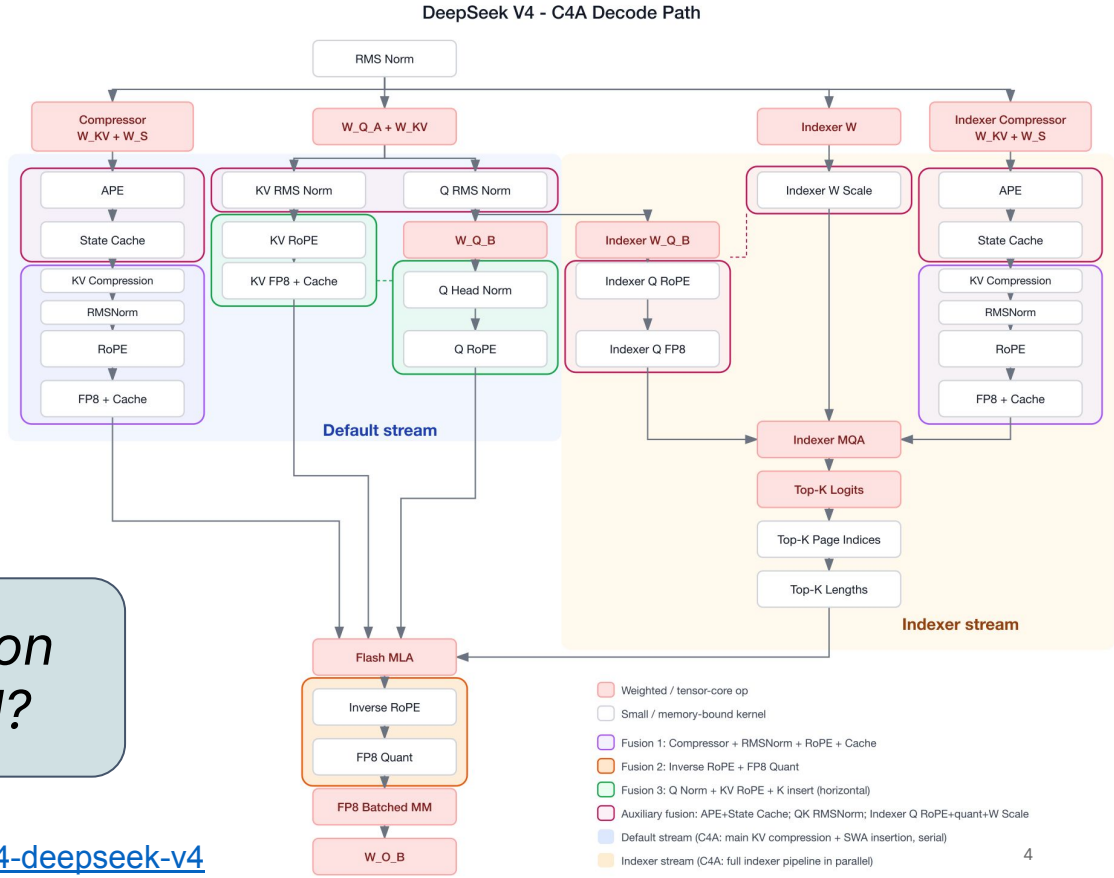
What We Have Seen Today: Combinatorial Explosion



Model Architecture Defines the Serving Path

"The key to understanding complicated things is knowing what not to look at."
– Gerald Jay Sussman

- Complex and diverse kernels
- Code path convoluted by advanced optimizations
  - Kernel fusion, multi-stream, etc.



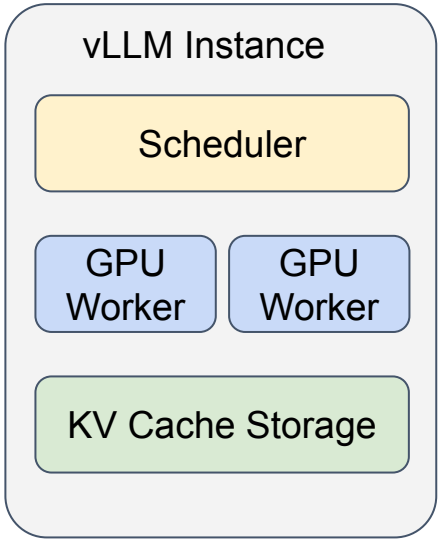
What's the right abstraction for defining a new model?

More details: https://vllm.ai/blog/2026-04-24-deepseek-v4

From Inference Engine to Distributed System

"You can have a second computer once you've shown you know how to use the first one."

– Paul Barham



x N

① Diverse SLAs require heterogeneous scheduling
- PD disaggregation, tiered SLAs

② Models grow larger than a single machine
- DeepSeek V4 Pro: 1.6T params

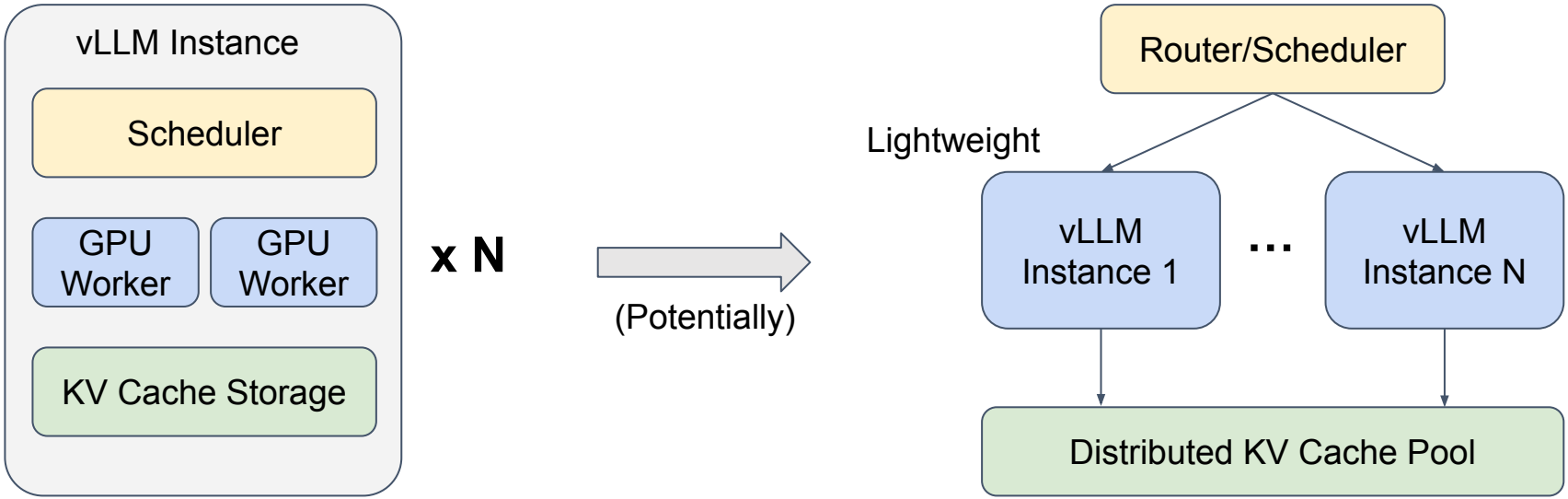
③ Surge traffic asks for cross-instance KV cache reuse
- Multi-turn agentic workloads

What's the right architecture for serving at scale?

From Inference Engine to Distributed System

"You can have a second computer once you've shown you know how to use the first one."

– Paul Barham



More details on distributed KV cache pool: https://vllm.ai/blog/2026-05-06-mooncake-store

inferact

vLLM

The most exciting work in LLM serving lives at the intersection of model architecture and systems.

- Please come talk to us at the booth! (Evergreen Ballroom 2)
- We're hiring!

CONTACT
contact@inferact.ai

WEB
inferact.ai

TWITTER
@inferact

inferact

vLLM

THANK YOU

Come build with us.

CONTACT
contact@inferact.ai

WEB
inferact.ai

TWITTER
@inferact