



Open Source Inference @ Production Scale



Stephen Watt

Vice President, Distinguished Engineer
Research and Emerging Technologies
Red Hat



Carlos Costa

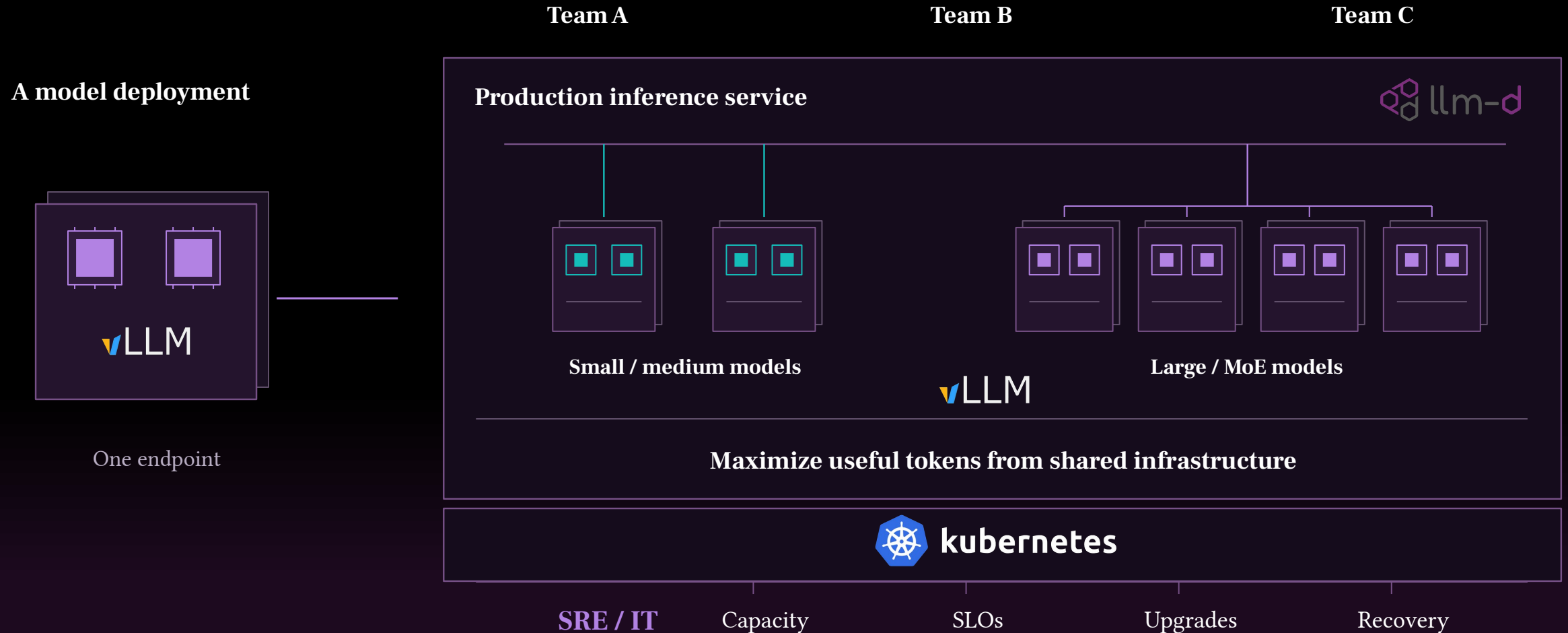
Distinguished Engineer
IBM Research



Abdullah Gharaibeh

Senior Staff Software Engineer
Google Cloud

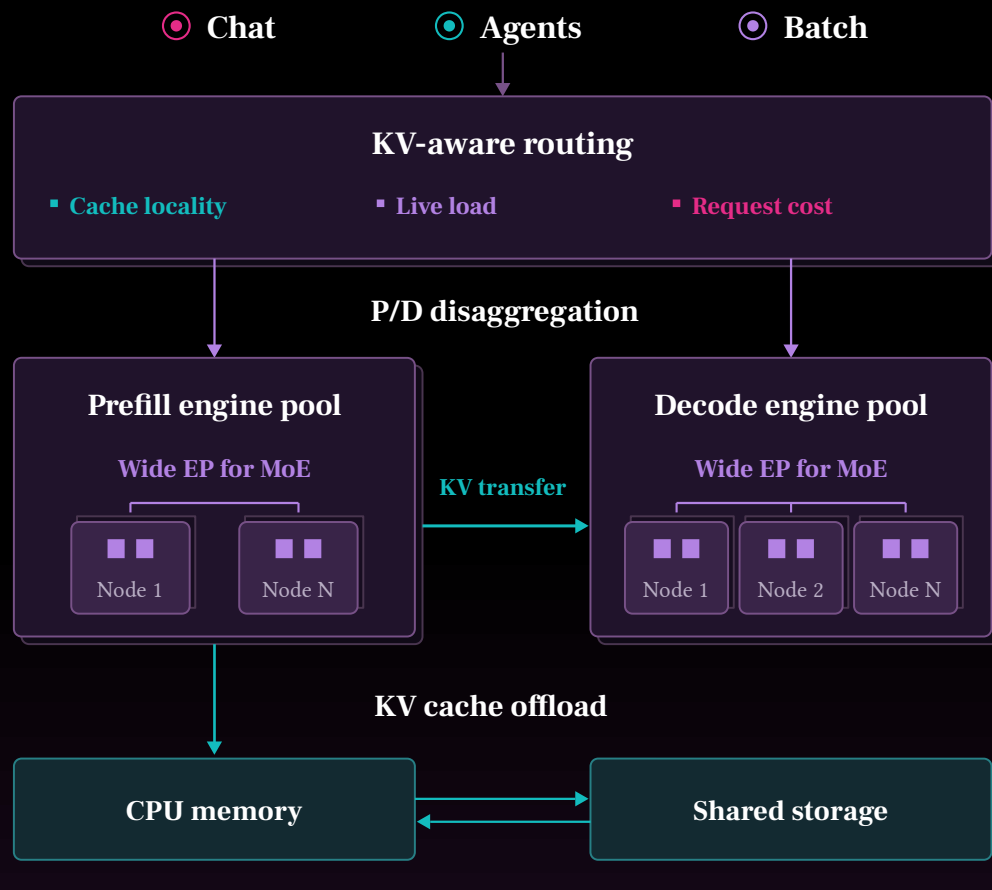
Inference at production scale takes an ecosystem





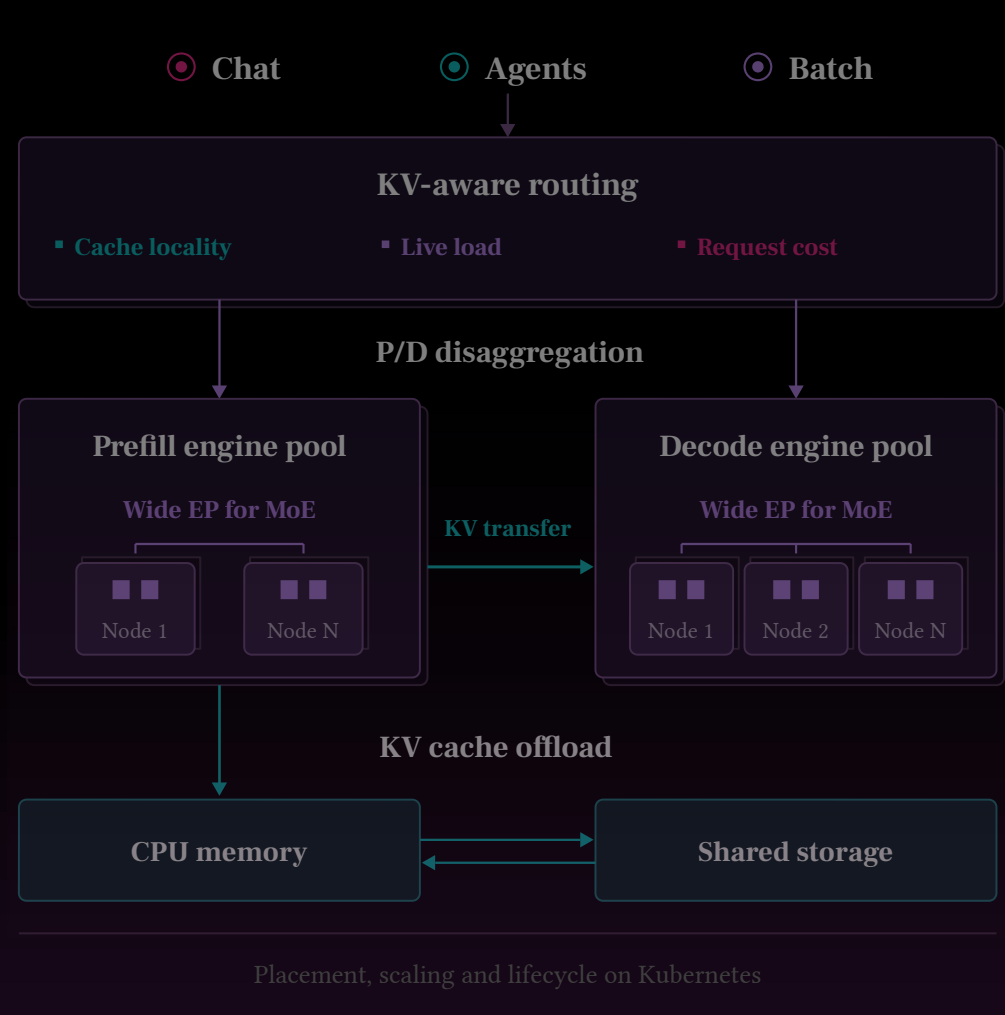
State-of-the-art
Distributed Inference
at Scale.

An industry coalition for inference at scale



Placement, scaling and lifecycle on Kubernetes

An industry coalition for inference at scale



Joined forces

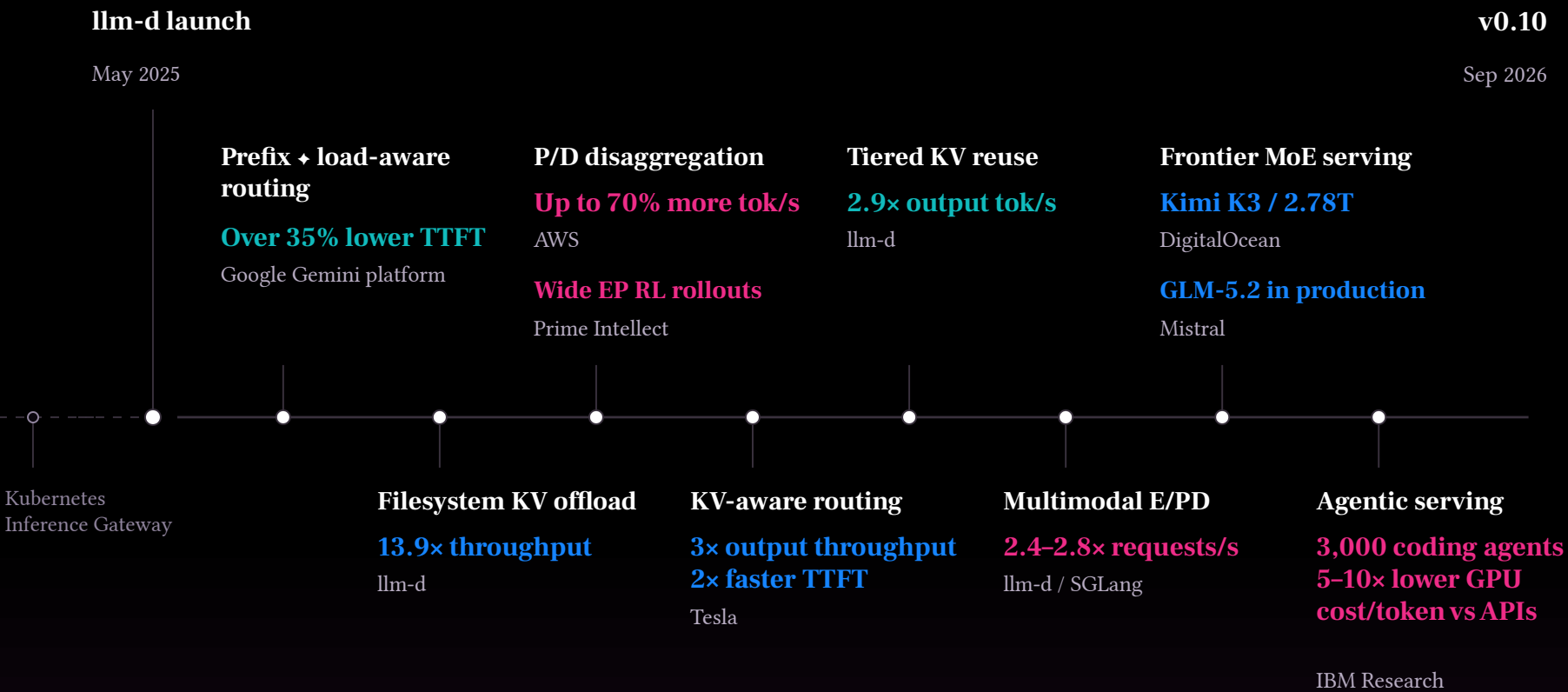
May 2025



Any model, any engine,
any accelerator, any cloud



Our journey to state-of-the-art inference



Any model, any engine,
any accelerator, any cloud



A growing community

 [llm-d/ADOPTERS.md](#)

Type	Name	Since	Website	Description
Founding Contributor	CoreWeave	2025	CoreWeave	CoreWeave is a founding contributor to llm-d, building on their long-standing commitment to open source AI and scalable inference infrastructure.
Founding Contributor	Google Cloud	2025	Google Cloud	Google Cloud is a founding contributor to llm-d, contributing to distributed AI inference at scale and support for Google Cloud TPUs in vLLM.
Founding Contributor	IBM Research	2025	IBM Research	IBM Research is a founding contributor to llm-d, focused on building a hardware-agnostic distributed AI inference platform for enterprise workloads.
Founding Contributor	NVIDIA	2025	NVIDIA	NVIDIA is a founding contributor to llm-d, accelerating the project with innovations from NVIDIA Dynamo including the NVIDIA Inference Xfer Library (NIXL).
Founding Contributor	Red Hat	2025	Red Hat	Red Hat is a founding contributor to llm-d, providing project infrastructure and community stewardship in collaboration with the broader contributor community.
Launch Partner	AMD	2025	AMD	AMD contributes high-performance GPU expertise to advance AI inference for evolving enterprise AI needs.
Launch Partner	Cisco	2025	Cisco	Cisco is exploring real-world use cases that help organizations apply AI more effectively and efficiently through llm-d.
Launch Partner	Hugging Face	2025	Hugging Face	Hugging Face supports llm-d to help developers take AI applications built on the Hugging Face ecosystem to scale.
Launch Partner	Intel	2025	Intel	Intel is a launch partner supporting llm-d as the latest milestone in a decades-long collaboration with Red Hat to empower enterprises with open source solutions.
Launch Partner	Lambda	2025	Lambda	Lambda supports llm-d to make state-of-the-art inference accessible, efficient, and open.
Launch Partner	Mistral AI	2025	Mistral AI	Mistral AI is a launch partner supporting the llm-d community for distributed generative AI inference at scale.
Contributor	DaoCloud	2025	DaoCloud	DaoCloud contributes to llm-d, leveraging P/D disaggregation and KV-cache architectures in its d.run MaaS platform.
Contributor	Moreh	2025	Moreh	Moreh contributes to llm-d, actively improving KV-cache management and vLLM rendering, along with P/D disaggregation for distributed inference.
Contributor	Rebellions	2025	Rebellions	Rebellions contributes to llm-d, enabling support for the NPU accelerator to bring high-performance, energy-efficient AI inference to the platform.
Contributor	Mooncake	2026	Mooncake	Mooncake is integrated with llm-d to support distributed KV cache pooling and prefill/decode disaggregation for large-scale inference.
University Supporter	University of California, Berkeley (Sky Computing Lab)	2025	Sky Computing Lab	The Sky Computing Lab at UC Berkeley, originators of vLLM, are founding academic supporters of llm-d.
University Supporter	University of Chicago (LMCache Lab)	2025	LMCache Lab	The LMCache Lab at the University of Chicago, originators of LMCache, are founding academic supporters of llm-d.
User	Tesla	2025	Tesla	The Tesla ML Platform team are users of llm-d.
User	Cohere	2026	Cohere	The Cohere model serving and RL team are users of llm-d.
User	JetBrains	2026	JetBrains	The JetBrains model serving team are users of llm-d.
User	Stelia.ai	2026	Stelia.ai	Stelia.ai's managed inference service is powered by llm-d.
User	Reflection	2026	Reflection.ai	Reflection uses llm-d's async processing for our internal inference platform.
User	Prime Intellect	2026	Prime Intellect	Prime Intellect uses llm-d in prime-rl for multi-node WideEP RL rollouts.
User	Snowflake	2026	Snowflake	The Snowflake Cortex team are users of llm-d.
User	Novita AI	2026	Novita	The Novita AI team are users of llm-d.
User / Contributor	DigitalOcean	2026	DigitalOcean	DigitalOcean uses llm-d for Dedicated and Serverless Inference endpoints and contributes actively to the project.
User	Capital One	2026	Capital One	The Capital One model serving team are users of llm-d.

Hyperscalers
& neoclouds

MaaS providers

AI labs

Enterprises running
their own inference

500+
contributors

25+
collaborating
companies

Any model, any engine,
any accelerator, any cloud



 **kubernetes**

State-of-the-art inference. **Ready for Day 2.**



1.0

Sustain performance
as the service **evolves**

Stable v1
configuration API

State-of-the-art inference. Ready for Day 2.



1.0

Sustain performance
as the service **evolves**

Stable v1
configuration API

Multi-tenant QoS

Priority + fairness-aware flow control



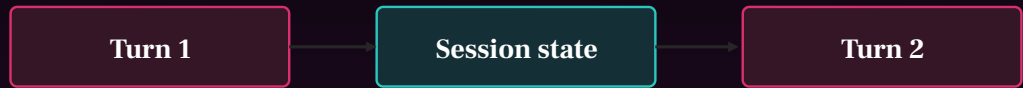
Zero-downtime ops

Router HA + coordinated P/D rollouts



Agent-aware serving

Session affinity, turn priority, Responses state



Any model, any engine,
any accelerator, any cloud





Let's see it in action...

Come build the future of distributed inference with us!



Explore the Well-Lit Paths

Learn more and
contribute

And leave us a star! 



<https://llm-d.ai>



Join Community Slack
llm-d.ai/slack