



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

Evolving KServe: The Unified Model Inference Platform for Both Predictive and Generative AI

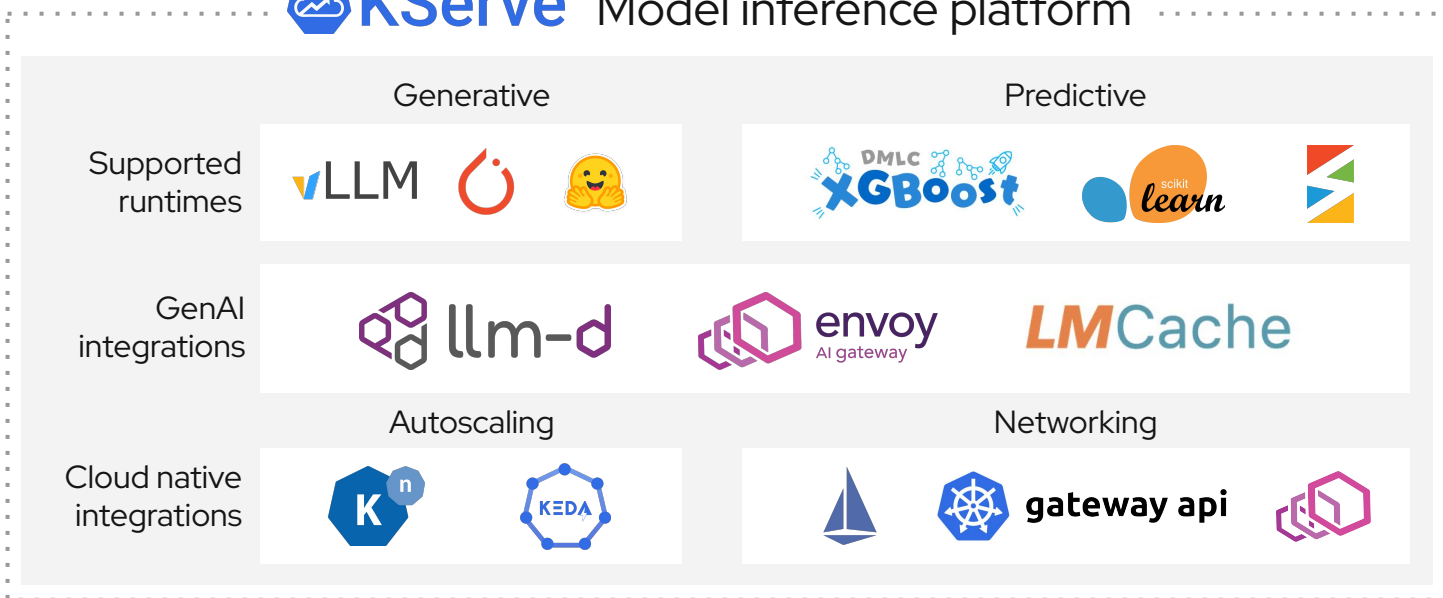
Yuan Tang

Senior Principal Software Engineer, Red Hat

Project Lead, KServe



KServe Model inference platform



Orchestration



Hardware accelerators (GPUs, CPUs, etc.)

KServe

CNCF **incubating**
project (Sept 2025)

30+ adopters

19 maintainers

300+ contributors

Trusted by industry leaders

KServe is used in production by organizations across various industries, providing reliable model inference at scale.

Bloomberg



CLOUDERA



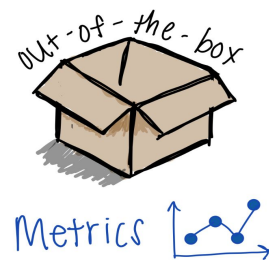
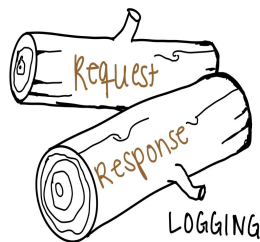
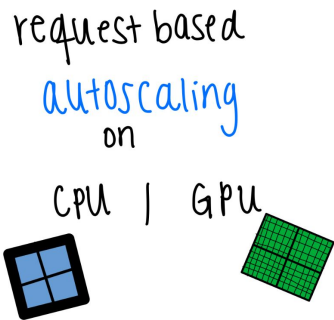
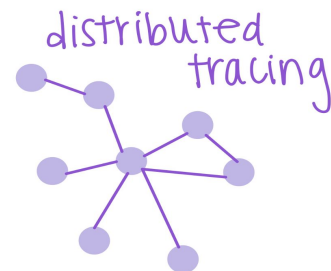
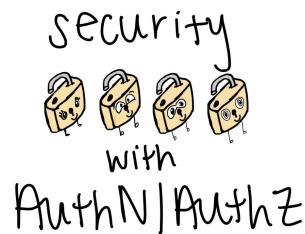
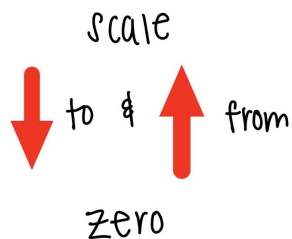
inspur



NAVER



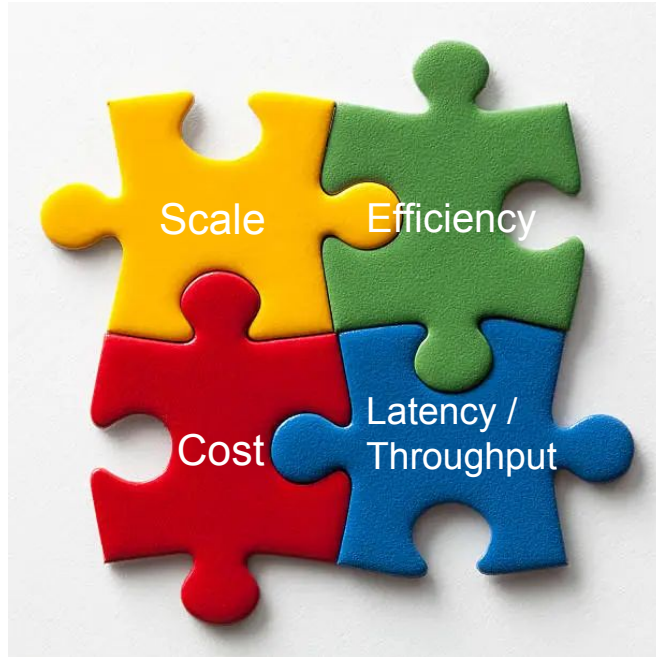
Predictive AI Features



GenAI Building Blocks

LLM Metric-Based
Autoscaling

Prompt Caching



GenAI Runtime

vLLM, TRT-LLM,
llm-d

Intelligent Routing,
Traffic Management

Optimized LLM Inference



A CNCF Sandbox project for distributed large language model inference that runs natively on Kubernetes.



Join Our Community!

- Repo: <https://github.com/kserve/kserve>
- Website: <https://kserve.github.io>
- Biweekly community meetings on Thursdays at 9 AM PST
- [#kserve](#) and [#kserve-contributors](#) channels in the CNCF Slack



<https://github.com/kserve/kserve>

Find Us This Week!

- Check out our **maintainer session** on Thursday! <https://sched.co/2EF54>
- We have a **project booth** throughout the week!
 - Kiosk Number: P-8A
 - Location: Halls 1-5 | Project Pavilion
 - Schedule:
 - Tuesday 10:15 - 14:40
 - Wednesday 10:00 - 13:30
 - Thursday 10:00 - 12:00



<https://github.com/kserve/kserve>