OSCHINA gitee gitee AI 模力方舟

万源共振 智构未来

2025 GOTC 全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

「云原生 AI」专场

本期议题：让 Kubernetes 在 AI 时代再次换发活力

张晋涛 2025 年 11 月 01 日

个人介绍

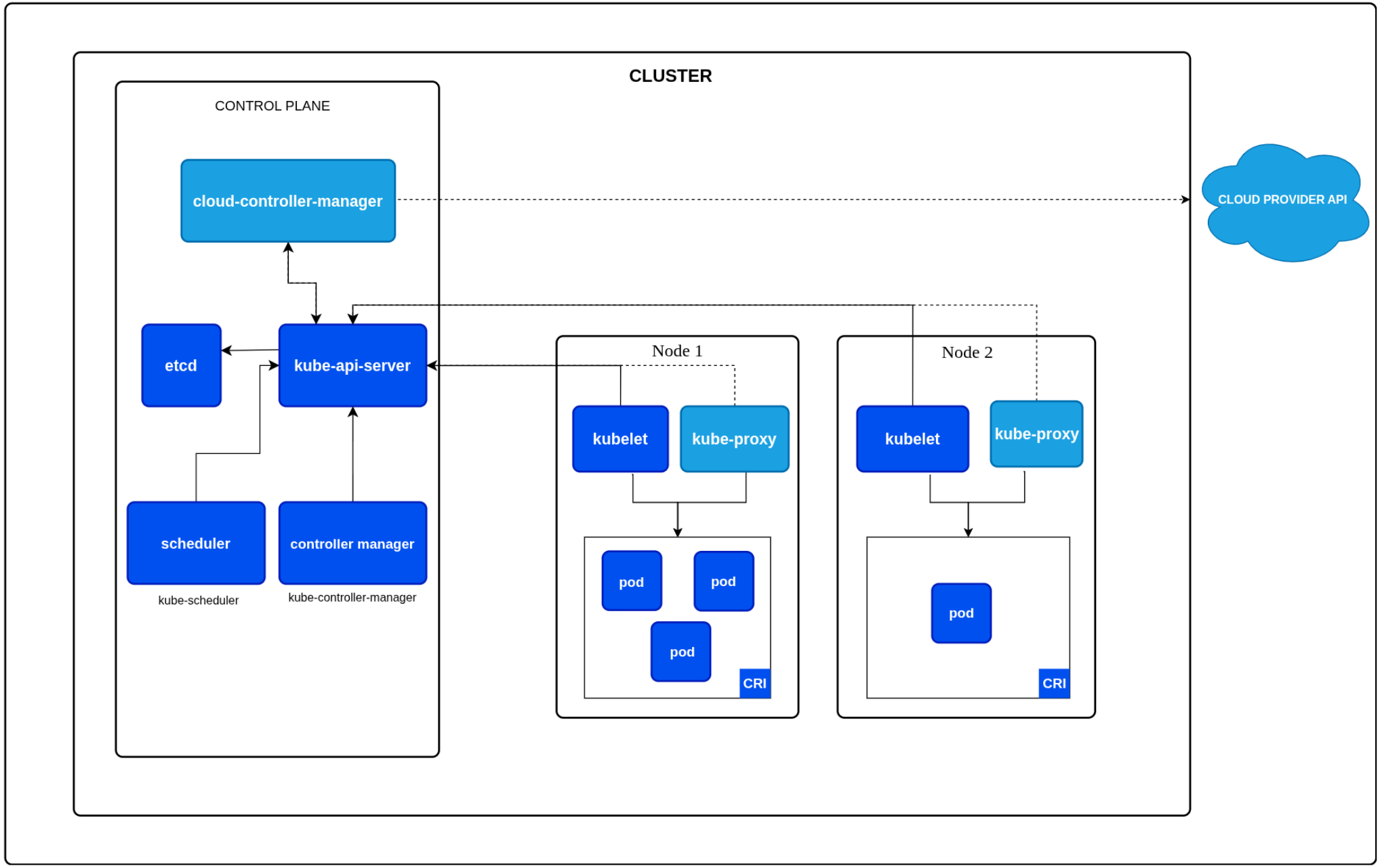
张晋涛 @ Kong Inc.

- Kubernetes Ingress-NGINX maintainer
- Microsoft MVP
- CNCF Ambassador
- LFAPAC Open Source Evangelist
- 公众号： MoeLove

Agenda

- Kubernetes 架构
- Kubernetes 在 AI 时代的挑战
- Kubernetes 在 AI 时代的变革和机遇
- Kubernetes 如何在 AI 时代迈向未来

Kubernetes 架构



Kubernetes 在 AI 时代的挑战和机遇

- 资源
- 成本

硬件资源

Device Plugin

- 非标准硬件资源（非 CPU/Memory 资源）
- 通常只能分配整块
- 不够灵活，资源浪费

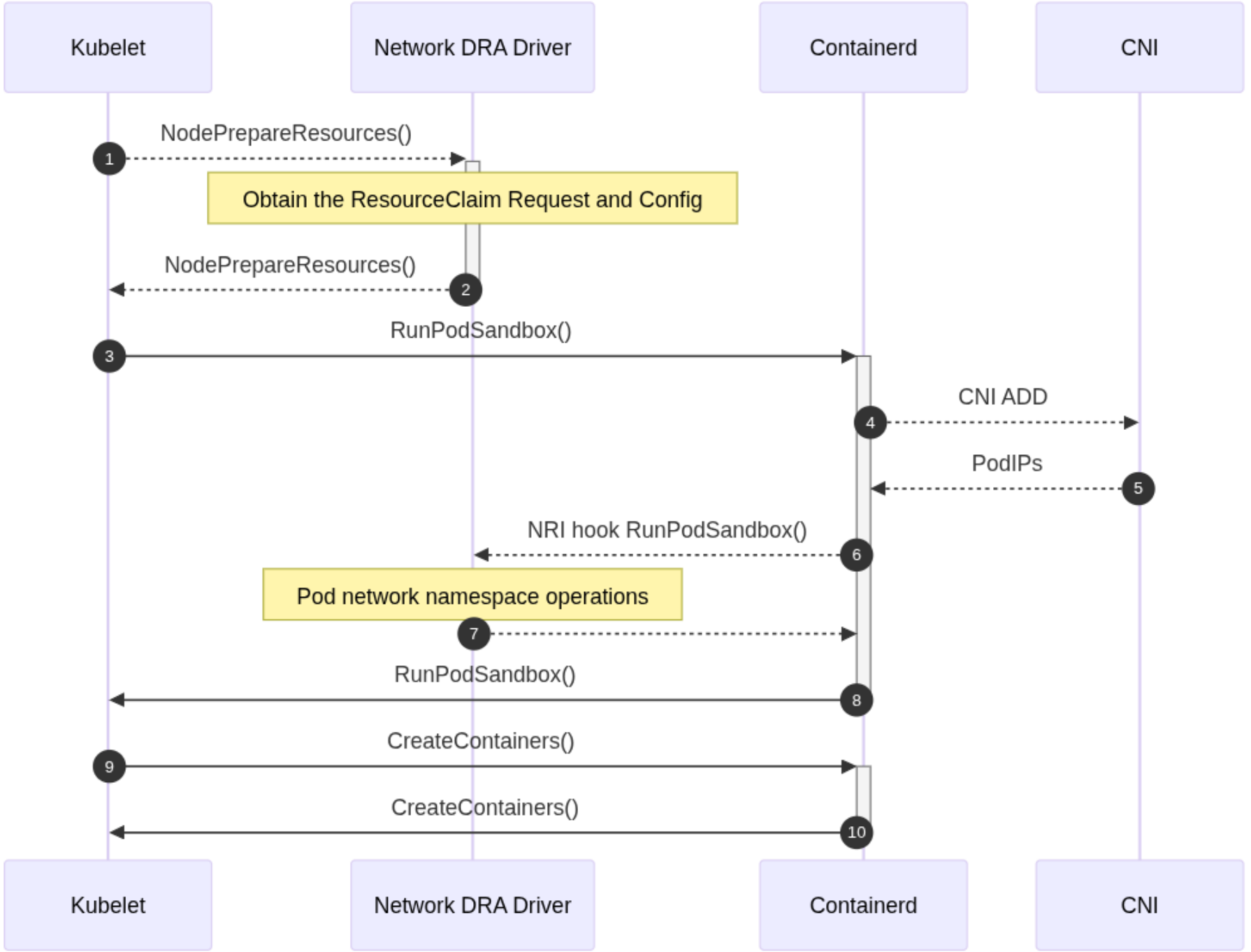
Dynamic Resource Allocation （ DRA ）

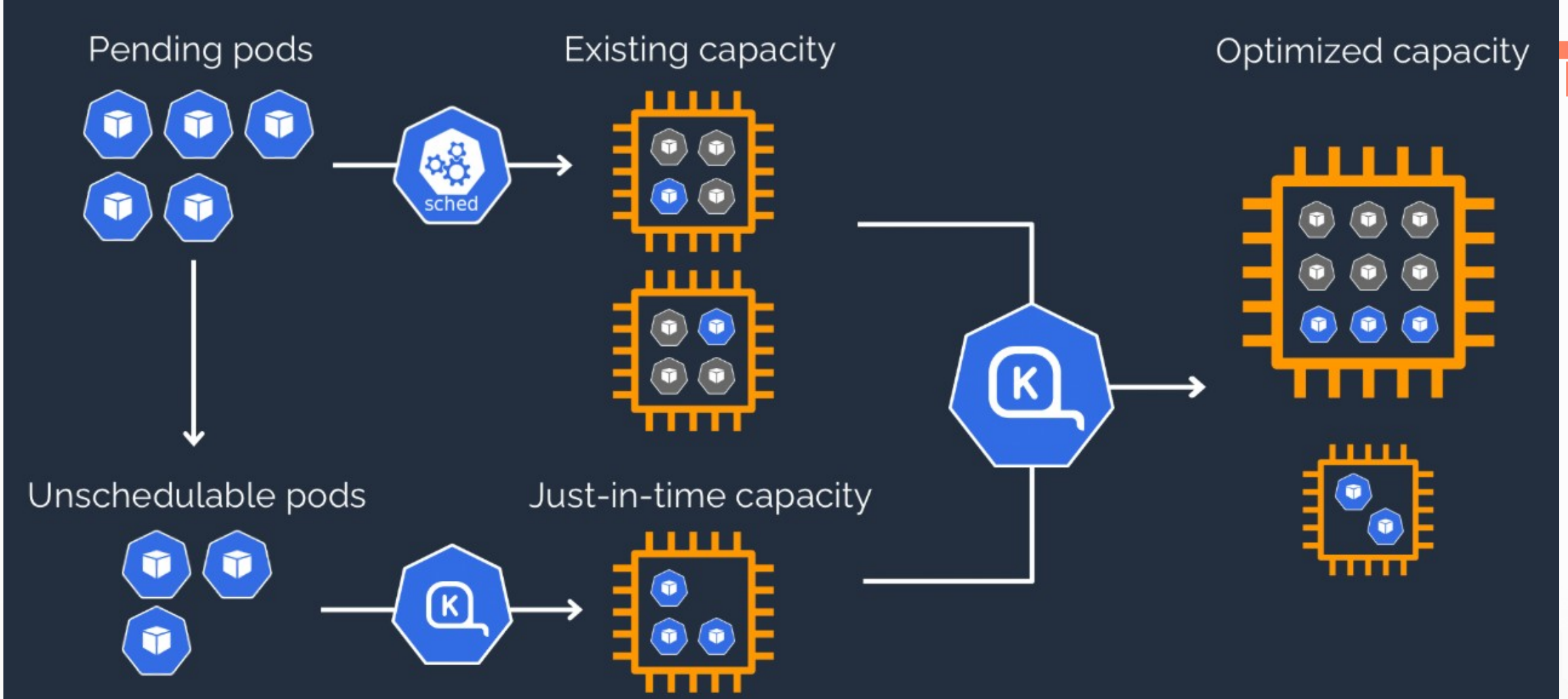
- 动态创建 / 分配
- 任意粒度
- 灵活，可扩展
- 仍在演进和扩展生态（厂商 / 插件）

DRANET

dranet.dev

- RDMA-NICs
- KND
- 高性能网络设备





Karpenter observes the aggregate resource requests of unscheduled pods and makes decisions to launch and terminate nodes to minimize scheduling latencies and infrastructure cost.

为什么选择Volcano

统一调度

支持 Kubernetes 原生负载及主流计算框架（如 TensorFlow、Spark、PyTorch、Ray、Flink等）的一体化作业调度。

队列管理

提供多层级队列管理能力，实现精细化资源配额控制和任务优先级调度。

异构设备支持

高效调度GPU、NPU等异构设备，充分释放硬件算力潜力。

网络拓扑感知

支持网络拓扑感知调度，显著降低跨节点间的应用通信开销，在AI分布式训练场景中大幅提升模型训练效率。

多集群调度

支持跨集群作业调度，提升资源池管理能力，实现大规模负载均衡。

在离线混部

实现在线与离线任务混合部署，提升集群资源利用率。

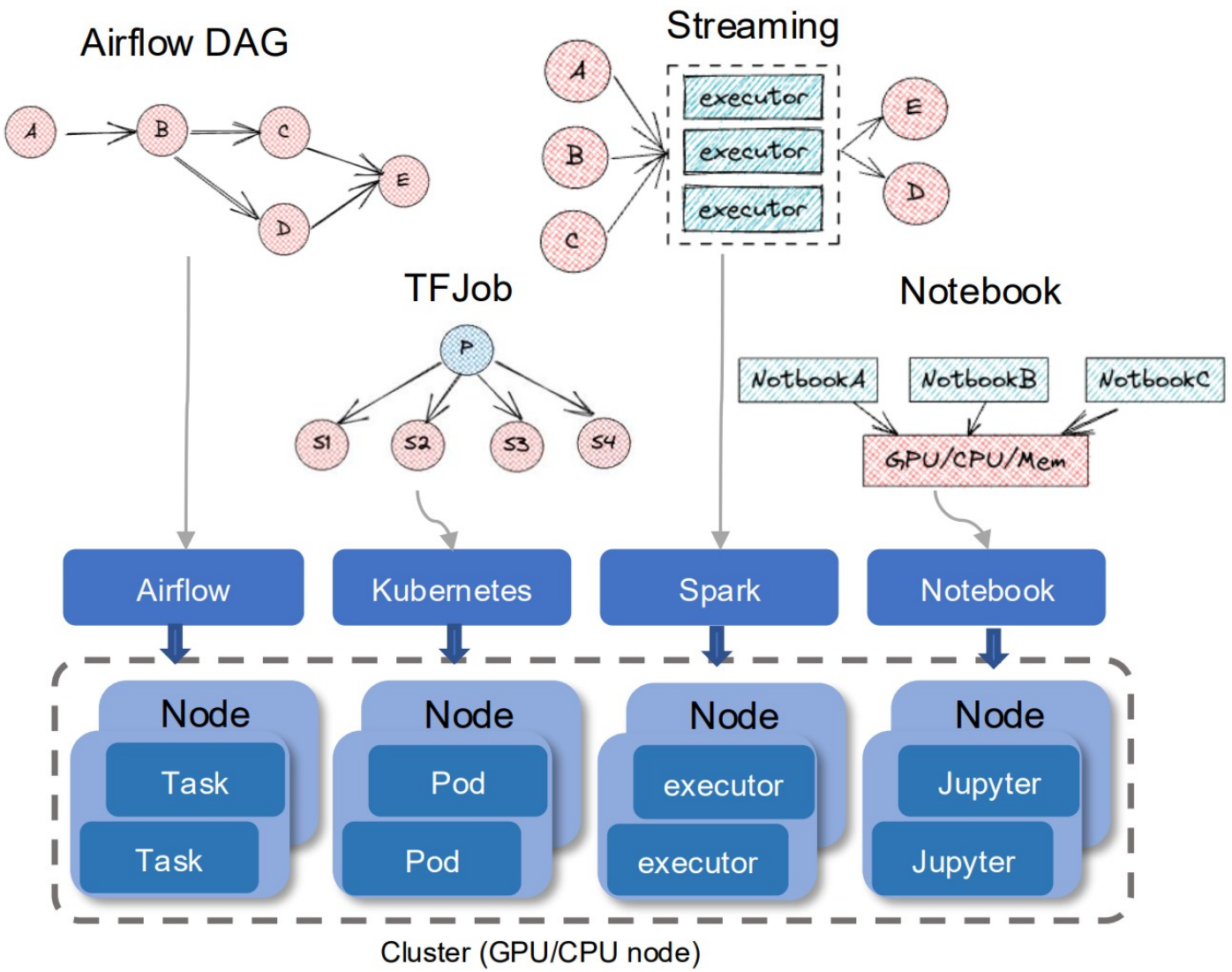
负载感知重调度

支持负载感知重调度，优化集群负载分布，提升系统稳定性。

多种调度策略

支持 Gang、Fair-Share、Binpack、DeviceShare、Capacity、Proportion、NUMA aware、Task Topology等多种调度策略，优化资源利用效率。

调度



Challenge

- **Scattered resources**

  Large-scale training cluster (thousands of CPU/GPU nodes)
- **Complex Job Management**

  Migration of legacy tasks (e.g., Airflow, Spark Streaming)
- **Diverse Workloads**

  ETL, model training (ML/DL), inference, etc.
- **Multi-Task Orchestration**

  Heterogeneous DAG scheduling & data transfer
- **PS-Worker Management Complexity**

  Contains hundreds of PS and workers

AI 企业为什么选择 JuiceFS?

千亿文件规模

JuiceFS 能在单一卷中管理高达千亿个文件，这一能力已在多家企业的生产环境中得到验证，适用于处理各类大规模 AI 数据集。

云原生设计

JuiceFS 专为云环境设计，可以在全球公有云上部署，并无缝集成到现有云基础设施中，适应不同的云平台和区域要求。

高吞吐、低延迟

在模型训练中，JuiceFS 可提供数百 GiB/s 的读吞吐量，支持每秒读取数十万个文件，并具有毫秒级的元数据响应时间。通过灵活的缓存配置，JuiceFS 能够提供无限的聚合吞吐能力，显著减少 GPU 的等待时间，从而提高整体的计算效率。

可支持混合云、多云架构

在跨多区域使用 GPU 资源时，JuiceFS 能保证数据在全球范围内的就近访问和一致性。有效减轻了跨区域访问的成本负担，并优化了数据调度与分发，适用于混合云、多云架构。

安全性

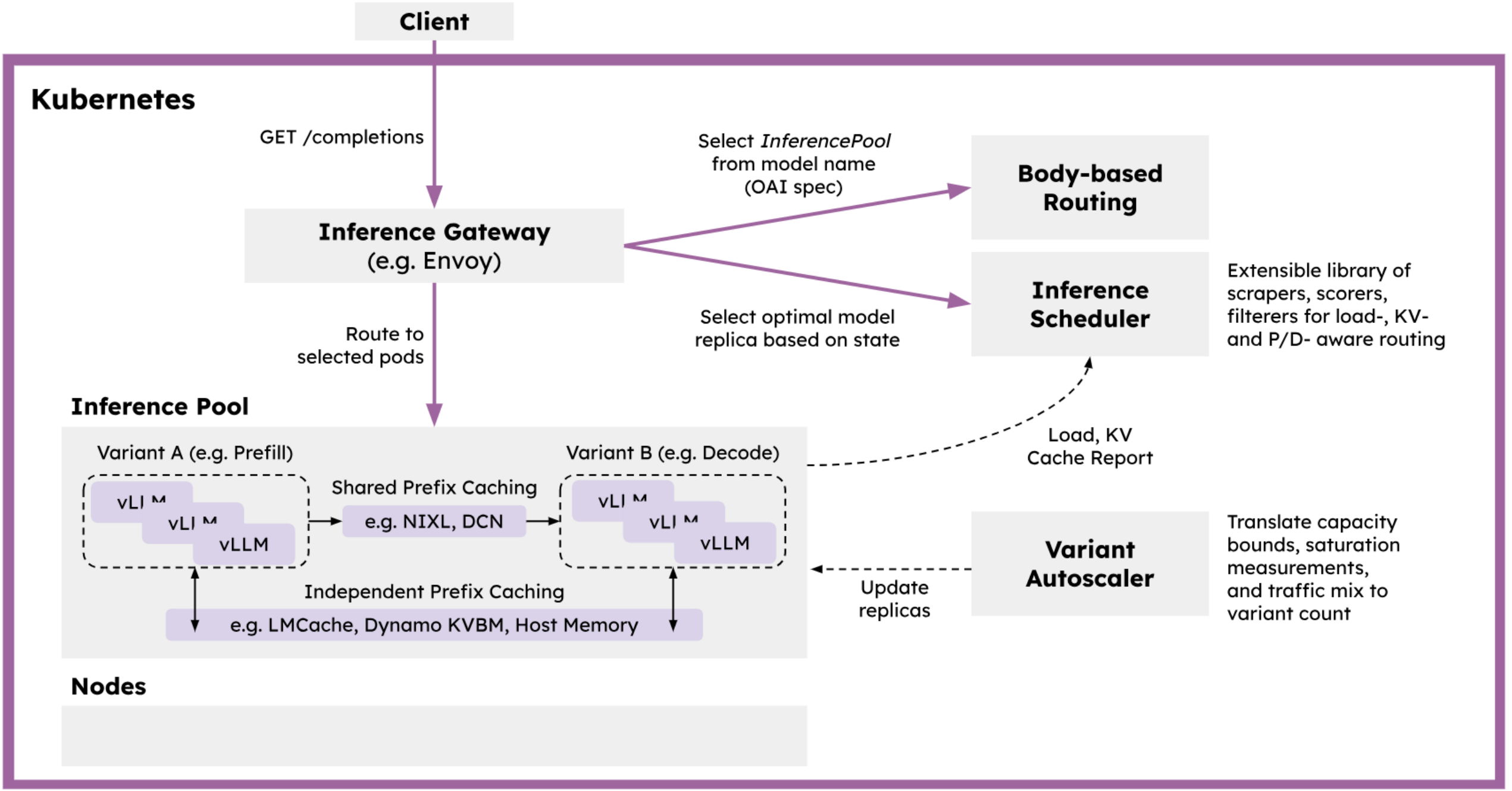
JuiceFS 可为不同团队共享存储系统提供数据隔离和安全性保障：基于访问令牌（Token）的挂载和访问控制、Linux 文件权限（File Permission）、POSIX ACL、子目录挂载、容量与 Inode 配额、流量 QoS 等能力。

成本优势

自动驾驶领域数据增长迅速，JuiceFS 通过对象存储作为底层存储，实现了容量的弹性扩展，并有效降低了存储成本。同时，JuiceFS 的灵活架构有助于降低学习、维护和迁移成本。


vLLM
Easy, fast, and cheap LLM serving for everyone


SGL
pypi v0.5.4.post1
downloads 4M
license Apache-2.0
closed issues 3.1k
open issues 546
Ask DeepWiki

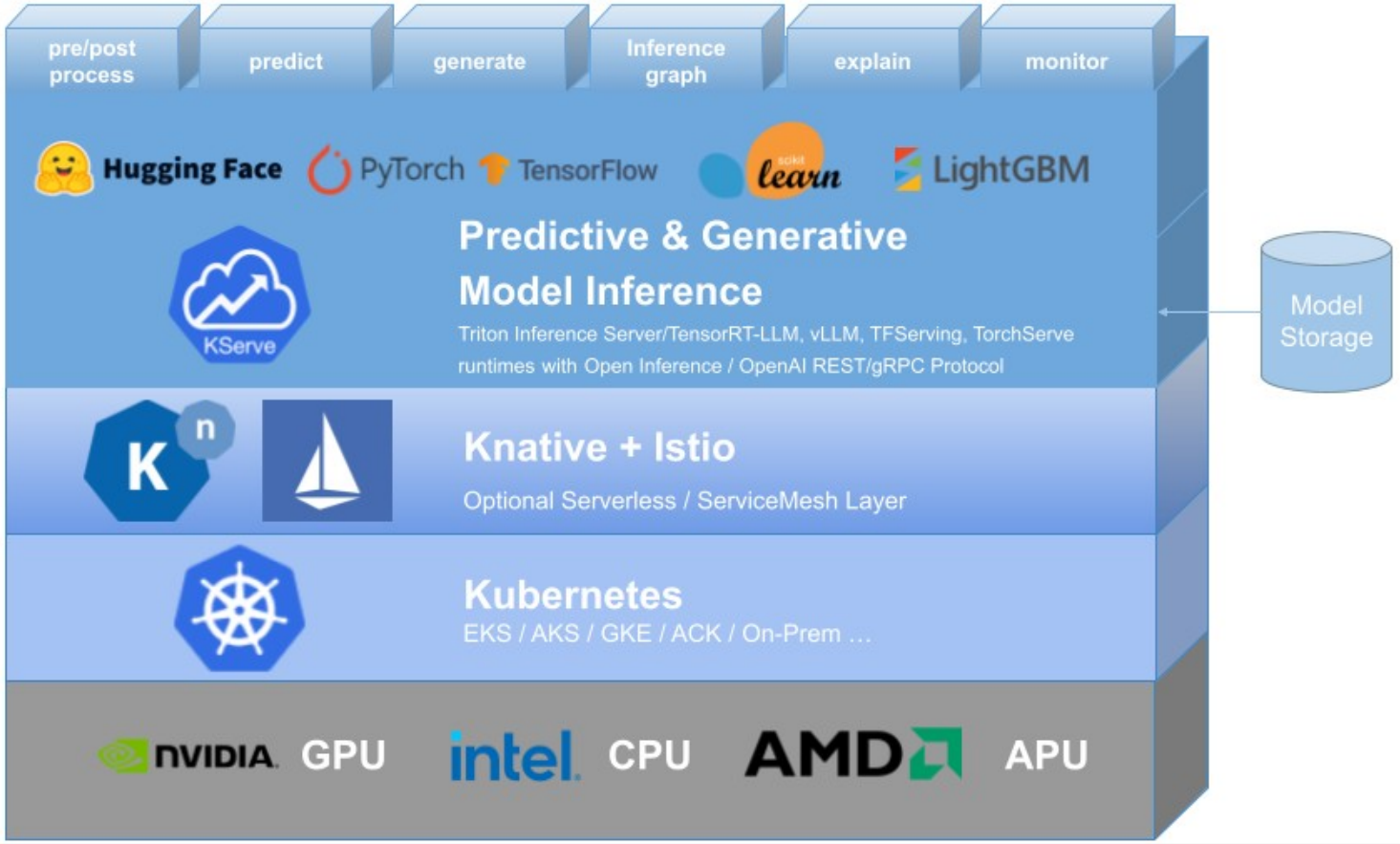

Client
Kubernetes
GET /completions
Inference Gateway
(e.g. Envoy)
Select InferencePool
from model name
(OAI spec)
Body-based
Routing
Select optimal model
replica based on state
Inference
Scheduler
Extensible library of
scrapers, scorers,
filterers for load-, KV-
and P/D- aware routing
Route to
selected pods
Inference Pool
Variant A (e.g. Prefill)
Variant B (e.g. Decode)
Shared Prefix Caching
e.g. NIXL, DCN
vLLM
vLLM
vLLM
vLLM
Load, KV
Cache Report
Independent Prefix Caching
e.g. LMCache, Dynamo KVBM, Host Memory
Update
replicas
Variant
Autoscaler
Translate capacity
bounds, saturation
measurements,
and traffic mix to
variant count
Nodes

KServe

CNCF incubating project

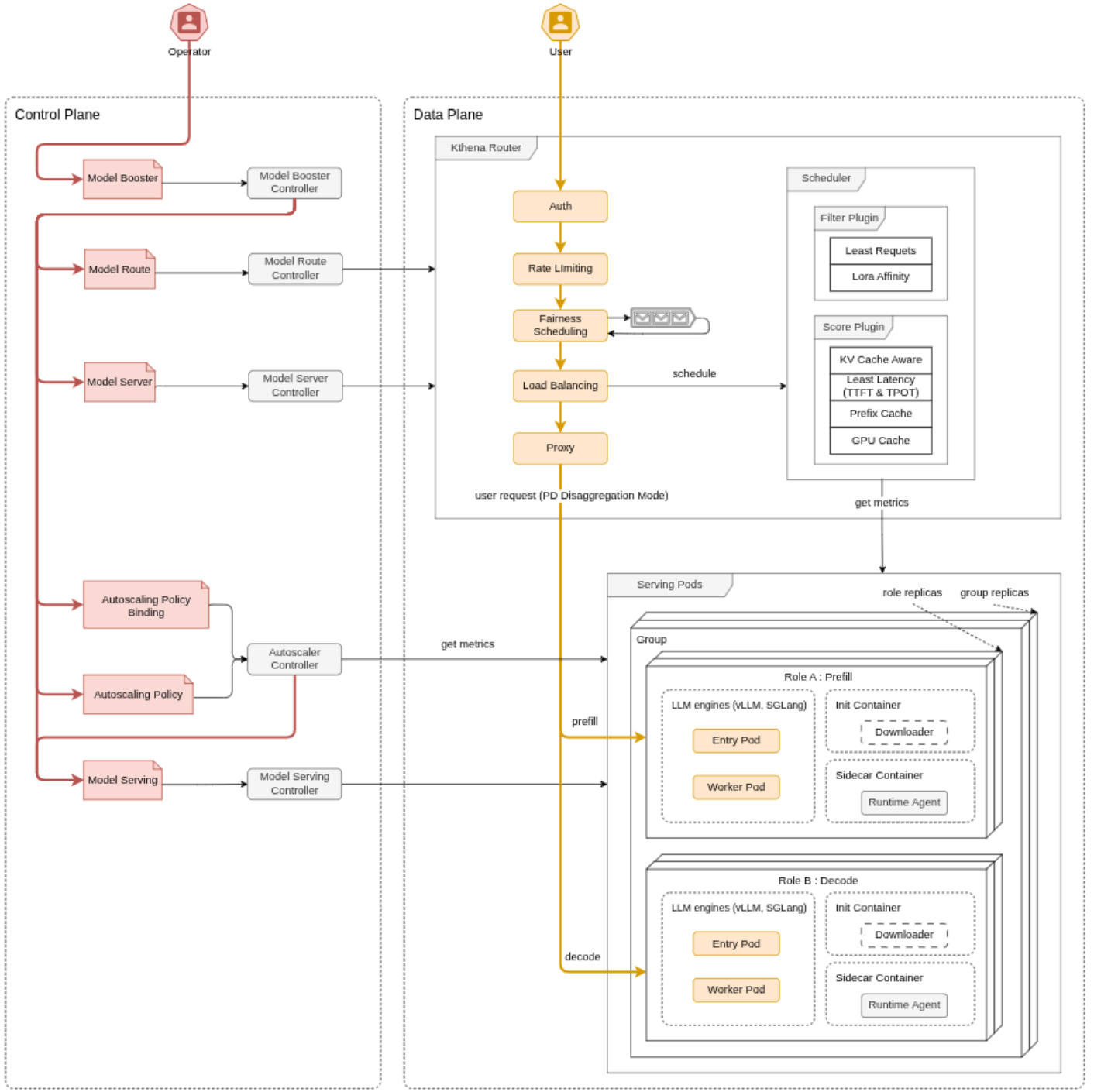
How KServe Works

KServe provides a Kubernetes custom resource definition for serving ML models on arbitrary frameworks, encapsulating complexity of autoscaling, networking, health checking, and server configuration to bring cutting edge serving features to your ML deployments.



Kthena

volcano-sh/kthena



CNCF CNAI LANDSCAPE

GOTC

CLOUD NATIVE LANDSCAPE

EXPLORE GUIDE STATS

Type / to search items

Filters GROUP: Projects and products | Members | Certified partners and providers | Serverless | Wasm | CNAI VIEW MODE: Grid | Card ZOOM: - +

CNAI

Data Architecture

ClickHouse, druid, cassandra, SCYLLA, HBase, presto, trino, Spark, Kubeflow, Flink, kafka, PULSAR, ALLUXIO, Kubeflow, FEAST, Fluid, redis, Valkey

Data Science

marimo, jupyter, Kubeflow, PyTorch, dmlc XGBoost, Apache Zeppelin

Vector Databases

Milvus

General Orchestration

kubernetes — CNCF GRADUATED

VOLCANO — CNCF INCUBATING

RAY, ARMADA, HAMi, COZYSTACK

Model/LLM Observability

trulens, deepchecks, Okahu, OpenLIT, DeepFlow

Agentic AI

kagent

ML Serving

KServe — CNCF INCUBATING

vLLM, KAITO, vLLM

CI/CD - Delivery

Kubeflow, mlflow, MLRun

Open Enterprise AI Blueprints

Open Platform for Enterprise AI

Distributed Training

Kubeflow, PyTorch, PyTorch, deepspeed, Megatron-LM, Kubeflow

Workload Observability

Grafana, W&B

Governance, Policy & Security

Kyverno — CNCF INCUBATING

spicedb, PERMIFY, JSON, stacklok

AutoML

HYPEROPT, Katib

THANKS