

vLLM processes push kv-cache ZMQ events (BlockStored, Block Removed, AllBlocksCleared)
cache-aware routing

LLMInferenceServiceConfig

common configuration patterns for intelligent routing, KV cache-aware scheduling, and disaggregated serving

EPP scrapes /metrics (e.g., queued, executing, kv-cache fullness) from leaders
load-aware routing

spec:
replicas: 2
size: 3

data: 6 → 6 ranks
dataLocal: 2 → 2 rank in each Pod
⇒
6/2 = 3 Pods, i.e. size of LWS

spec:
replicas: 2

parallelism:
data: 6
dataLocal: 2
tensor: 3

worker: {}

scaling*:
minReplicas: 2
maxReplicas: 5
wva:
keda:
pollingInterval: 10

tensor:
parallelism
for each
rank over 3
GPUs

4. leader coordinates execution
across local/distributed workers:

- **simple case:** one worker/rank handles it
- **distributed case:** multiple workers/ranks participate (e.g. **expert parallelism**)

© ai-infra.dev

Serving and Parallelism

Prometheus

Routing

Autoscaling

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller

ScaledObject

KEDA Metric API Server

HPA Controller

HorizontalPodAuto scaler

Prometheus

Prometheus Controller

ServiceMonitor

KEDA Controller