



KubeCon



CloudNativeCon

THE LINUX FOUNDATION



China 2024





KubeCon



CloudNativeCon



China 2024

# Empower Large Language Models (LLMs) in Production With Cloud Native AI Technologies

Lize Cai, Senior Software Engineer, SAP  
Yang Che, Senior Engineer, Alibaba Cloud

# About us



**Lize Cai**  
**Senior Software Engineer in SAP**



**Yang Che**  
**Senior Engineer in Alibaba Cloud**

# Agenda

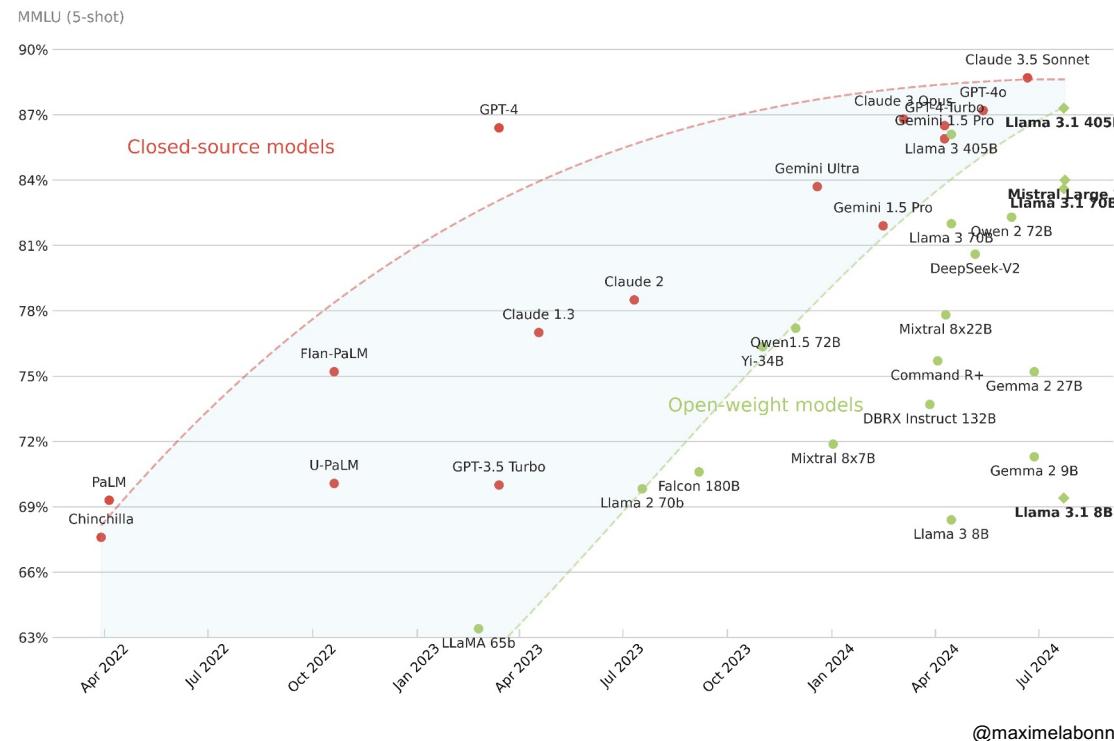


- Introduction
- LLM Challenges in Production
- Manages LLM lifecycle in the K8s way - KServe
- Accelerates LLM scaling from data perspective - Fluid
- Demo
- Future Works
- QA

# Introduction

## Closed-source vs. open-weight models

Llama 3.1 405B closes the gap with closed-source models for the first time in history.



Models 132,425 Filter by name Full-text search Edit filters Sort: Trending

Active filters: text-generation Clear all

- microsoft/Phi-3.5-MoE-instruct Text Generation Updated about 16 hours ago ↓ 1.4k ❤ 221
- microsoft/Phi-3.5-mini-instruct Text Generation Updated about 20 hours ago ↓ 4.1k ❤ 189
- meta-llama/Meta-Llama-3.1-8B-Instruct Text Generation Updated about 23 hours ago ↓ 1.68M ❤ 1.95k
- microsoft/Phi-3.5-vision-instruct Text Generation Updated about 18 hours ago ↓ 1.07k ❤ 167
- NousResearch/Hermes-3-Llama-3.1-8B Text Generation Updated 2 days ago ↓ 13.3k ❤ 117
- meta-llama/Meta-Llama-3.1-8B Text Generation Updated 28 days ago ↓ 452k ❤ 603
- tiiuae/falcon-mamba-7b Text Generation Updated 3 days ago ↓ 5.05k ❤ 156
- akjindal53244/Llama-3.1-Storm-8B Text Generation Updated about 15 hours ago ↓ 1.14k ❤ 60

# Introduction



It is a common use case to provide a playground to try out different models ...

Inference API ⓘ

Text Generation Examples

Input a message to start chatting with microsoft/Phi-3.5-mini-instruct.

Your sentence here... Send

View Code

Maximize

This screenshot shows the Azure AI Studio Inference API playground. It features a text input field with placeholder text "Input a message to start chatting with microsoft/Phi-3.5-mini-instruct." and a "Send" button. Below the input field is a text area labeled "Your sentence here..." with a "Send" button. At the bottom left is a "View Code" link. On the right side, there's a "Maximize" button.

Azure AI Studio / Model catalog Sign in ⓘ

Model catalog

Find the right model to build your custom AI solution

Announcements

Phi-3.5 models are here! Microsoft's latest Phi-3.5 MoE and Mini models now support 20+ languages. View models Read blog

Try improved GPT-4o Latest version of GPT-4o, the most advanced multimodal model from OpenAI, is now available. Try early access Read blog

Cohere Rerank is here! Cohere Rerank, the leading AI model for reranking, is now available on Azure. View models Read blog

Mistral Large (2407) & Nemo are here! Large (2407) and Nemo, Mistral AI's latest models, now available in Azure AI Studio. View models Read blog

All filters Collections Deployment options Inference tasks Fine-tuning tasks Licenses

Search Models 1737

gpt-4o-mini	gpt-4o	gpt-4	whisper	tts-hd
Chat completion	Chat completion	Chat completion	Speech recognition	Text to speech
tts	text-embedding-3-small	text-embedding-3-large	dall-e-2	dall-e-3
Text to speech	Embeddings	Embeddings	Text to image	Text to image
gpt-35-turbo-instruct	davinci-002	text-embedding-ad-002	gpt-4-32k	gpt-35-turbo-16k
Chat completion	Completions	Embeddings	Chat completion	Chat completion
Phi-3.5-turbo	babbage-002	Phi-3.5-vision-instruct	Phi-3.5-mini-instruct	Phi-3.5-MoE-instruct
Chat completion	Completions	Chat completion	Chat completion	Chat completion
Phi-3-small-8k-instruct	Phi-3-small-128k-instruct	Phi-3-mini-4k-instruct	Phi-3-mini-128k-instruct	Phi-3-medium-4k-instruct
Chat completion	Chat completion	Chat completion	Chat completion	Chat completion
Phi-3-medium-128k-instruct	Phi-3-vision-128k-instruct	Llama-2-7b	Llama-2-70b	Llama-2-13b
Chat completion	Chat completion	Text generation	Text generation	Text generation

Privacy & cookies Terms & conditions Trademarks

This screenshot shows the Azure AI Studio Model catalog. It features a grid of cards for various AI models. Each card includes the model name, a small icon, a brief description, and links to "View models" and "Read blog". The models listed include gpt-4o-mini, gpt-4o, gpt-4, whisper, tts-hd, tts, text-embedding-3-small, text-embedding-3-large, dall-e-2, dall-e-3, davinci-002, text-embedding-ad-002, gpt-4-32k, gpt-35-turbo-16k, Phi-3.5-turbo, babbage-002, Phi-3.5-vision-instruct, Phi-3.5-mini-instruct, Phi-3.5-MoE-instruct, Phi-3-small-8k-instruct, Phi-3-small-128k-instruct, Phi-3-mini-4k-instruct, Phi-3-mini-128k-instruct, Phi-3-medium-4k-instruct, Llama-2-7b, Llama-2-70b, and Llama-2-13b. The catalog also includes sections for announcements and deployment options like Inference tasks and Fine-tuning tasks.

# Introduction



But it is not so easy...

The image displays a comparison between two screenshots of an AI inference API interface. Both screenshots show a purple progress bar at the bottom with the text "Model is loading". A large blue arrow points from the left screenshot to the right one, indicating a progression or flow.

**Left Screenshot (Initial State):**

- Downloads last month: 4,098
- Safetensors logo
- Model size: 3.82B params
- Tensor type: BF16
- Inference API logo
- Text Generation tab selected
- Input field: "Input a message to start chatting with microsoft/Phi-3.5-mini-instruct."
- Message input: "Can you provide ways to eat combinations of bananas and dragonfruits?"
- Text area placeholder: "Your sentence here..."
- Progress bar: "Model is loading"
- Buttons: </> View Code, Maximize

**Right Screenshot (Final State):**

- Downloads last month: 4,098
- Safetensors logo
- Model size: 3.82B params
- Tensor type: BF16
- Inference API logo
- Text Generation tab selected
- Input field: "Input a message to start chatting with microsoft/Phi-3.5-mini-instruct."
- Text area placeholder: "Your sentence here..."
- Progress bar: "Model microsoft/Phi-3.5-mini-instruct time out"
- Buttons: </> View Code, Maximize

# LLM Challenges in Production



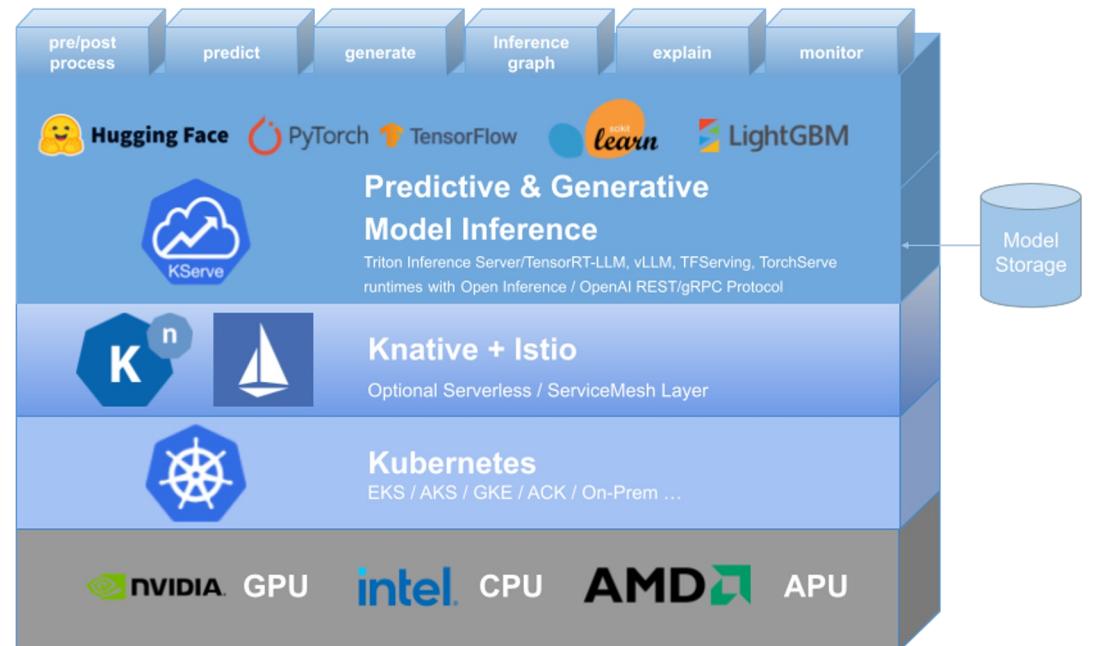
- **New requirements on serving LLM**
  - New inference APIs like text generation, embeddings.
  - Streaming response is required for real-time user experience.
- **Variety of models and runtimes**
  - TGI, vLLM, TRT-LLM etc.
  - Llama, Mistral, Phi, Qwen etc.
- **LLM services from cloud providers**
  - Different providers have their own spec (api and token calculation) which leading to a poor user experience and increased maintenance efforts.
- **High computing cost**
  - the need for expensive hardware, high energy consumption, and associated infrastructure expenses.
- **Data privacy**
  - Model and request data can be sensitive and private for inference.



Manages LLM lifecycle in the K8s  
way - KServe

# What is KServe?

**Highly scalable** and **standards-based** **cloud-native model inference platform** on **Kubernetes** for trusted AI that encapsulates the complexity of deploying AI models to production.



# What is KServe?



KubeCon



CloudNativeCon



THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI dev  
OPEN SOURCE DATA & ML SUMMIT

China 2024

## Core Inference

- Transformer/Predictor
- Serving Runtimes
- Custom Runtime SDK
- Open Inference Protocol
- Serverless Autoscaling
- Cloud/PVC Storage

## Advanced Inference

- ModelMesh for Multi-Model Serving
- Inference Graph
- Payload Logging
- Request Batching
- Canary Rollout

## Model Explanability & Monitoring

- Text, Image, Tabular Explainer
- Bias Detector
- Adversarial Detector
- Outlier Detector
- Drift Detector

# What is KServe?



# Serving runtime support matrix

# KServe on LLM



## Inference Service and Serving Runtime for LLM

### KServe User

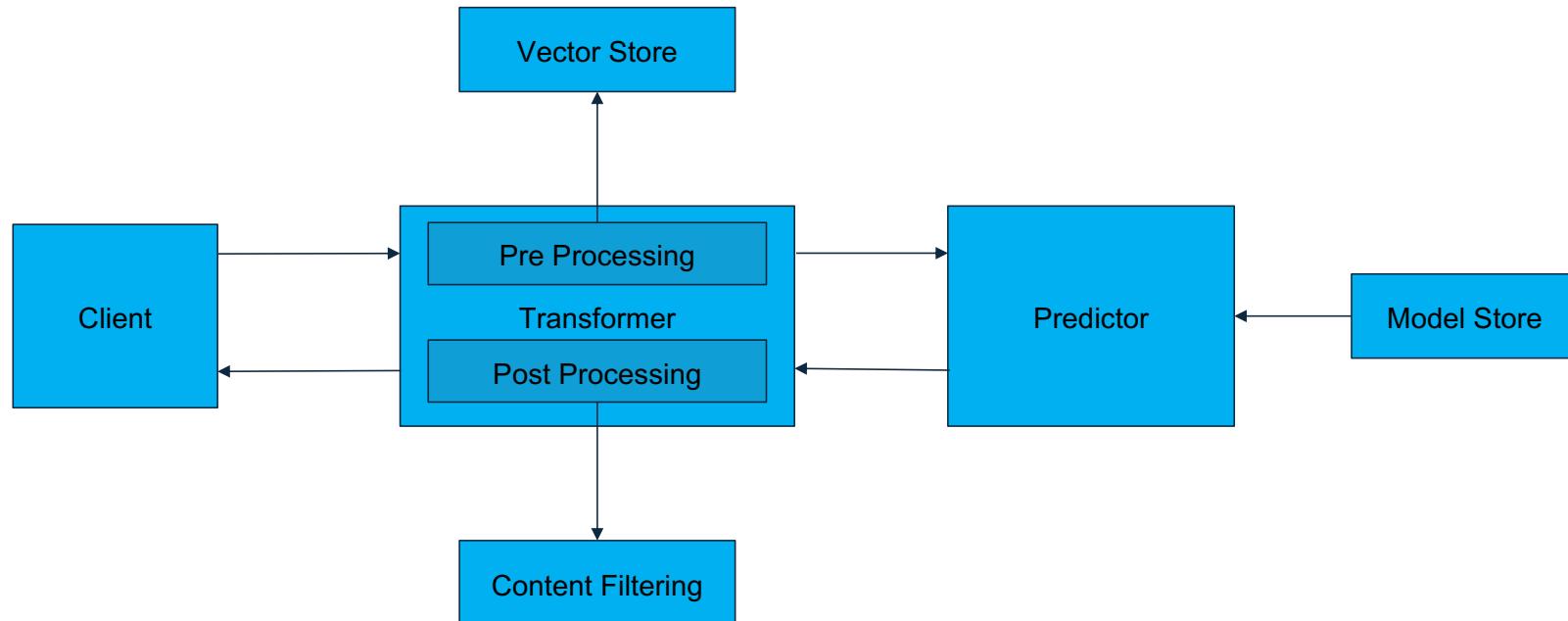
```
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: huggingface-llama3
spec:
  predictor:
    model:
      modelFormat:
        name: huggingface
      args:
        - --model_id=meta-llama/meta-llama-3-8b-instruct
    resources:
      limits:
        cpu: "6"
        memory: 24Gi
        nvidia.com/gpu: "1"
      requests:
        cpu: "6"
        memory: 24Gi
        nvidia.com/gpu: "1"
```

### KServe Admin

```
apiVersion: serving.kserve.io/v1alpha1
kind: ClusterServingRuntime
metadata:
  name: kserve-huggingfaceserver
spec:
  annotations:
    prometheus.kserve.io/port: '8080'
    prometheus.kserve.io/path: "/metrics"
  supportedModelFormats:
    - name: huggingface
      version: "1"
      autoSelect: true
      priority: 1
  protocolVersions:
    - v1
    - v2
  containers:
    - name: kserve-container
      image: "kserve/huggingfaceserver:latest"
      args:
        - --model_name={{ .Name }}
      resources:
        requests:
          cpu: "1"
          memory: 2Gi
        limits:
          cpu: "1"
          memory: 2Gi
```

# KServe on LLM

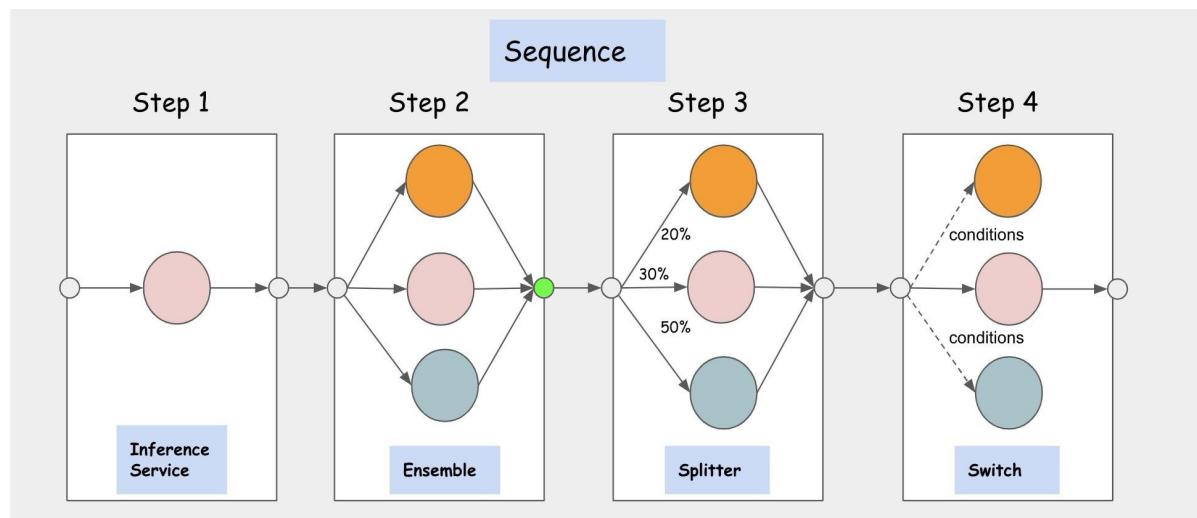
## Support of common LLM Use cases



# KServe on LLM

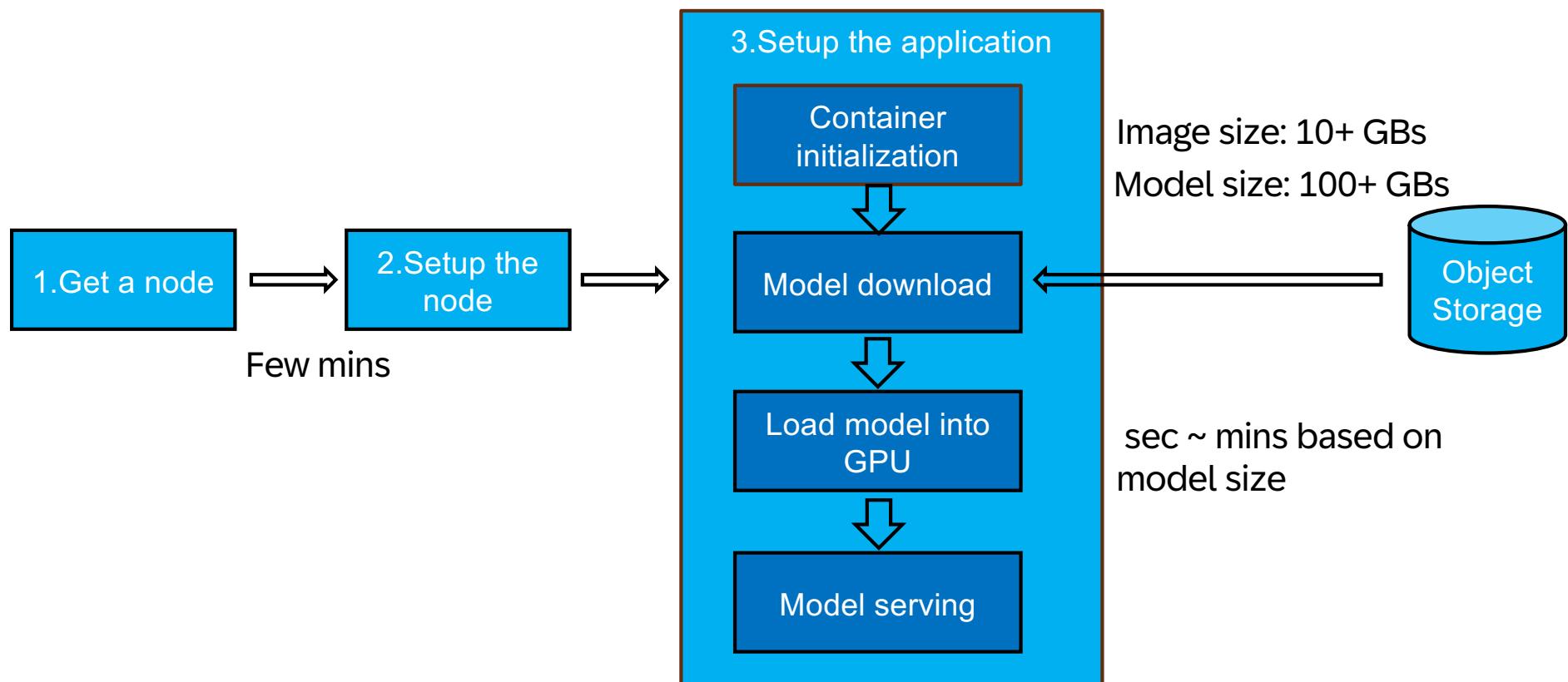
## Inference Graph:

- Inference Graph is built for more complex **multi-stage inference pipelines**.
- Inference Graph is deployed in a **declarative way and highly scalable**.
- Inference Graph supports **Sequence, Switch, Ensemble and Splitter** nodes.
- Inference Graph is **highly composable**. It is made up with a list of routing nodes and each node consists of a set of routing steps which can be either route to an InferecenService or another node.



```
apiVersion: "serving.kserve.io/v1alpha1"
kind: "InferenceGraph"
metadata:
  name: "dog-breed-pipeline"
spec:
  nodes:
    root:
      routerType: Sequence
      steps:
        - serviceName: cat-dog-classifier
          name: cat_dog_classifier # step name
        - serviceName: dog-breed-classifier
          name: dog_breed_classifier
          data: $request
          condition: "[@this].#(predictions.0=='dog')"
```

# Process of Deploying LLM

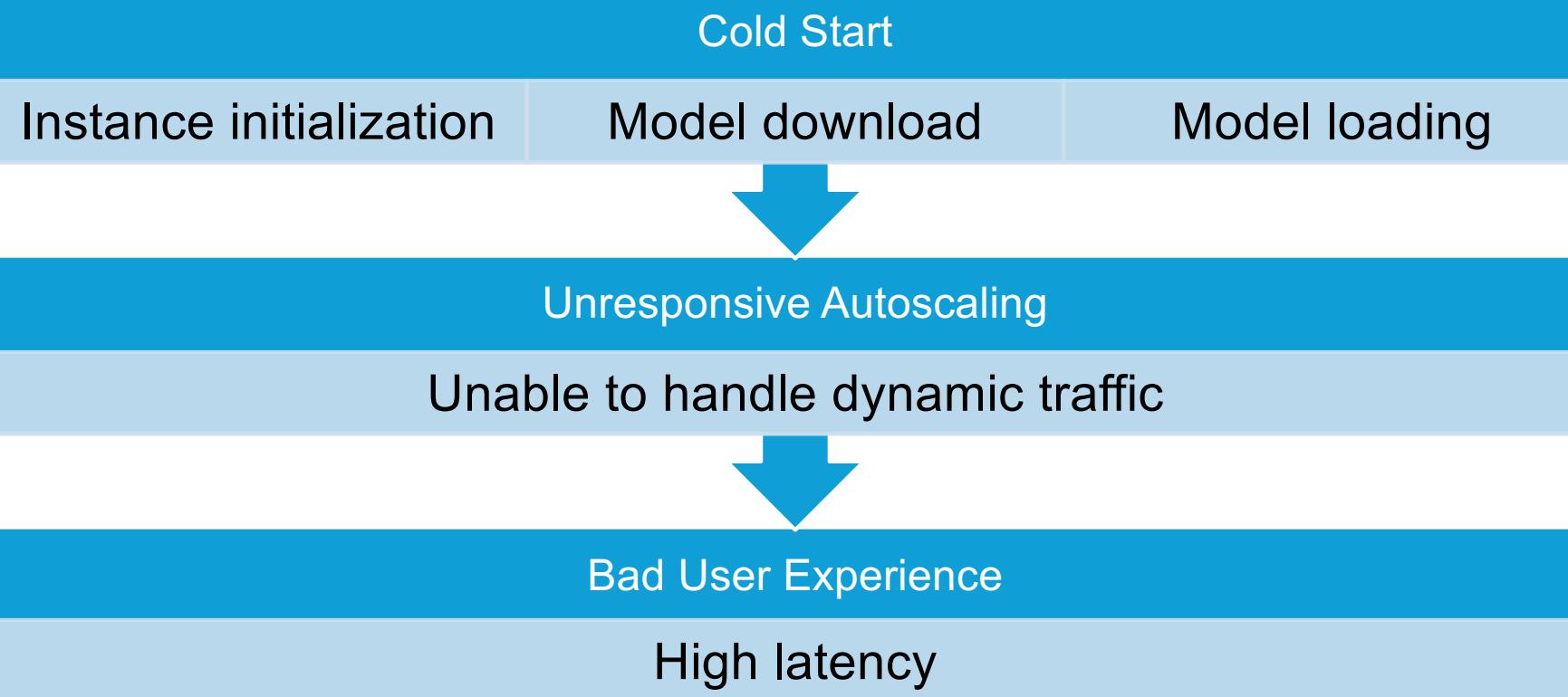


# Challenges of Autoscaling in LLM



AI.dev  
Open Source Dev & ML Summit

China 2024





Accelerates LLM scaling from data perspective - Fluid

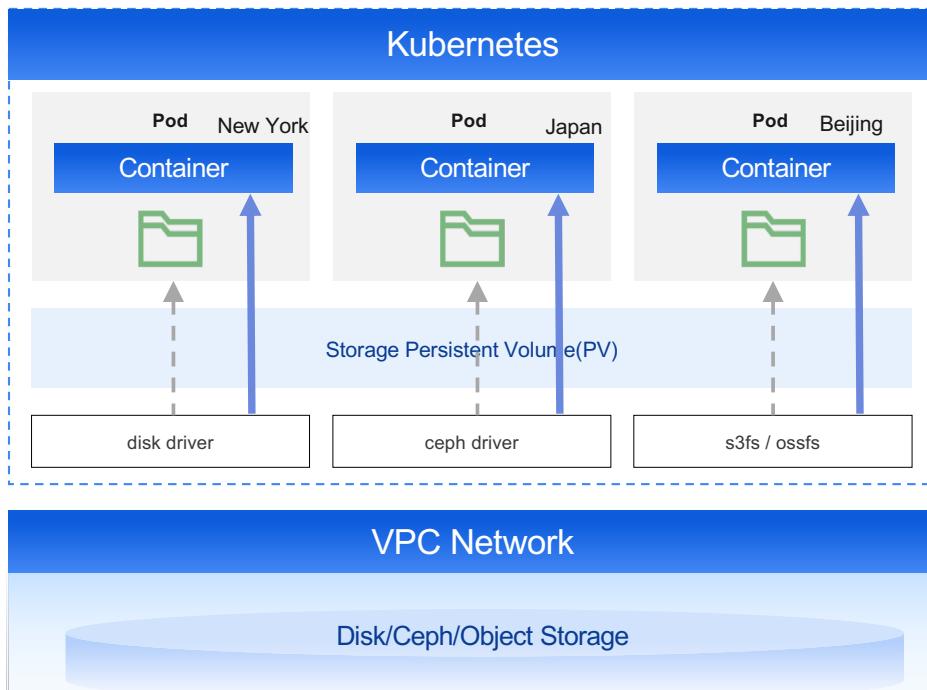
# Pains of LLM inference services in Kubernetes



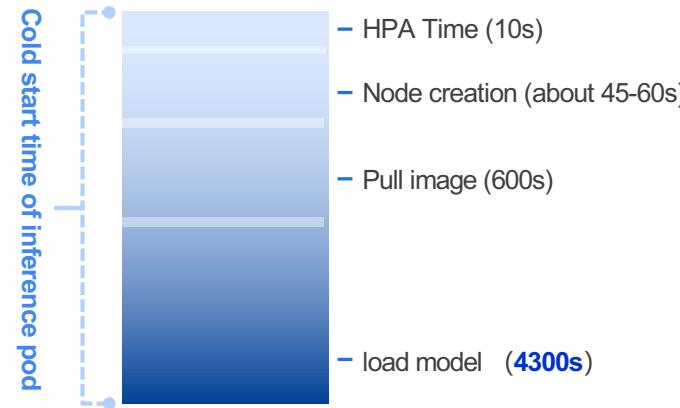
AI\_dev  
OPEN SOURCE SUMMIT

China 2024

Compute-storage separation causes high data access latency and limited bandwidth in large AI models, exacerbating cost, performance, and efficiency issues.



A Pod starts the Bloom-175B model (FP16, approximately 340 GiB), it takes about 4970 seconds, with model loading taking 4300 seconds (85%).



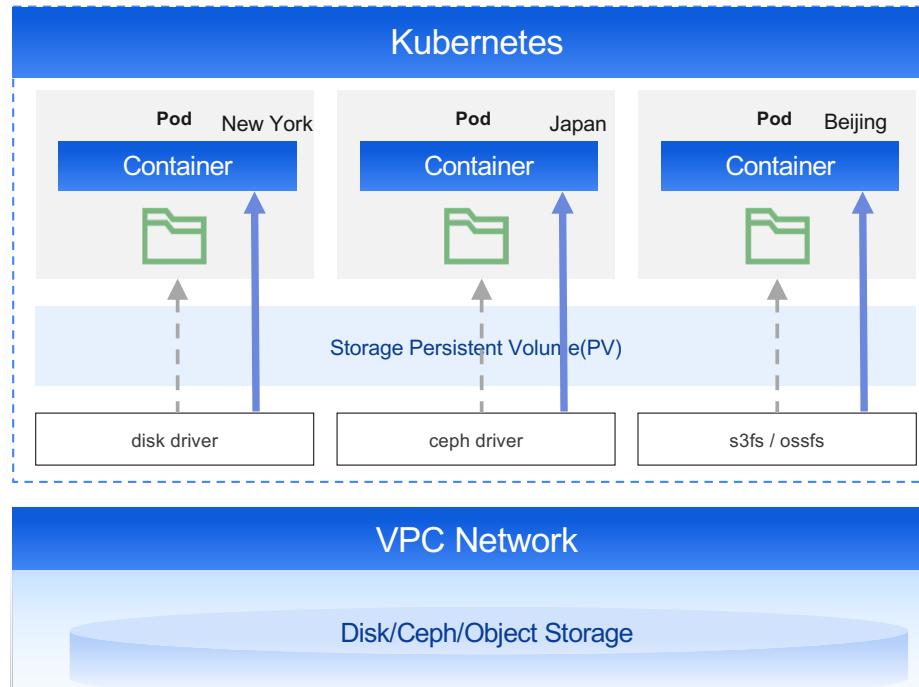
# Pains of LLM inference services in Kubernetes



AI dev  
OPEN SOURCE SUMMIT

China 2024

Cache can't always speed up data access due to engineering challenges in model inference.



## Cache Scheduling:

- How to configure affinity scheduling?

## User Experience:

- How to use data in cache?
- How to integrate with applications?

## Configuration:

- How to distribute cached data?
- What media to use for storage?

## Resource Cost :

- Extra compute and storage needs for cache. How to cut costs?
- Cross-region traffic costs

## Operations and Management:

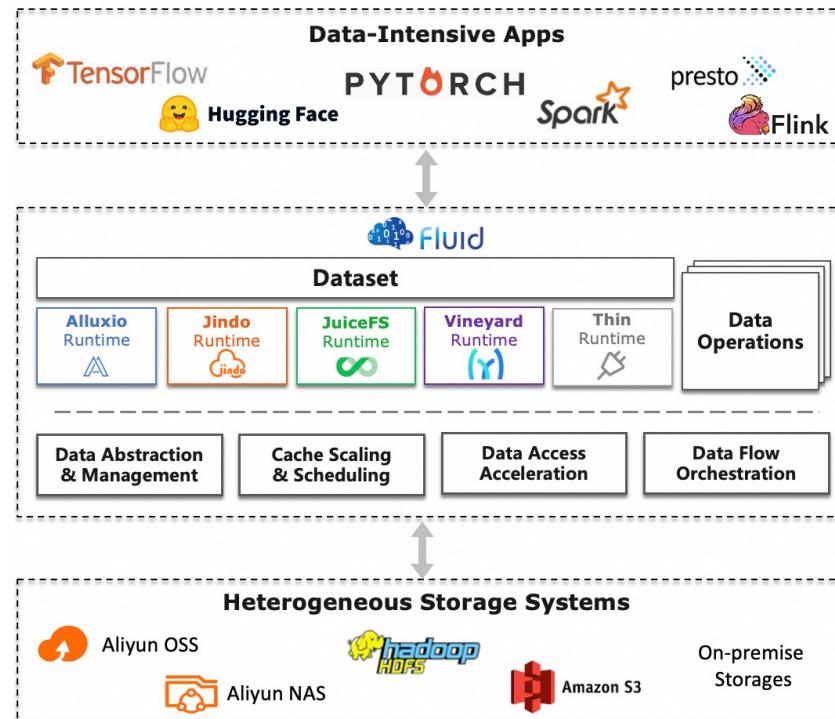
- How to manage cache system lifecycle?
- Operations observability?
- Cross-region model synchronization

## Performance Tuning:

How to optimize cache parameters for specific scenarios?



# Fluid: Data and Task Orchestrator in Kubernetes



- **Standardized:** K8s Native APIs for **data access** and **distributed cache management**.
- **Extensible:** Runtime plugins for different distributed cache and storage backends.
- **Elasticity:** Scale out and in the distributed cache on demand.
- **Performance:** Accelerate data access via elastic distributed cache
- **Automation:** Operation for Data like. *prefetching processing, migration and cache scaling*
- **Orchestration:** Data and task co-aware scheduling

Joint launched by Nanjing University, Alibaba Cloud and Alluxio

<https://github.com/fluid-cloudnative/fluid>

# Fluid Optimization for LLMs



## Characteristics and Current Issues in LLMs:

Distributed caching are *complex* and *vary greatly* across environments

How to balancing Performance and Cost

Cross-region/zone data access affects performance

Data Operations are complex and time-consuming.

## Capabilities Provided by Fluid:

Out-of-the-box acceleration capabilities.

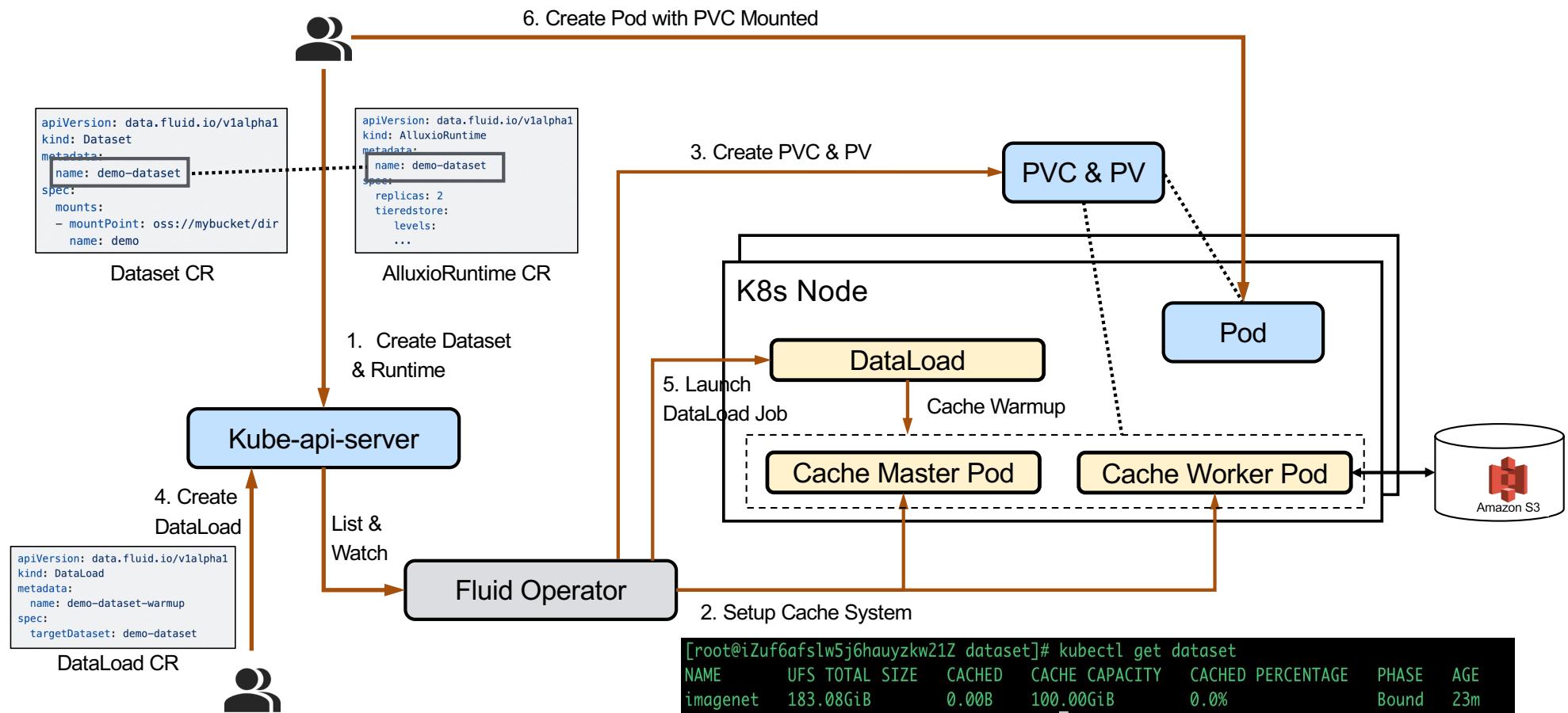
Elastic Compute-Side Distributed Cache

Affinity Scheduling for Data and Workloads

Data Flow for Automated Data Management and Consumption Processes



# Out-of-the-Box Distributed Cache



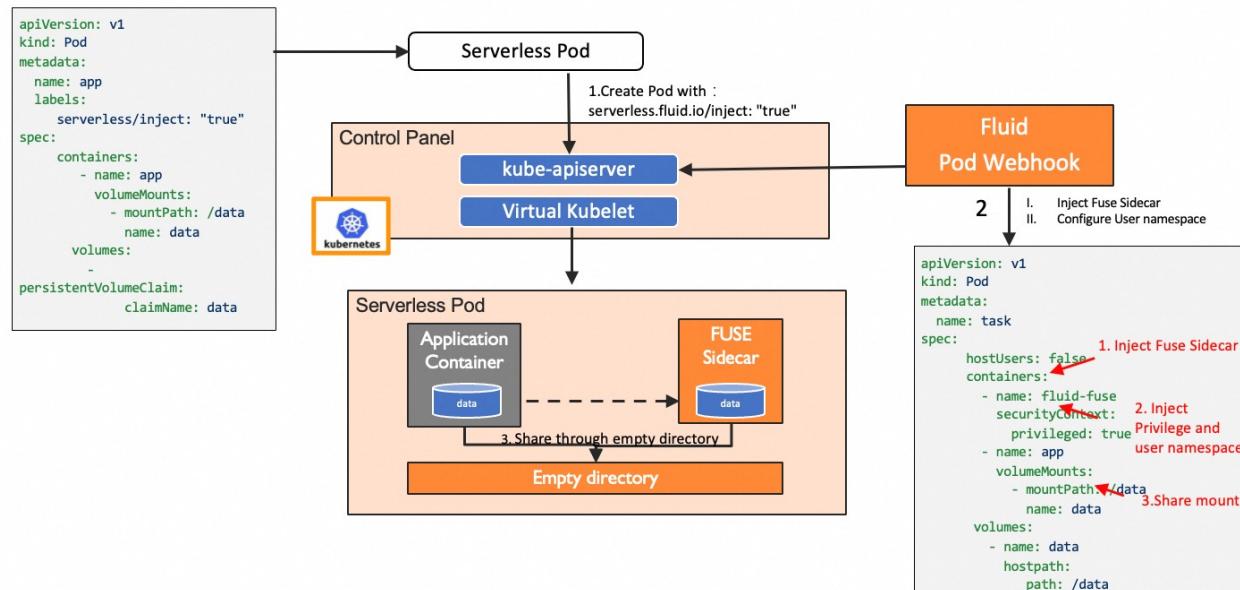
# Distributed Cache can Run Anywhere

## Problem:

LLMs often run in serverless environments cannot use third-party distributed caches due to *extensibility* and *security* limitations.

## Fluid:

- ✓ Inject a sidecar to replace PVC with a fuse sidecar
- ✓ control the container startup sequence
- ✓ Also stop sidecar when main container exit



# Elasticity of Distributed Cache is important for LLMs



AI\_dev  
OPEN SOURCE SUMMIT

China 2024

## Problem:

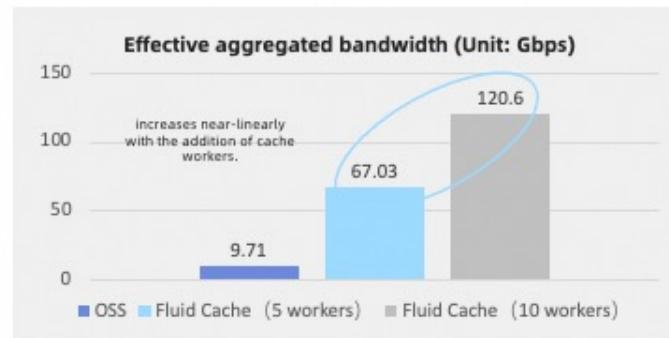
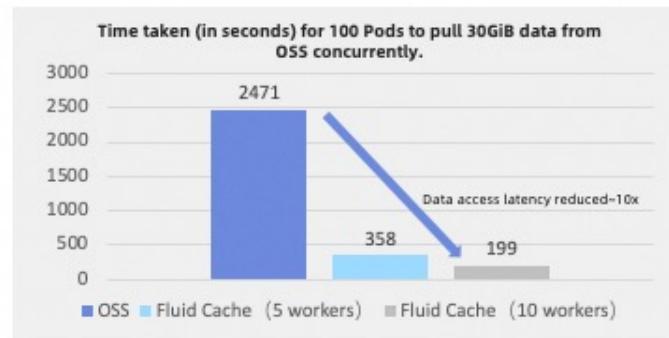
Limited storage-side bandwidth causes ‘cold start’ issues in large language model inference.

## Example:

If 10Gbps OSS Bucket bandwidth is equally split among 100 instances, with 0.1Gbps each, sequential read of a 30GB file could take 2400s.

## Fluid:

Using a [scalable distributed cache](#) optimizes bandwidth, which directly relates to [the size](#) and [number of cache nodes](#).



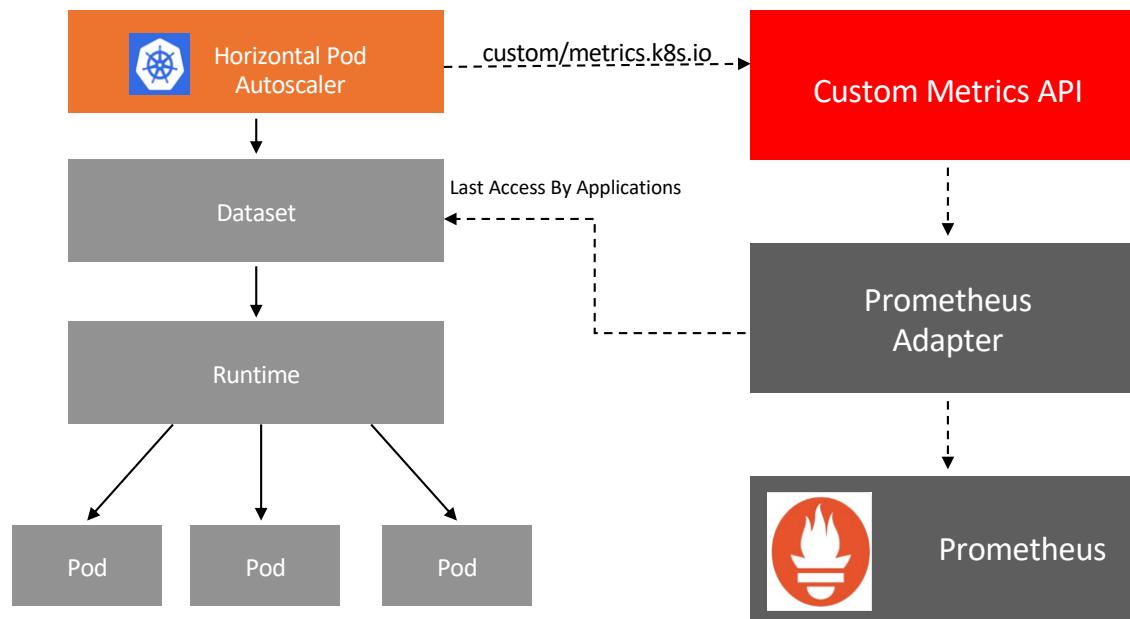
# Elasticity of Distributed Cache is important for LLMs

## Cost Strategies Across Different cases:

- ✓ One-time intensive loading for LLM models without repetition
- ✓ Frequent loading for image-text models, requiring hot caches
- ✓ Temporary cache expansion needed during business peaks

## Fluid:

- ✓ Fine-tuned control over data cache lifecycle, with elastic expansion and contraction based on business needs



```
apiVersion: autoscaling/v2beta2
kind: HorizontalPodAutoscaler
metadata:
  name: imagenet-hpa
spec:
  scaleTargetRef:
    apiVersion: data.fluid.io/v1alpha1
    kind: JindoRuntime
    name: imagenet
  minReplicas: 2
  metrics:
  - type: Object
    object:
      metric:
        name: capacity_used_rate
        describedObject:
          apiVersion: data.fluid.io/v1alpha1
          kind: Dataset
          name: imagenet
    target:
      type: Value
      value: "90"
```

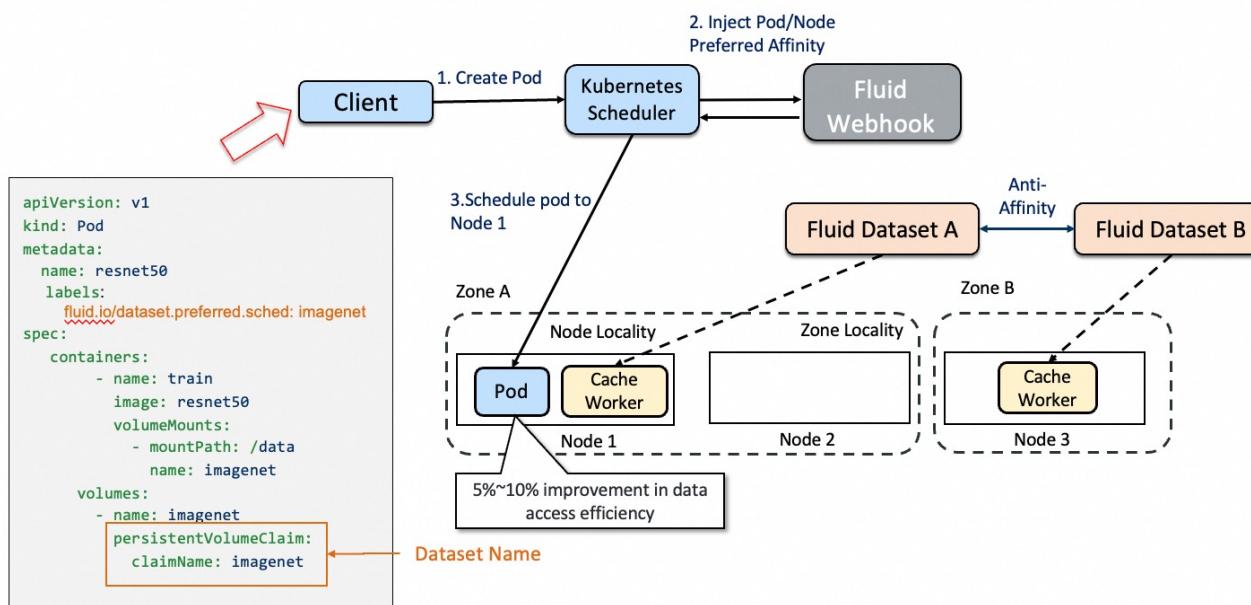
# Hierarchical Cache aware scheduling

## Problem:

LLMs often run in serverless environments cannot use third-party distributed caches due to *extensibility* and *security* limitations.

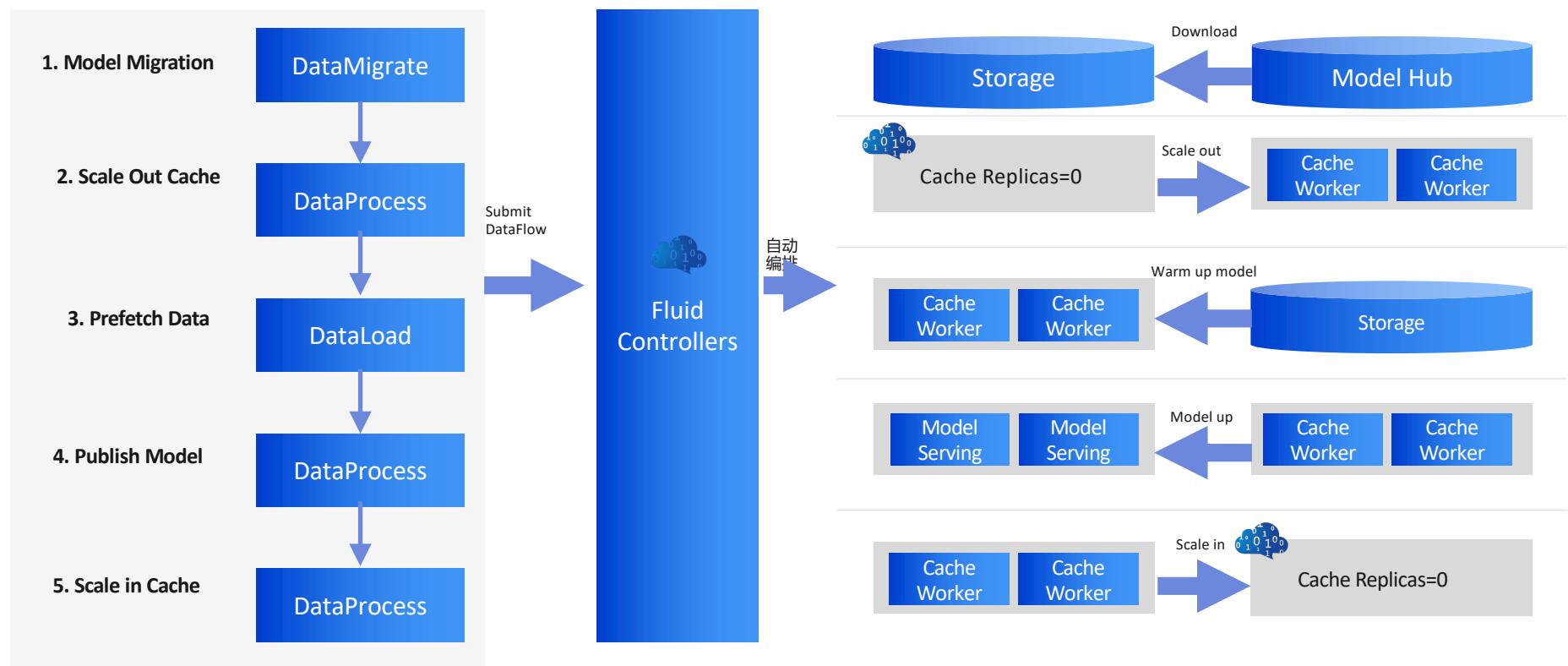
## Fluid:

- ✓ Inject a sidecar to replace PVC with a fuse sidecar
- ✓ control the container startup sequence
- ✓ Also stop sidecar when main container exit



# Dataflow

Use Fluid for Automated Operations and Maintenance to Balance Performance and Cost in LLMs Daily Work





A dark, abstract background featuring glowing, translucent geometric shapes in shades of pink, blue, and purple. These shapes include various polygons and arrows pointing towards the center. A prominent diagonal band of pink and blue light runs from the bottom left to the top right. In the center-left area, the word "Demo" is written in a large, white, sans-serif font.

Demo

# Demo



- Cache cluster by Fluid
- Cluster serving runtime by KServe
- Inference service by KServe
- Environment\*:
  - m5n.xlarge for cache cluster
  - g5.8xlarge for inference service
- Goals:
  - Fluid integration with KServe
  - OpenAI compatible API in KServe

\* The performance might be affected by the environment and setup.

# Demo



- Patch min replicas to 2
- Environment\*:
  - g5.8xlarge for inference service
- Goals:
  - Scaling feature of Inference Service
  - Speed up scaling by distributed cache

\* The performance might be affected by the environment and setup.

# Future Works



KubeCon



CloudNativeCon



THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI dev  
OPEN SOURCE DATA & ML SUMMIT

China 2024

- LLM Serving Runtimes: TGI, TRT-LLM etc
- LLM RAG Pipeline Orchestration
- GenAI Task APIs
- LLM Gateway
- Enhance Resilience of distributed cache



KubeCon



CloudNativeCon



China 2024



# Thank you & QA

Lize Cai  
Yang Che