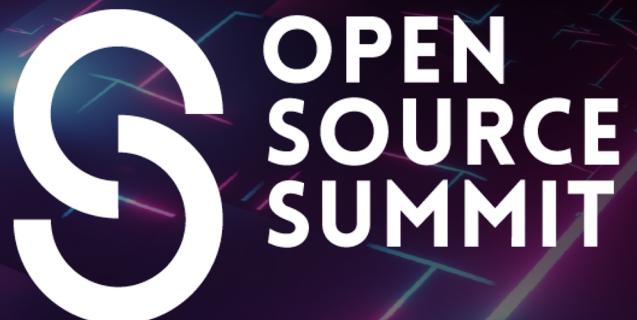


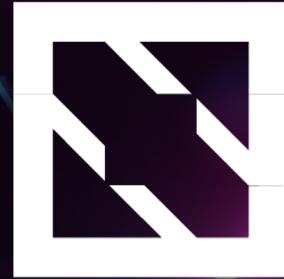


# KubeCon

THE LINUX FOUNDATION



China 2024



# CloudNativeCon





KubeCon



CloudNativeCon



China 2024

# JuiceFS CSI in Multi-Thousand Node Kubernetes Clusters for LLM Pre-Training

*Weiwei Zhu, Juicedata*

# About me



KubeCon



CloudNativeCon



China 2024



Weiwei Zhu

Fullstack Engineer at Juicedata

Maintainer of JuiceFS CSI Driver, Maintainer of Fluid

# Agenda

- Storage challenges for LLM training in K8s
- How JuiceFS addresses these challenges
- Optimizations for multi-thousand node clusters
- Demo with JuiceFS



KubeCon



CloudNativeCon



THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI\_dev  
Open Source DevOps & ML Summit

China 2024



KubeCon



CloudNativeCon



THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI\_dev  
Open Source Dev & ML Summit

China 2024

# 1. What are Storage challenges for LLM training in K8s?

# LLM Models



KubeCon



CloudNativeCon

THE LINUX FOUNDATION  
OPEN SOURCE SUMMITAI\_dev  
Open Source Dev & ML Summit

China 2024

Model	Parameters	Size
GLM	9 B	5.5 GB
Yi 1.5	34 B	19 GB
qwen2	72 B	41 GB
Llama 2	70 B	39 GB
Llama 3.1	70 B	40 GB
Llama 3.1	405 B	231 GB

# Storage for LLM training



KubeCon



CloudNativeCon



OPEN  
SOURCE  
SUMMIT



AI Dev  
Open Source Dev & ML Summit

China 2024



Tens of billions of files

Multi-cloud architecture

Mix of large and small files

Cost control

POSIX compliance

High data security

# Storage for LLM training in K8s



KubeCon



CloudNativeCon



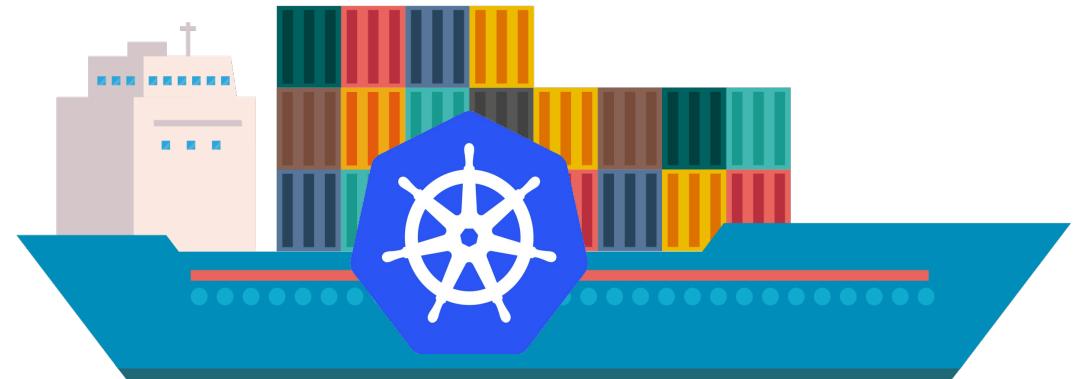
China 2024



Data **security**

Data **elasticity** in elastic clusters

Data **consistency** in multi-cloud



## 2. How JuiceFS addresses these challenges

# What is JuiceFS



China 2024

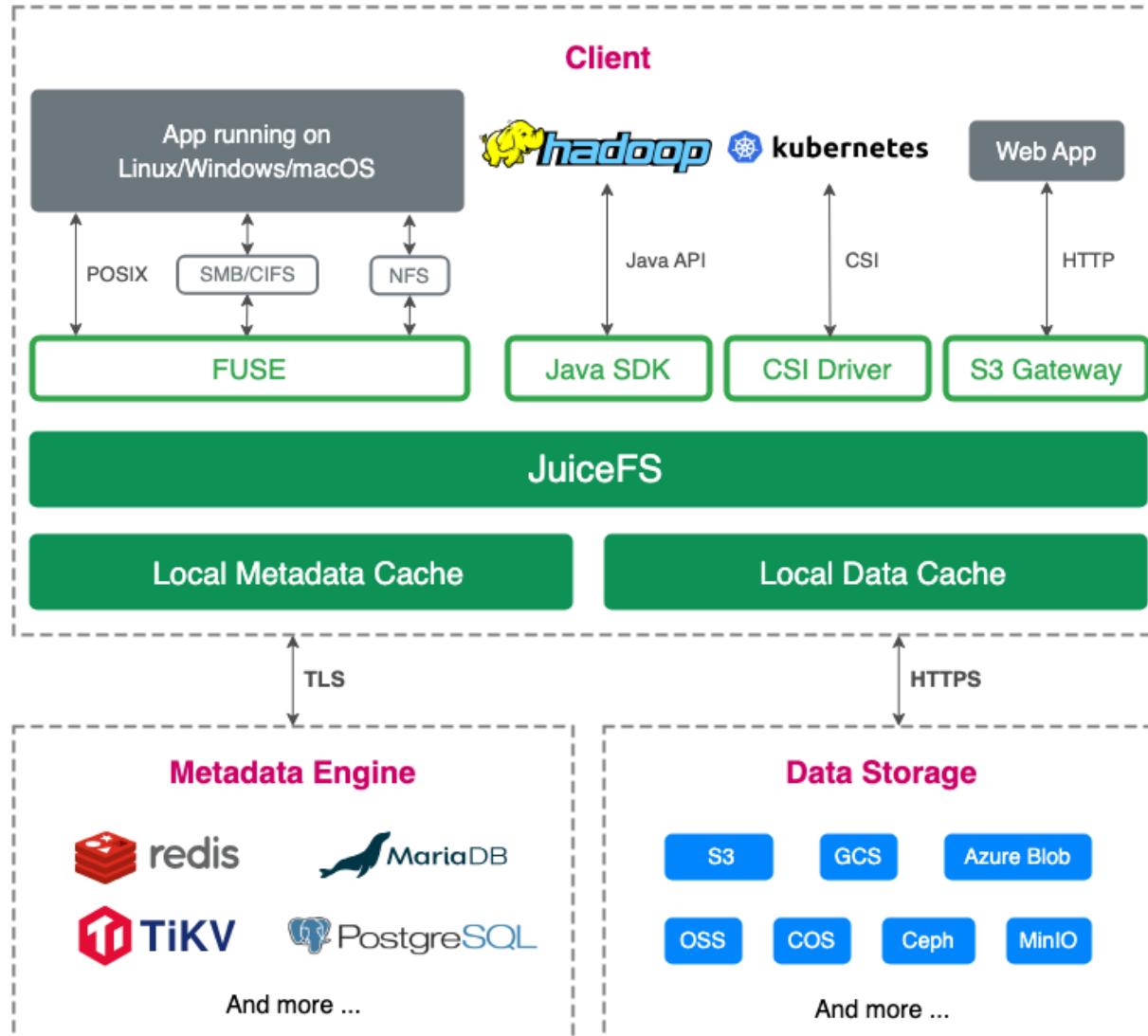


- An open-source, high-performance distributed file system designed for cloud
- Apache License 2.0
- **10k+ stars**

# What is JuiceFS



China 2024



## Client

- Handles all file I/O operations
- Multiple protocols

## Data Storage

- File data is split and stored in object storage
- Supports almost all types of object storage

## Metadata Engine

- Stores file metadata
- A variety of common databases, like Redis, TiKV, MySQL/MariaDB and PostgreSQL
- An in-house high-performance metadata engine

# JuiceFS in Kubernetes



KubeCon



CloudNativeCon

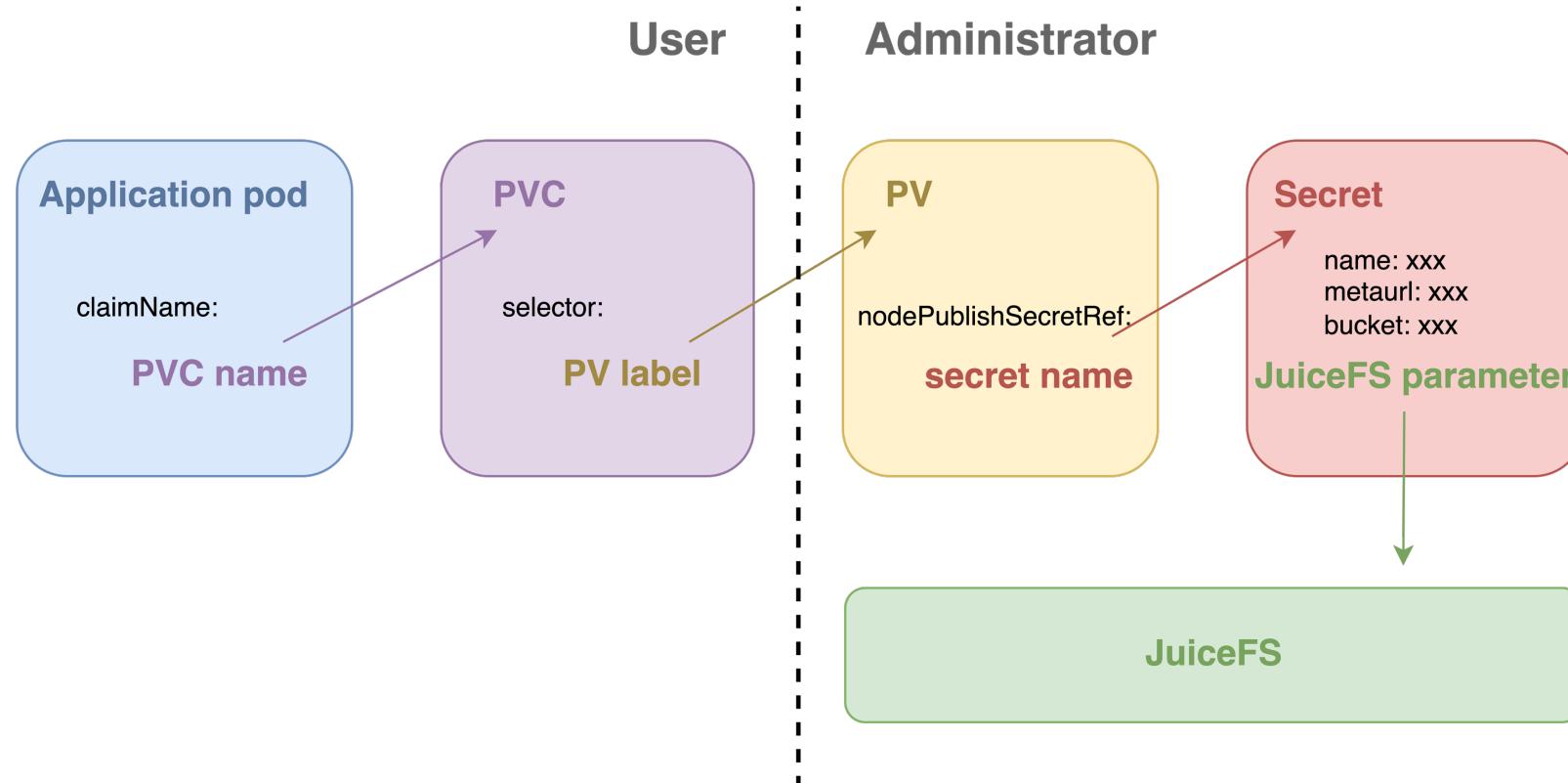


THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



China 2024

## Static Provision



# JuiceFS in Kubernetes



KubeCon



CloudNativeCon

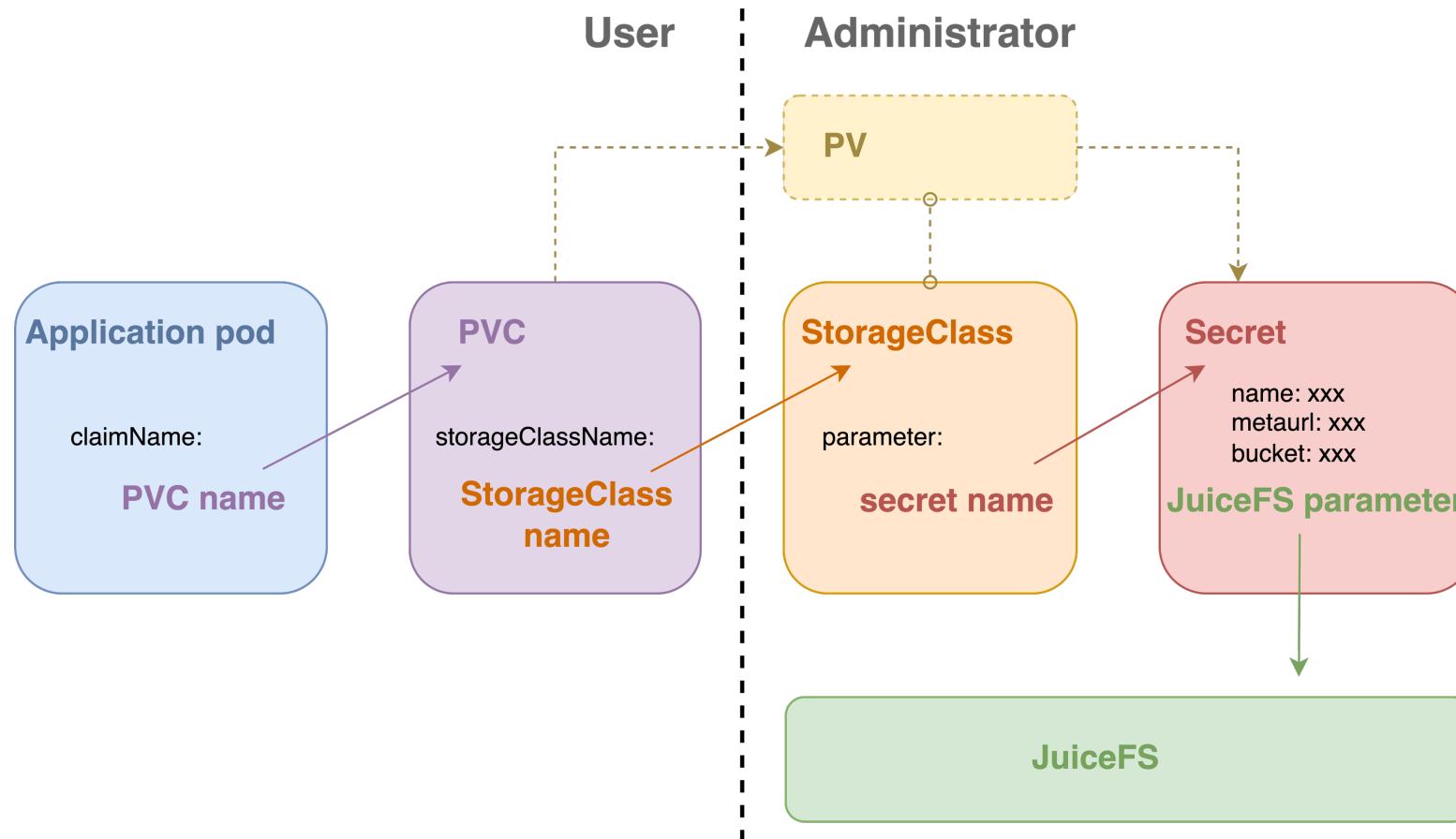


THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



China 2024

## Dynamic Provision



# JuiceFS in Kubernetes



KubeCon



CloudNativeCon

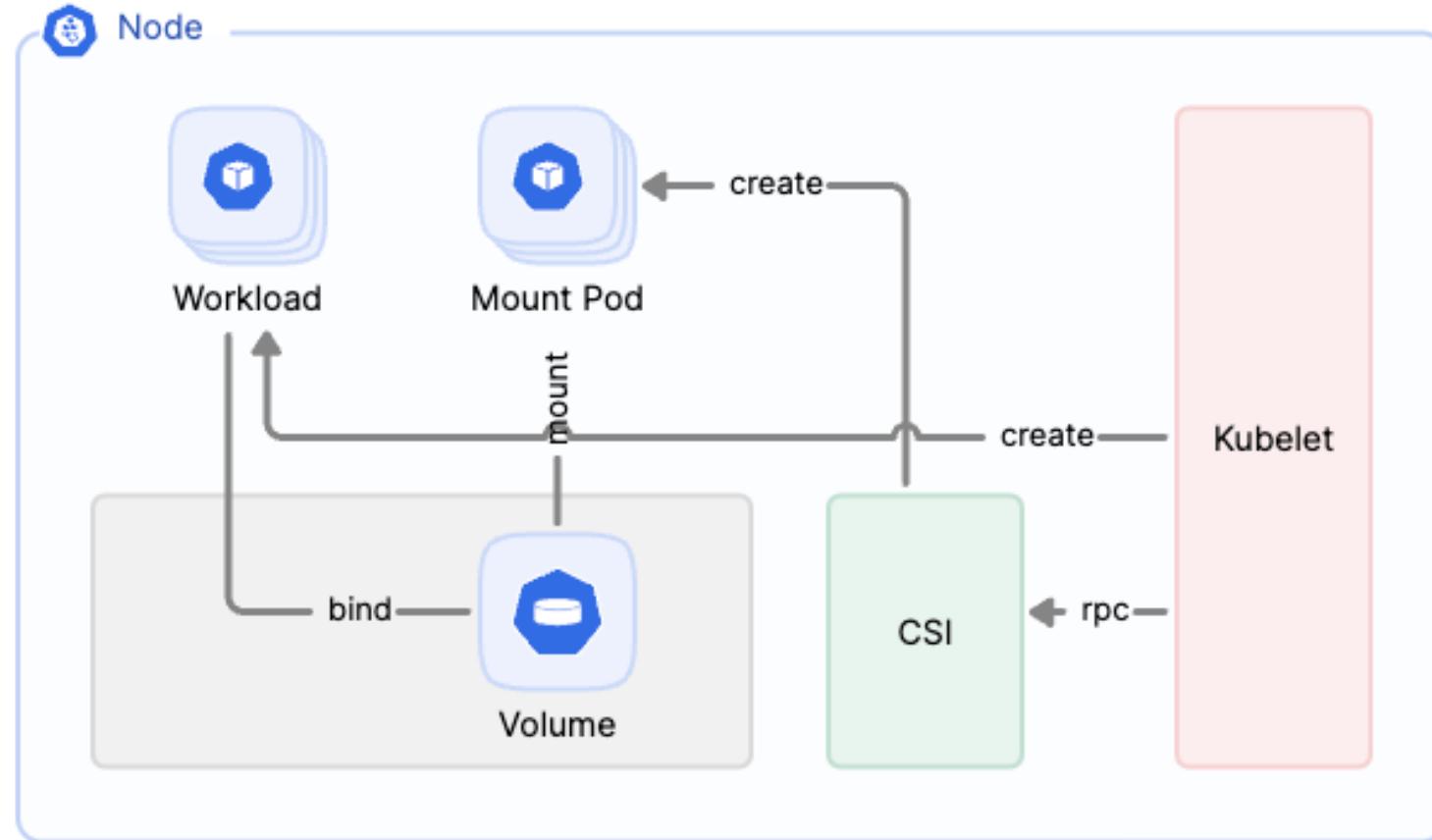


THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI\_dev  
Open Source Dev & ML Summit

China 2024



# JuiceFS in Serverless



KubeCon



CloudNativeCon

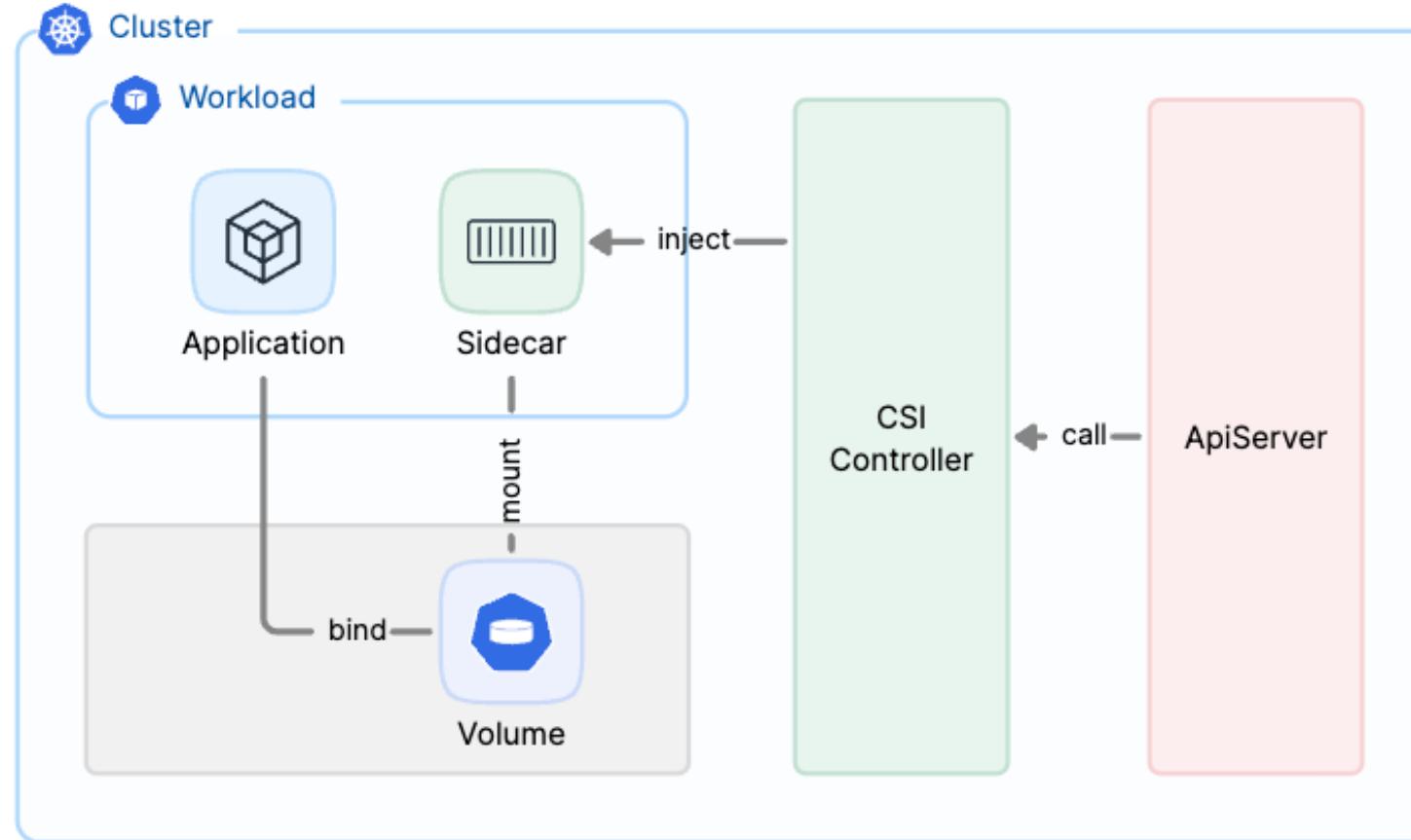


THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



Open Source Dev & ML Summit

China 2024



## Data isolation

For dynamic PVC, data will be stored in different directories.

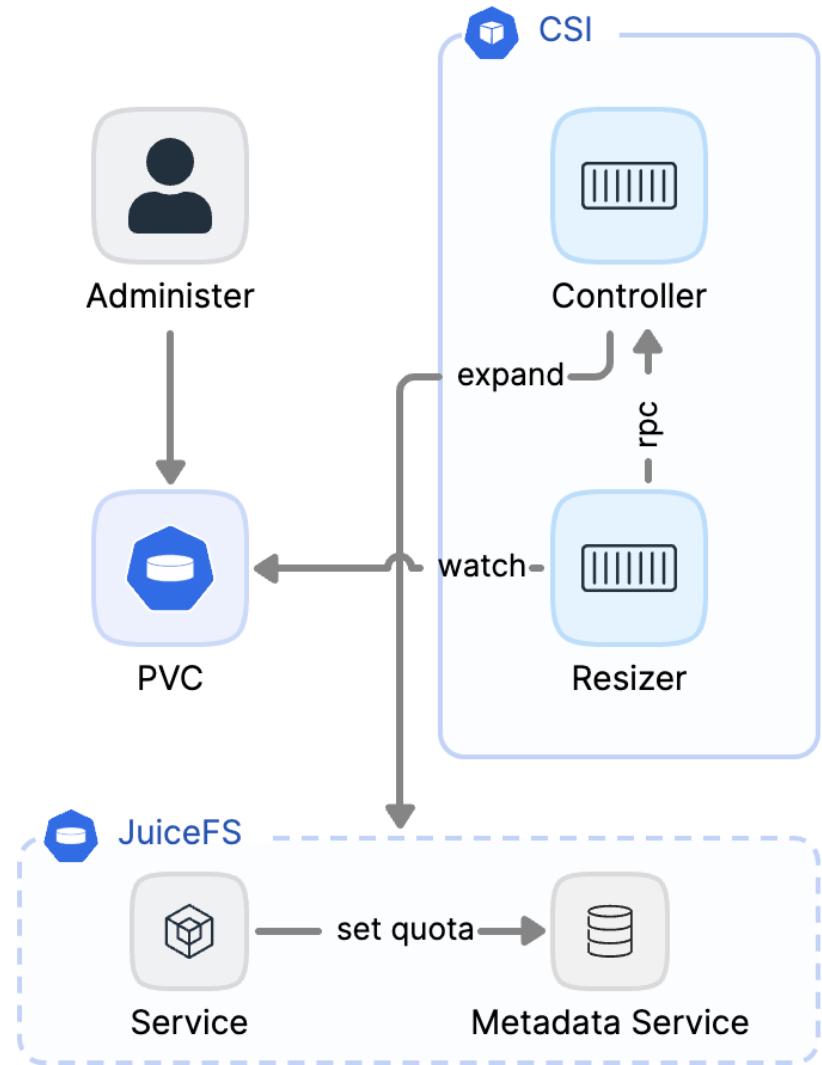
## Data encryption

Set passphrase and RSA private key in Secret which is used by StorageClass

## Permission control

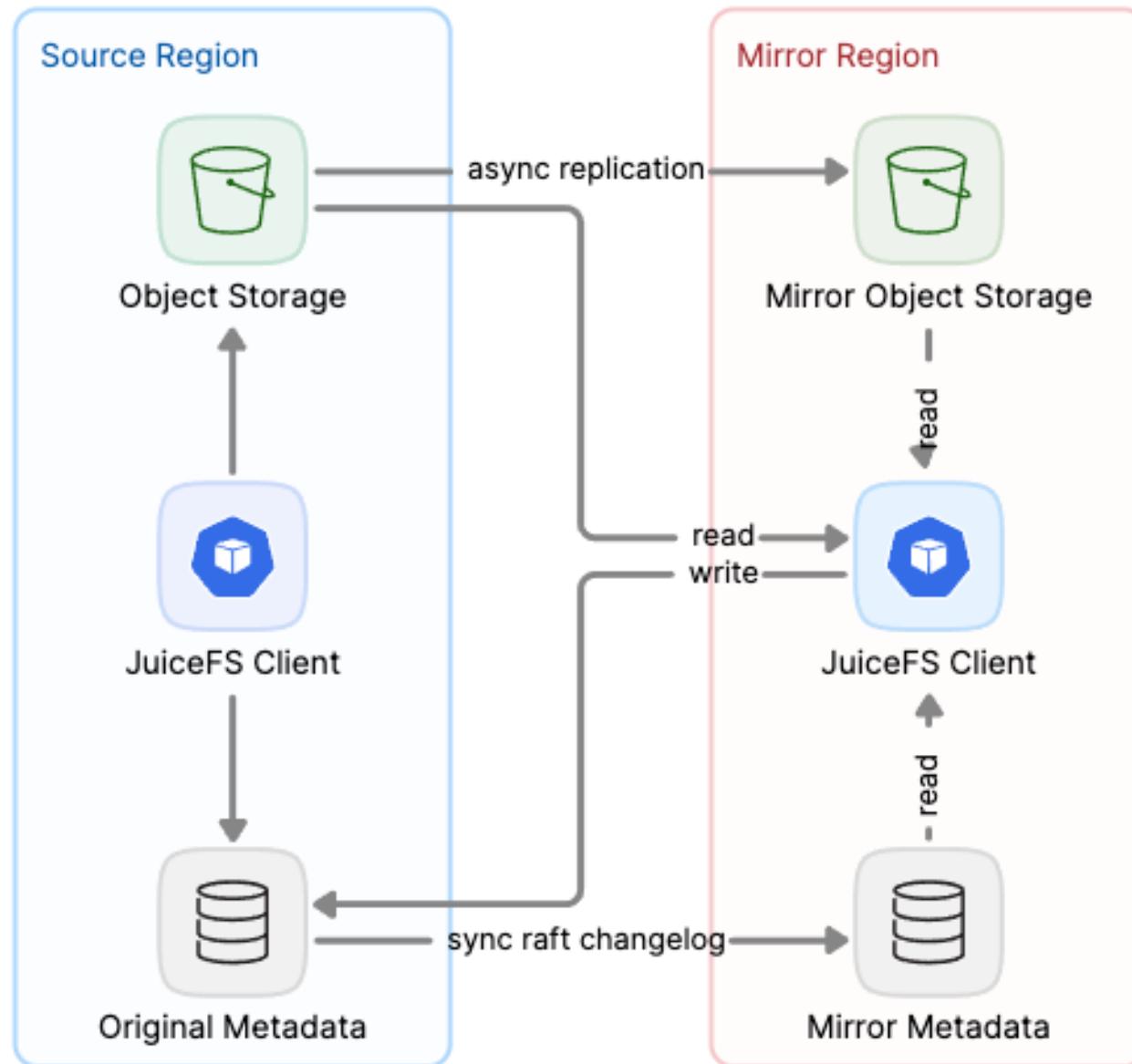
UID and GID of Unix-like systems to manage file permissions  
POSIX ACL permissions

# Data expansion in JuiceFS PV



```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: juicefs-pvc
  namespace: default
spec:
  accessModes:
    - ReadWriteMany
  volumeMode: Filesystem
  storageClassName: juicefs-sc
  resources:
    requests:
      storage: 10Gi
```

# Mirror file system



- Data in multi-cloud
- Metadata sync raft changelog to mirror
- Object Storage async to mirror

# Cache Group of JuiceFS



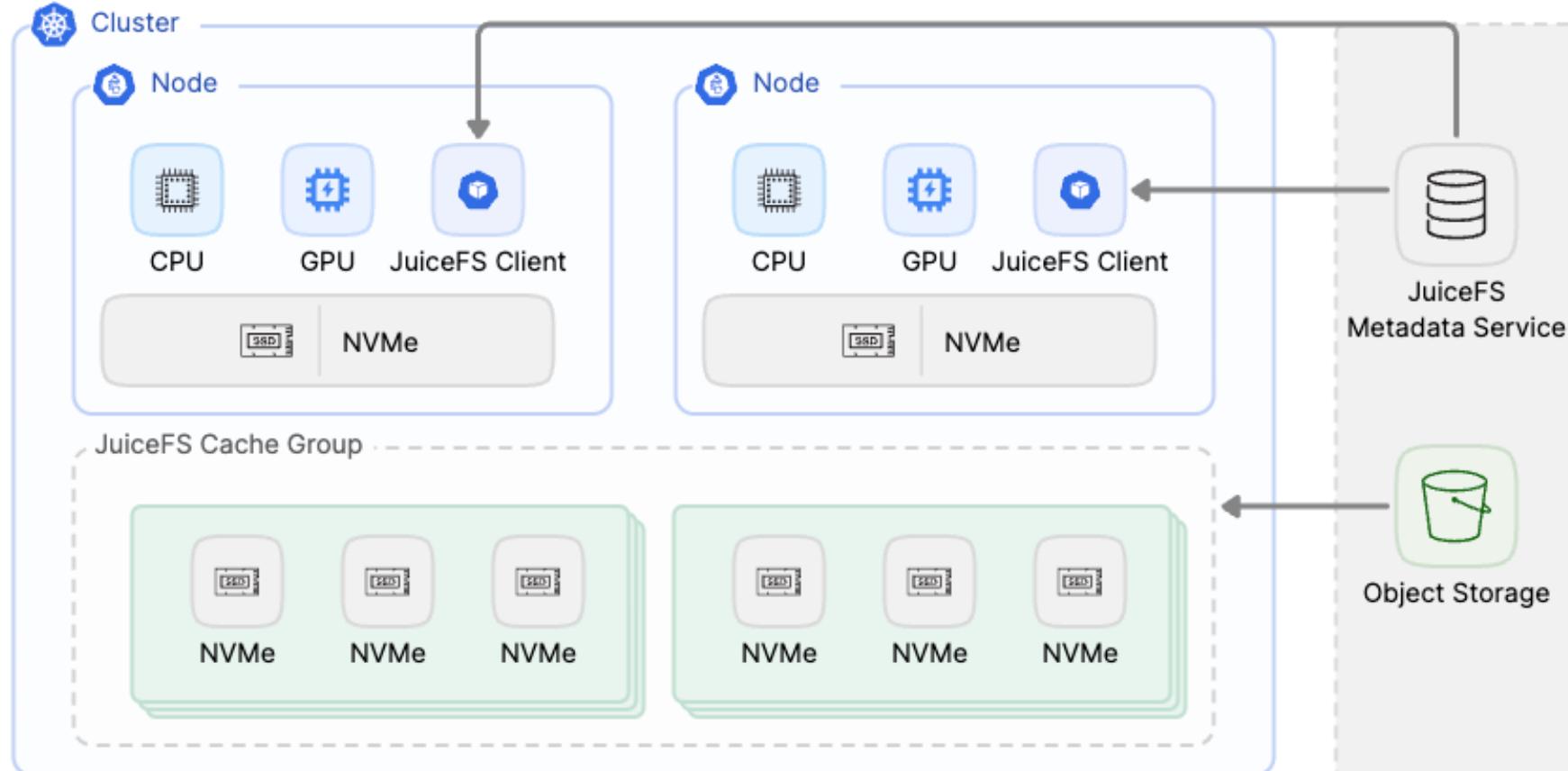
KubeCon



CloudNativeCon

THE LINUX FOUNDATION  
OPEN SOURCE SUMMITAI\_dev  
Open Source Dev & ML Summit

China 2024



# Benchmark of Cache Group



KubeCon

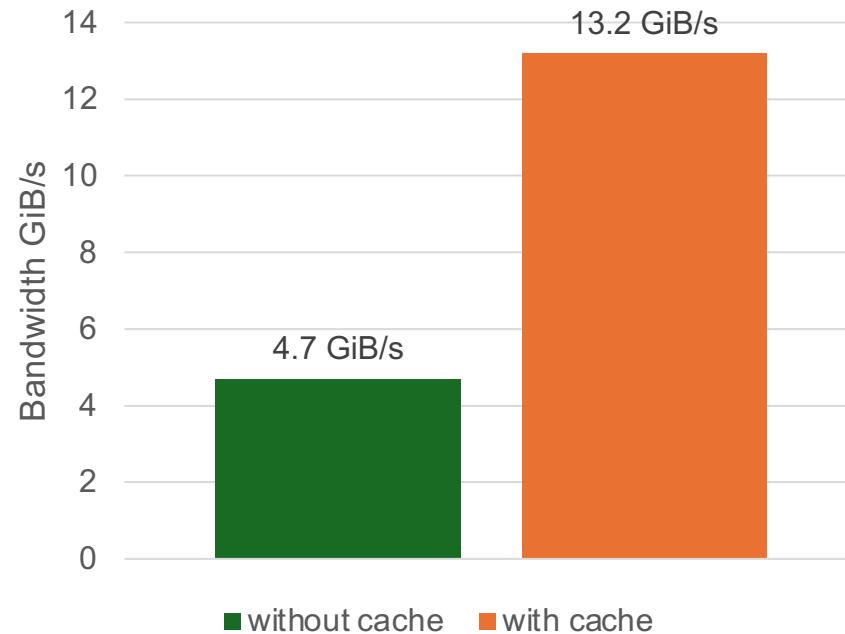


CloudNativeCon

THE LINUX FOUNDATION  
OPEN SOURCE SUMMITAI\_dev  
Open Source Dev & ML Summit

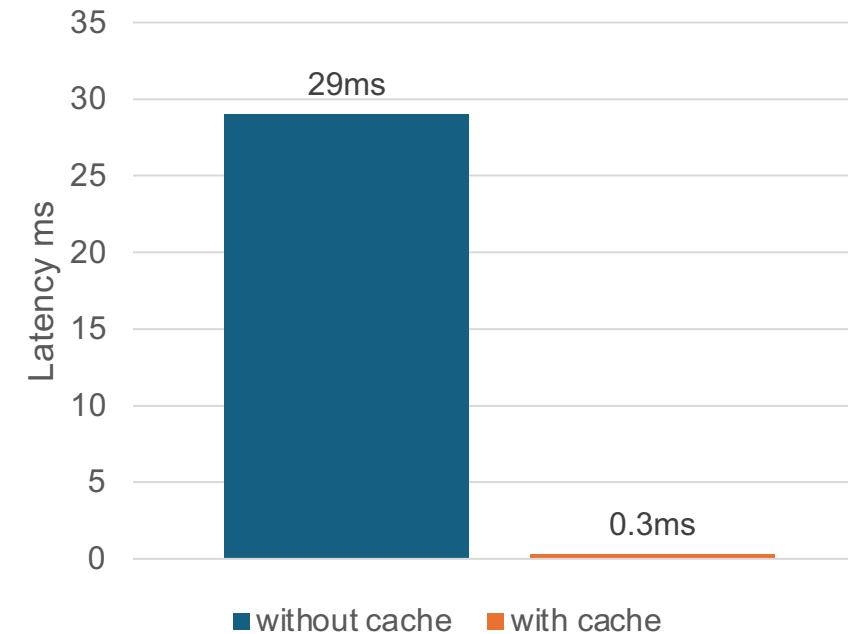
China 2024

## Big file sequential read



File: 10 GB  
Network speed: 200Gbps

## Big file random read



File: 10 GB  
Network speed: 200Gbps

### 3. Optimizations for multi-thousand node clusters

# Multi-Thousand Node Clusters



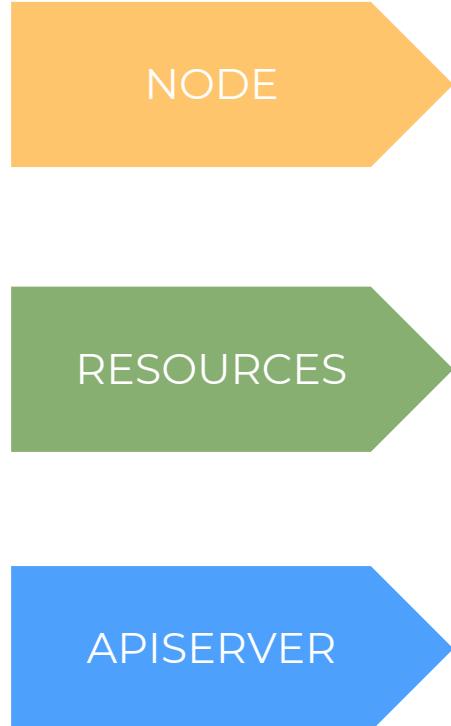
KubeCon



CloudNativeCon



China 2024



Too many nodes, difficult to operate and maintain

Too many resources

ApiServer is under great pressure

# Visualization



KubeCon



CloudNativeCon

THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT

China 2024

JuiceFS CSI

② ⚙️ 🔍

Application Pod

System Pod

PV

PVC

Storage Class

## Application Pod

Name: Please enter

Namespace: Please enter

PV: Please enter

Mount Pods: Please enter

Status: Please select

## Application Pod List ⓘ

C I ⚙️

Name	Namespace	PV	Mount Pods	Status	CSI Node	CreateTime
ce-dynamic-6c6b54478d-pwd6b	kube-system	• pvc-9a09305d-5004-45ec-83c9-0e626ae3b59d	• juicefs-cn-hangzhou.10.0.1.84-dynamic-ce-drkagd	• Running	• juicefs-csi-node-p6ccl	2024/8/2 14:25:50
ce-dynamic-6c6b54478d-mrh87	default	• pvc-a77b0ac6-7a24-4545-8484-76f9f2fac5db	• juicefs-cn-hangzhou.10.0.1.84-dynamic-ce-drkagd	• Running	• juicefs-csi-node-p6ccl	2024/8/2 11:11:44
ce-static-54445fc7dbd-r7s98	default	• ce-static	• juicefs-cn-hangzhou.10.0.1.84-ce-static-handle-bspgdx	• Running	• juicefs-csi-node-p6ccl	2024/7/31 17:46:45
normal-664f8b8846-mt4fb	default	• ce-static	• juicefs-cn-hangzhou.10.0.1.84-ce-static-handle-kuudkc	• Running	• juicefs-csi-node-p6ccl	2024/7/17 10:21:56
cn-wrong-7b7577678d-l7pcz	default	• ce-static	• juicefs-cn-hangzhou.10.0.1.84-ce-static-handle-kuudkc	• CrashLoopBackOff	• juicefs-csi-node-p6ccl	2024/7/17 10:21:56
pending	sidecar	• ce-sidecar	• juicefs-cn-hangzhou.10.0.1.84-ce-sidecar-handle-cltobf	• Pending	• juicefs-csi-node-p6ccl	2024/7/17 10:21:56
res-err	default	• ce-static	-	• Pending	-	2024/6/24 11:38:00

PVC which it uses was not successfully bound, please click "PVC" to view details.

# Resource and performance optimization



KubeCon



CloudNativeCon

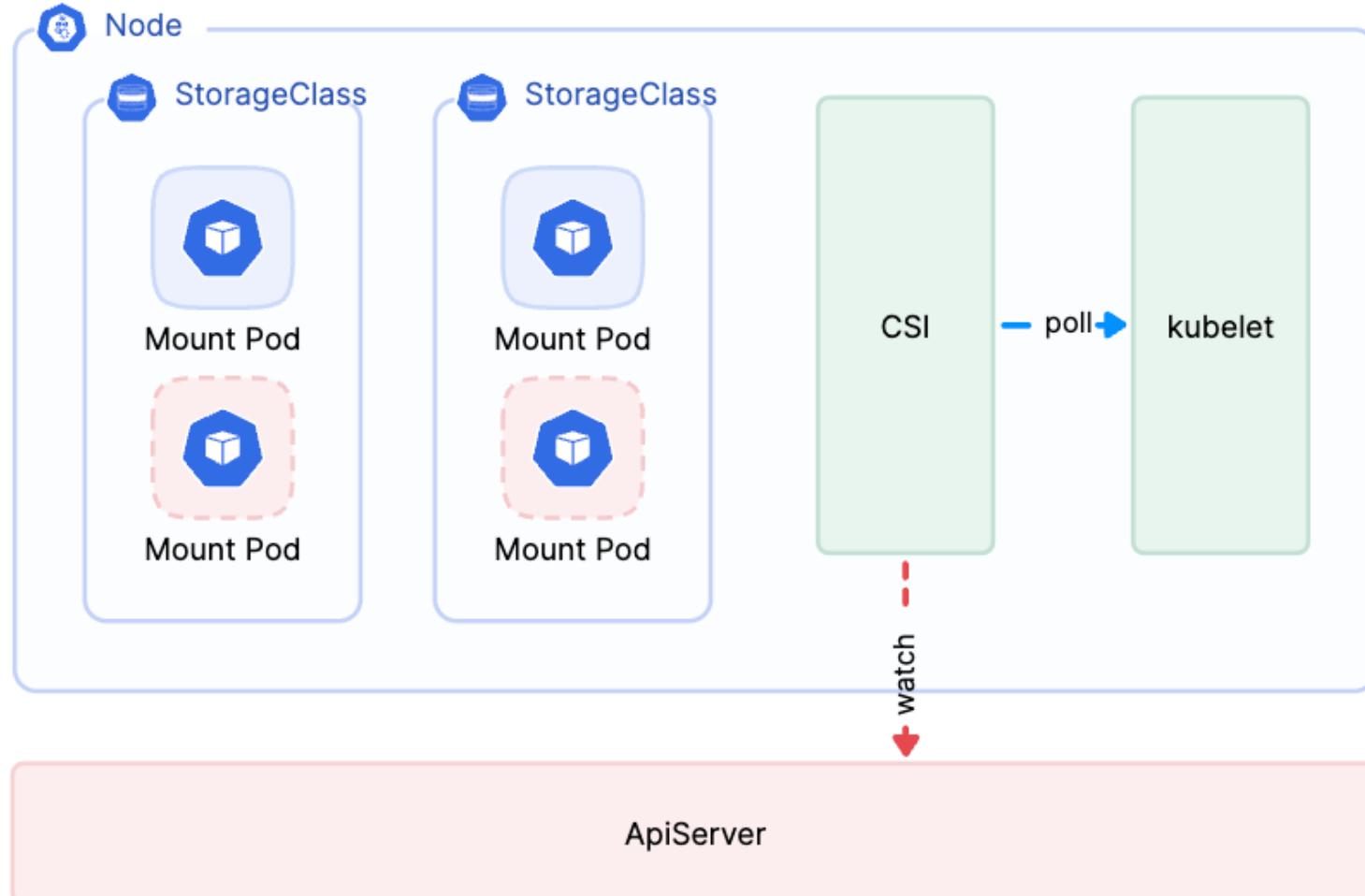


THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI\_dev  
Open Source Dev & ML Summit

China 2024



- PVC/StorageClass
- Shared Mount Pod
- Poll kubelet instead of watching ApiServer

# Stability optimization



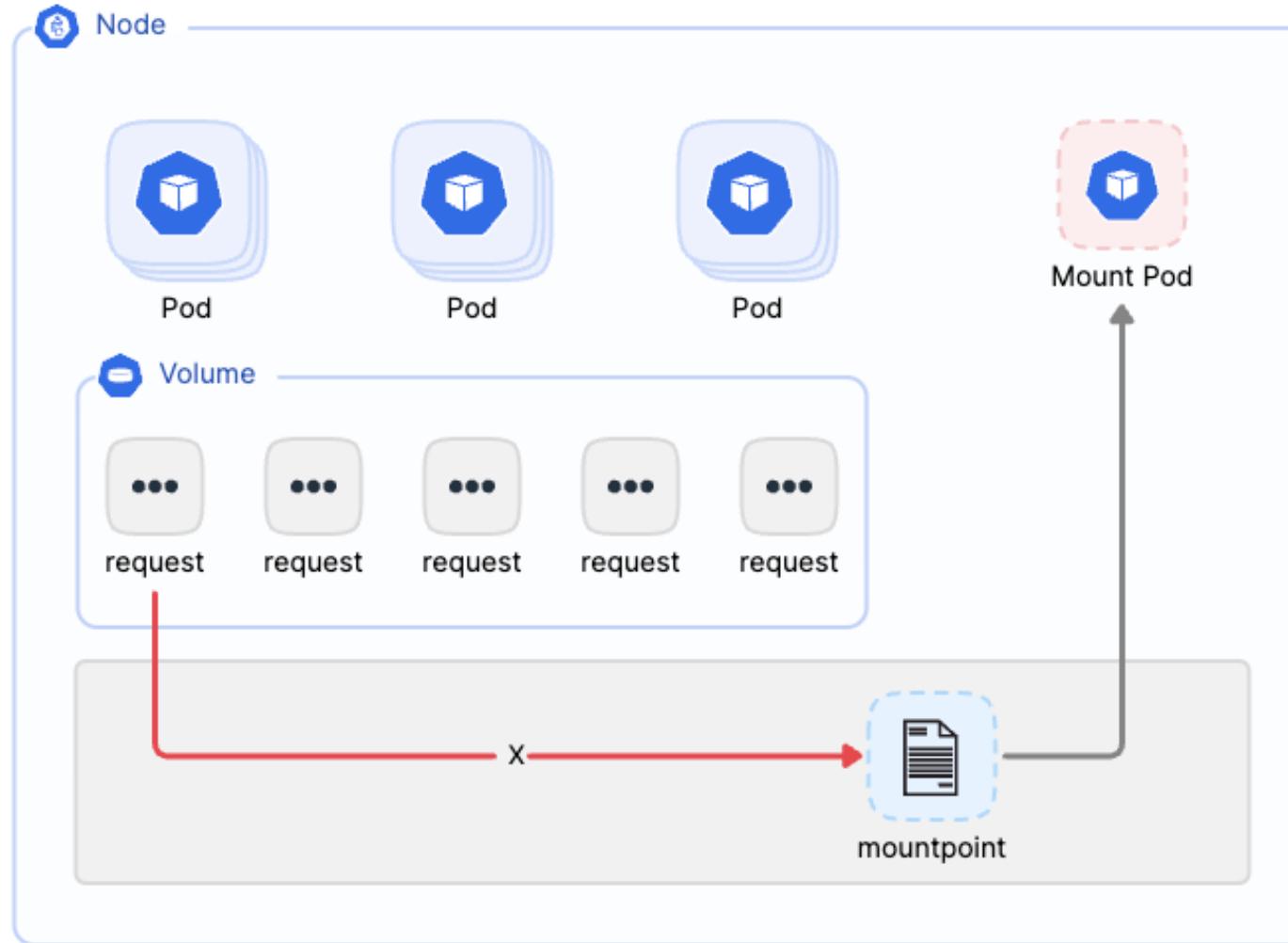
KubeCon



CloudNativeCon

THE LINUX FOUNDATION  
OPEN SOURCE SUMMITAI\_dev  
Open Source Dev & ML Summit

China 2024



Data requests in application pod will be broken when mount pod is deleted for OOM or some other reasons

# Stability optimization



KubeCon



CloudNativeCon

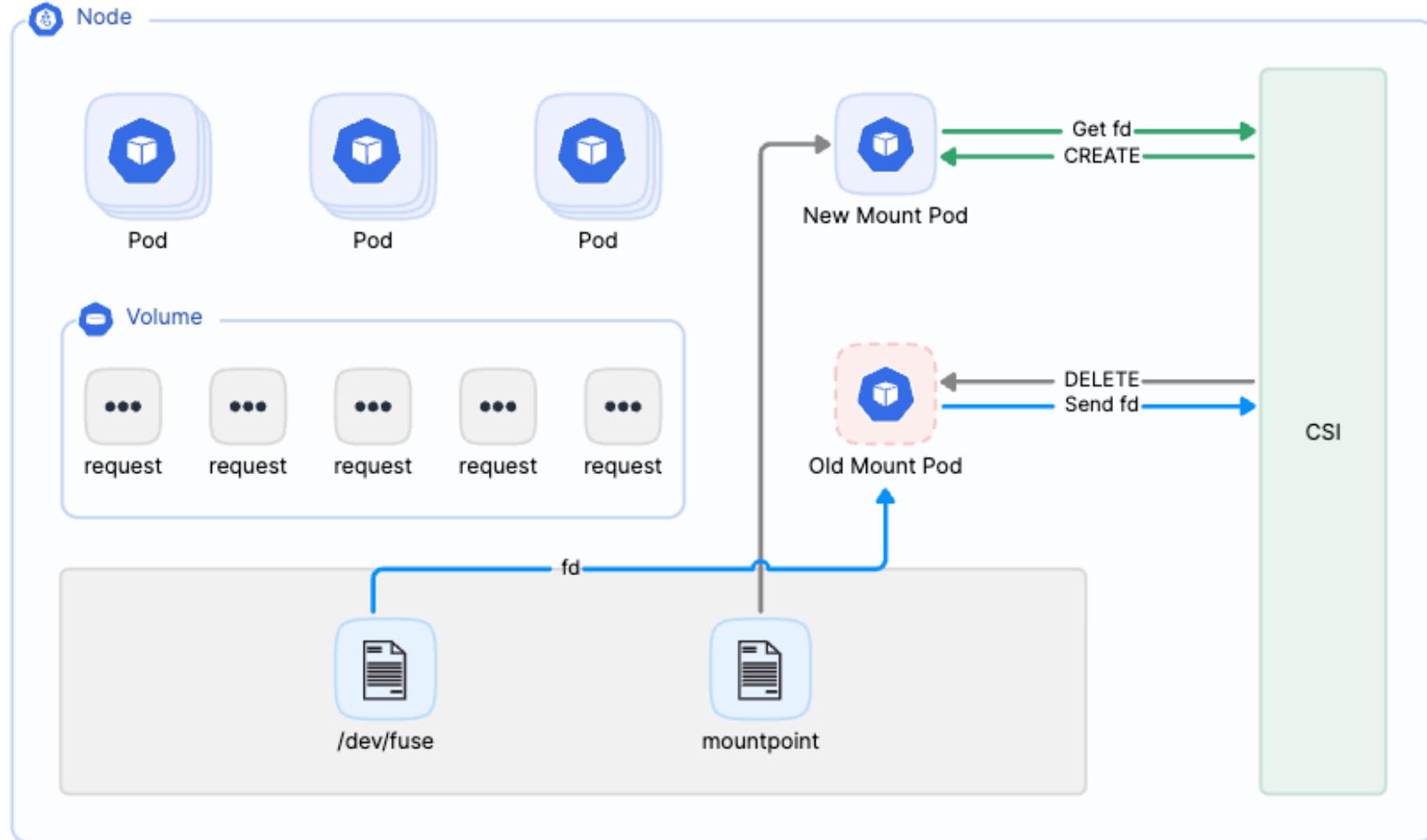


THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



Open Source Dev & ML Summit

China 2024



## Mount pod up:

- Gets FUSE fd from /dev/fuse
- Sends FUSE fd to CSI

## Old mount pod:

- CSI deletes it

## New mount pod:

- CSI creates it
- Gets FUSE fd from CSI

# Smoothly upgrade



KubeCon



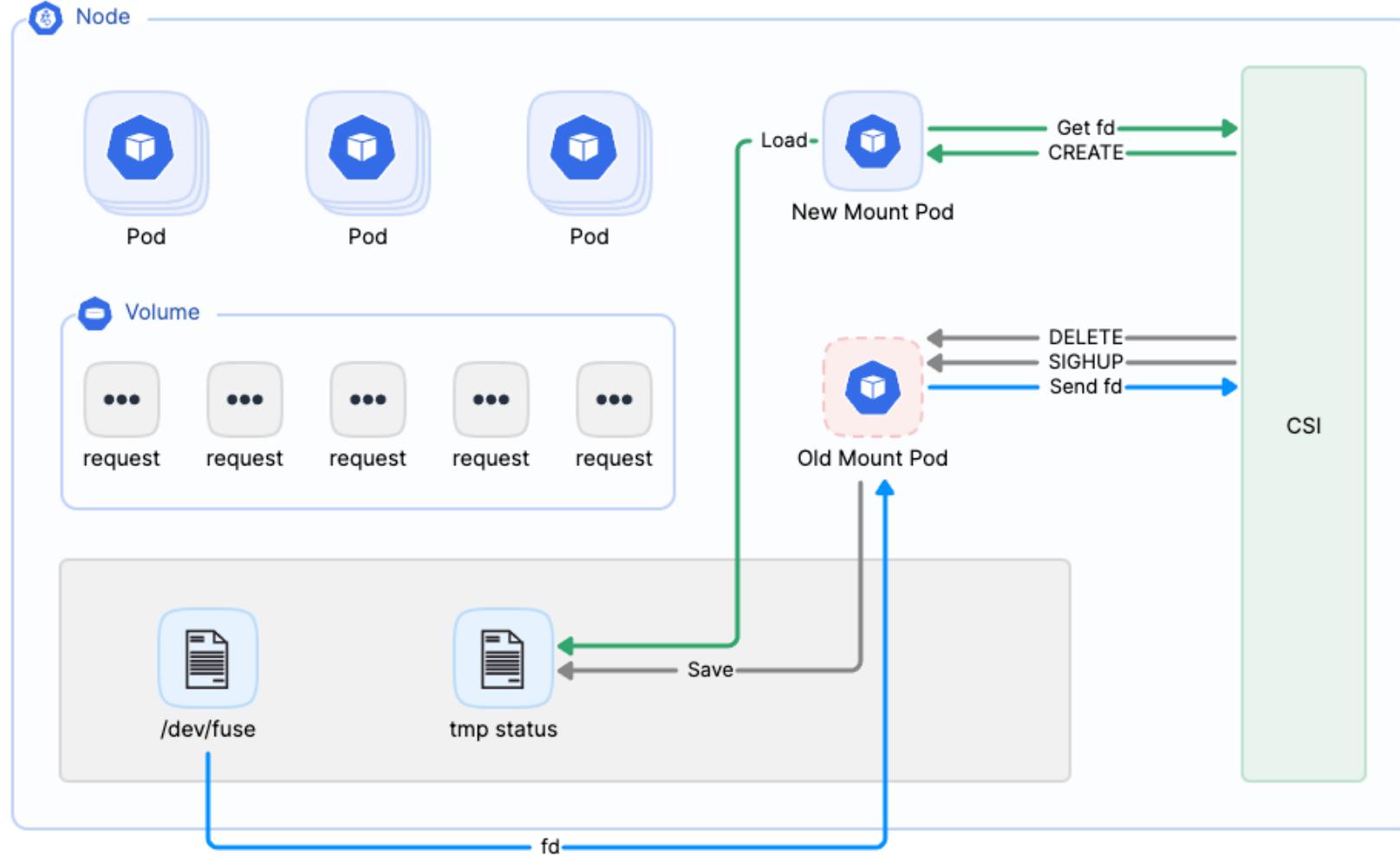
CloudNativeCon



THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



China 2024



## Mount pod up:

- Get FUSE fd from /dev/fuse
- Send FUSE fd to CSI

## Old mount pod:

- CSI send SIGHUP
- Save status in tmp file
- CSI delete it

## New mount pod:

- CSI create it
- Get FUSE fd from CSI
- Load status from tmp file



KubeCon



CloudNativeCon



THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



China 2024

## 4. Demo with JuiceFS

# Demo with JuiceFS



KubeCon



CloudNativeCon



China 2024



- How to use JuiceFS in Kubernetes
- Impact of JuiceFS on reading speeds for large models

[root@iZbp1hizg4bntjyyiu82qlZ demo]#

I

# Conclusion



KubeCon



CloudNativeCon



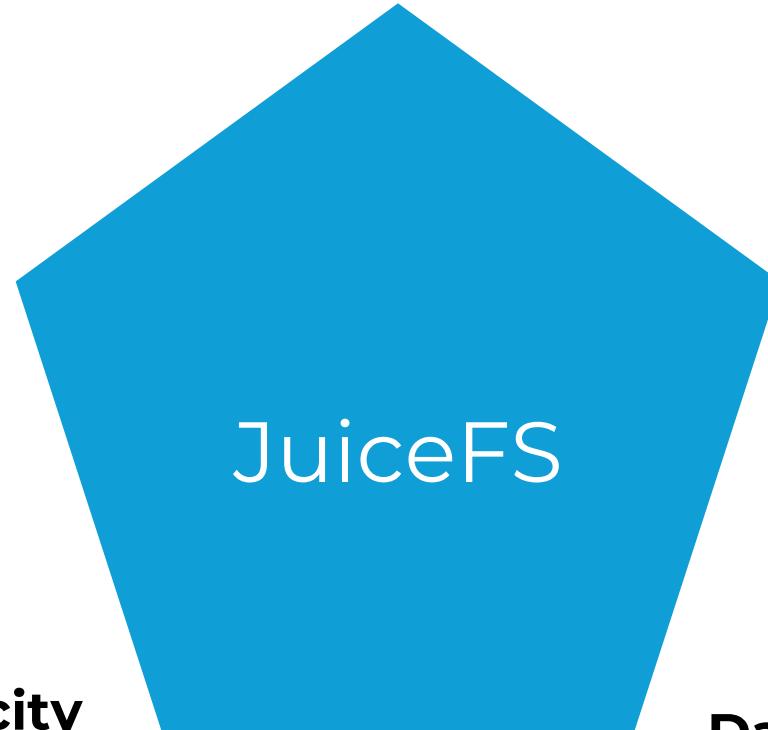
THE LINUX FOUNDATION  
OPEN SOURCE SUMMIT



AI\_dev  
Open Source Dev & ML Summit

China 2024

## POSIX compliance



### High performance & Cost control

Distributed cache group

### Data elasticity

Sidecar in Serverless /  
Data expansion

### High data security

Isolation / encryption /  
Permission control

### Data consistency

Mirror file system

# Thanks



<https://juicefs.com>



<https://github.com/juicedata/juicefs>

