KubeCon CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

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

Weiwei Zhu, Juicedata

About me

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

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

LLM Models

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

- 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

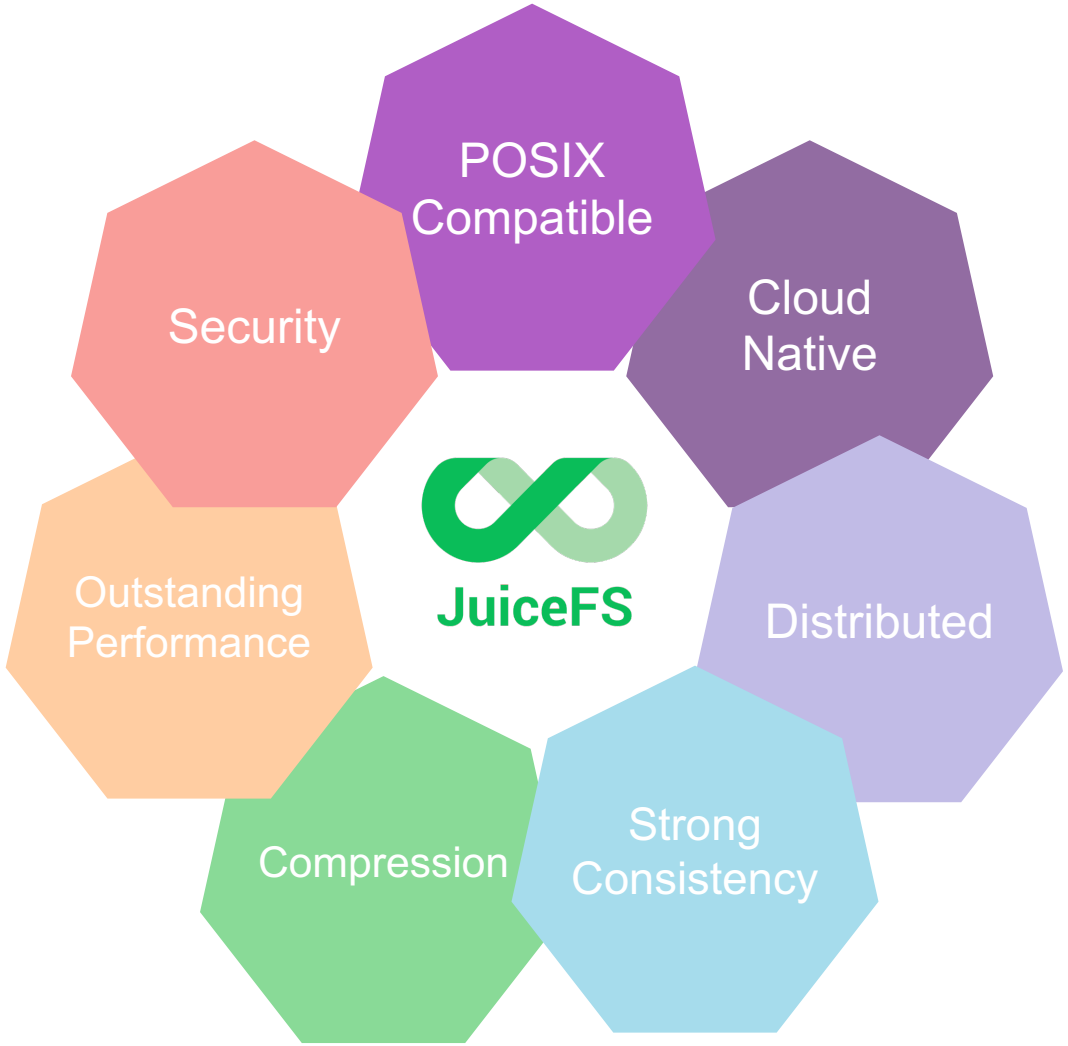
Data security

Data elasticity in elastic clusters

Data consistency in multi-cloud

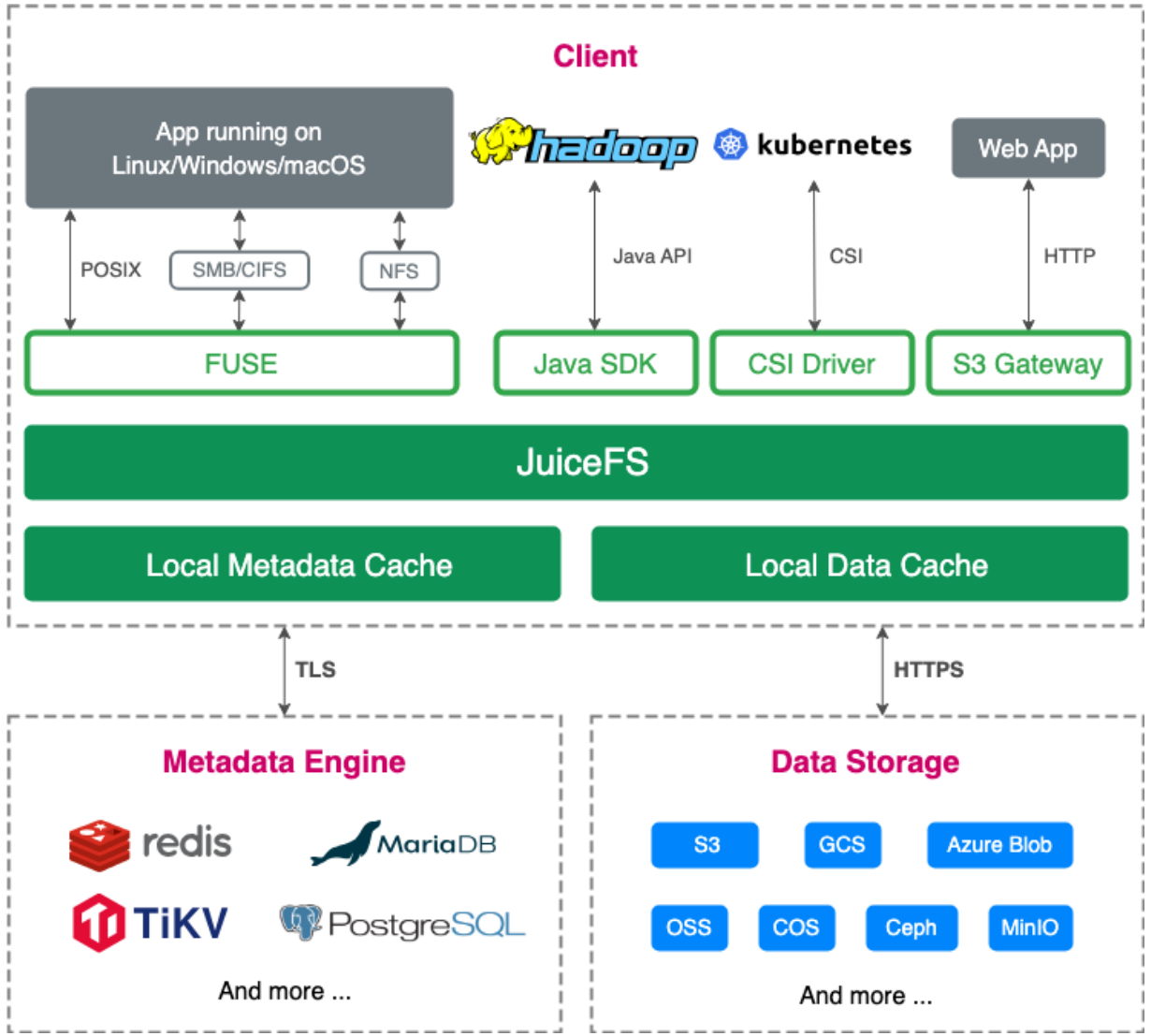
2. How JuiceFS addresses these challenges

What is JuiceFS



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

What is JuiceFS



Client

- Handles all file I/O operations
- Multiple protocols

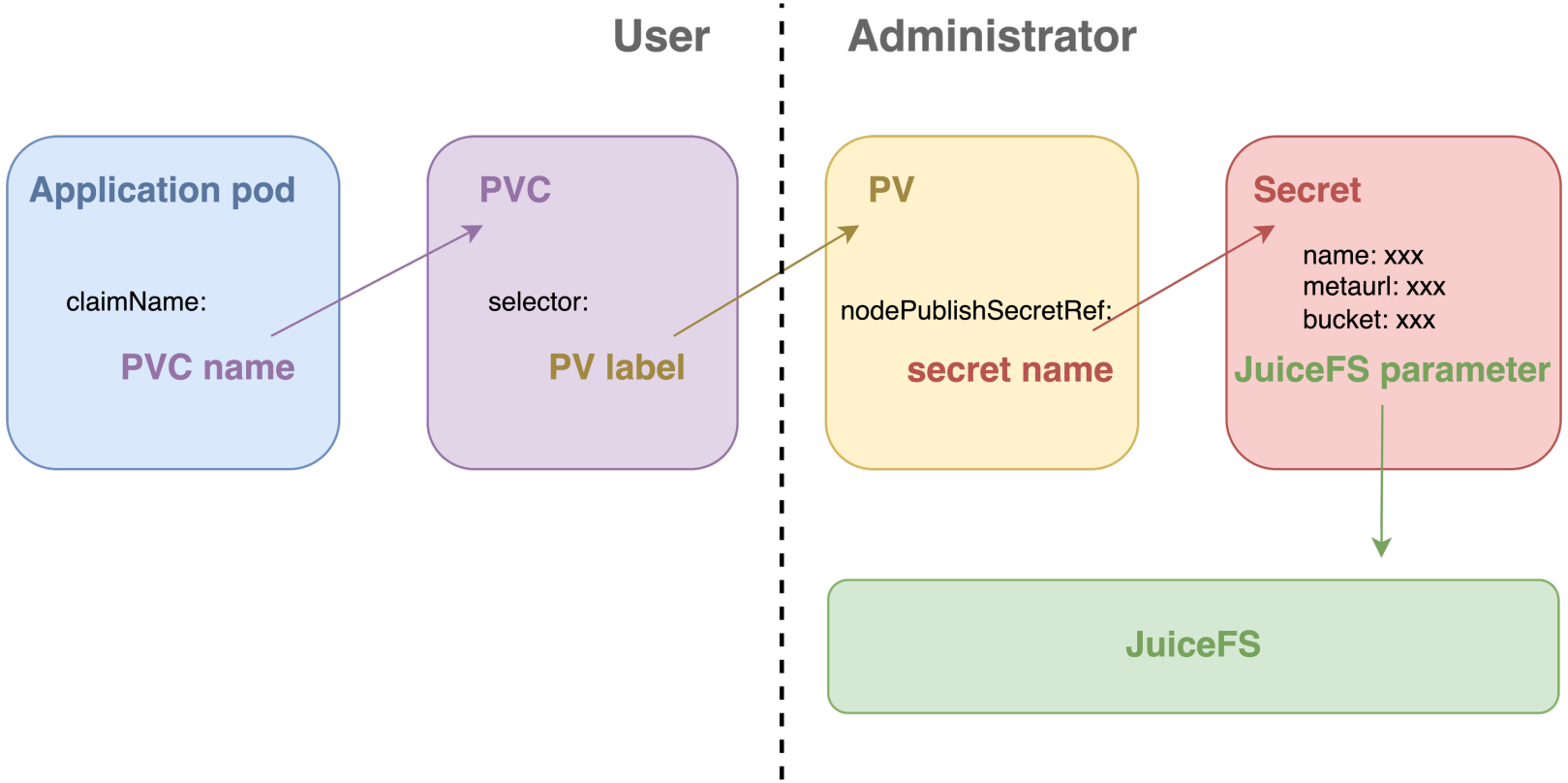
Data Storage

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

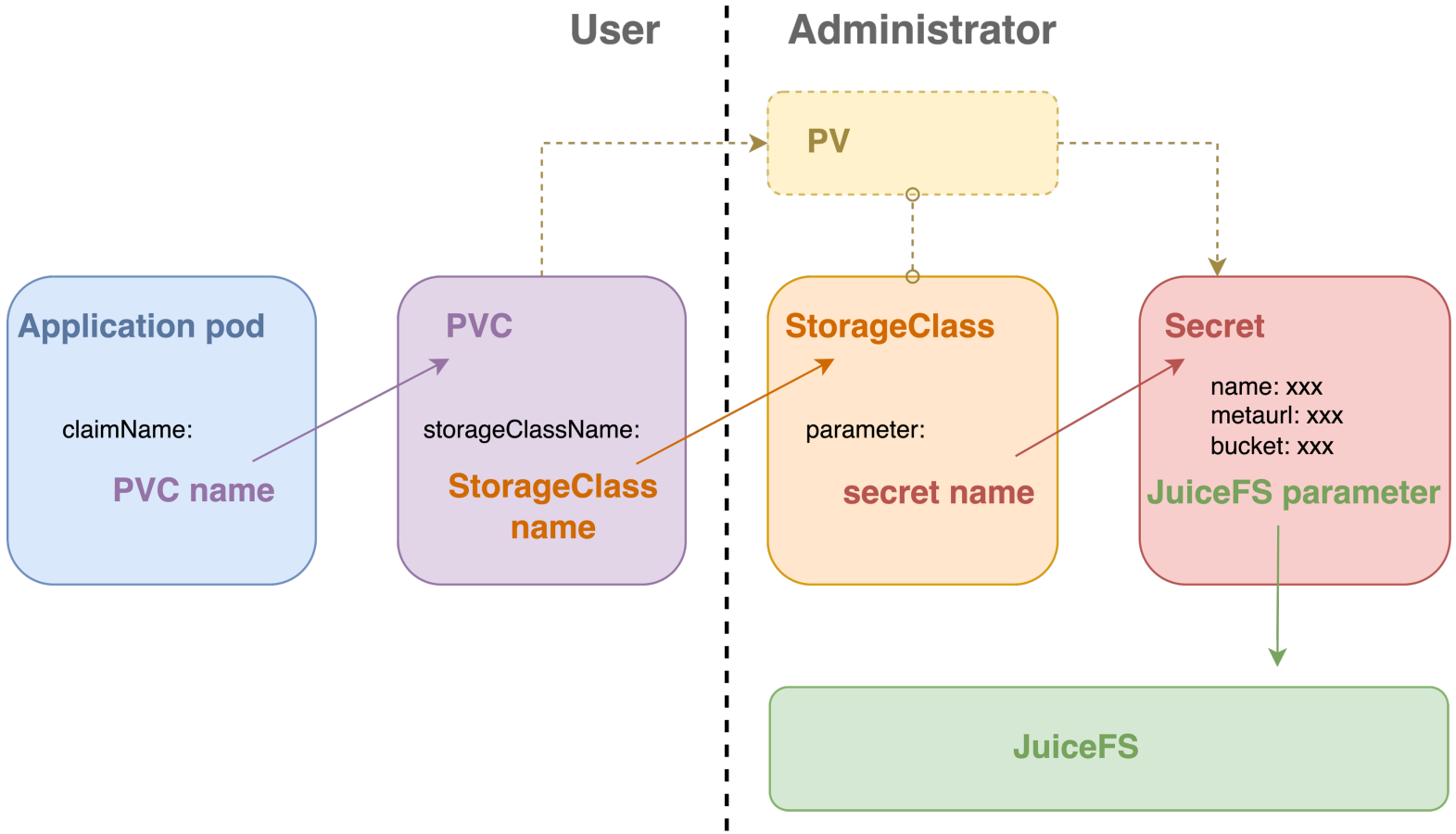
Metadata Engine

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

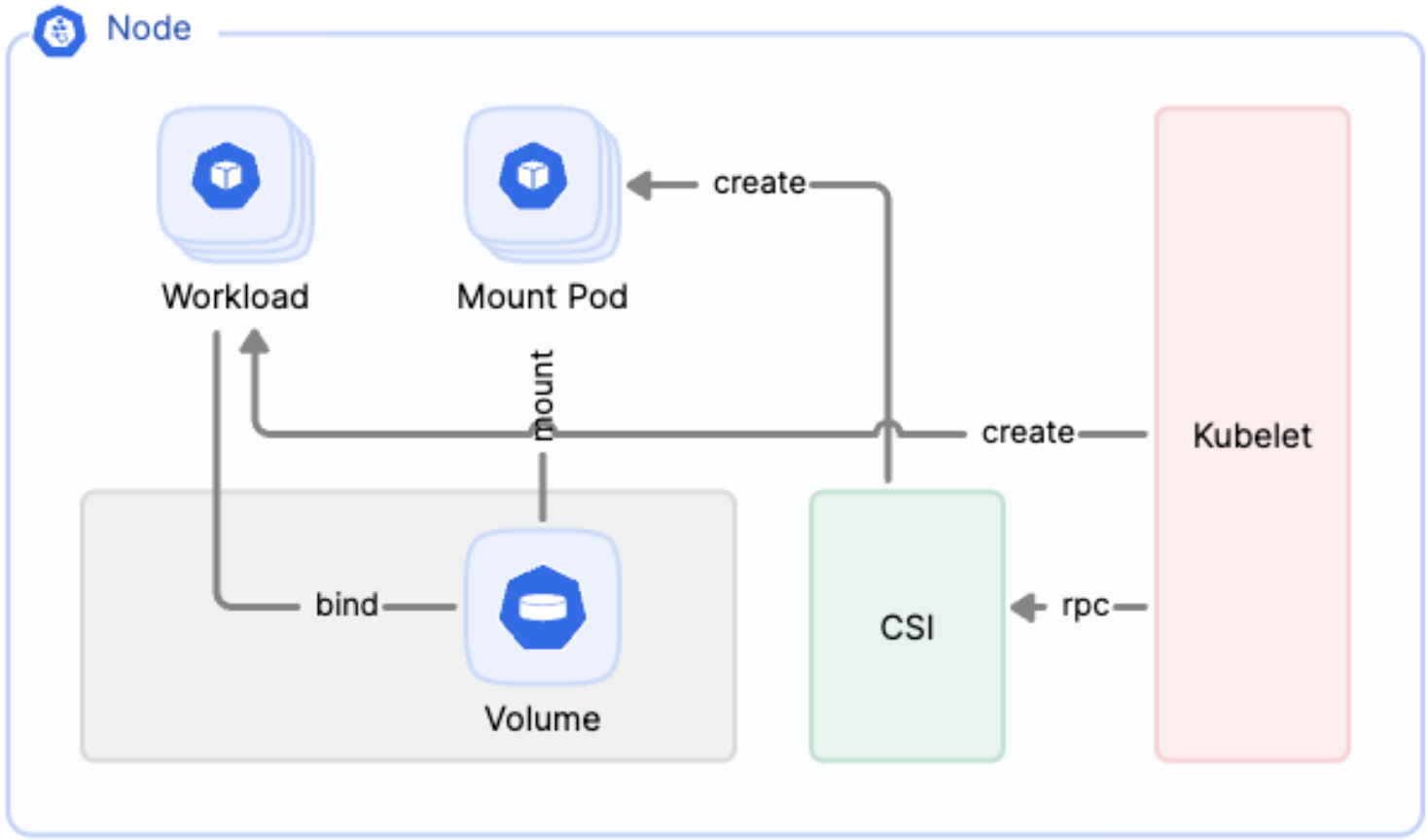
Static Provision



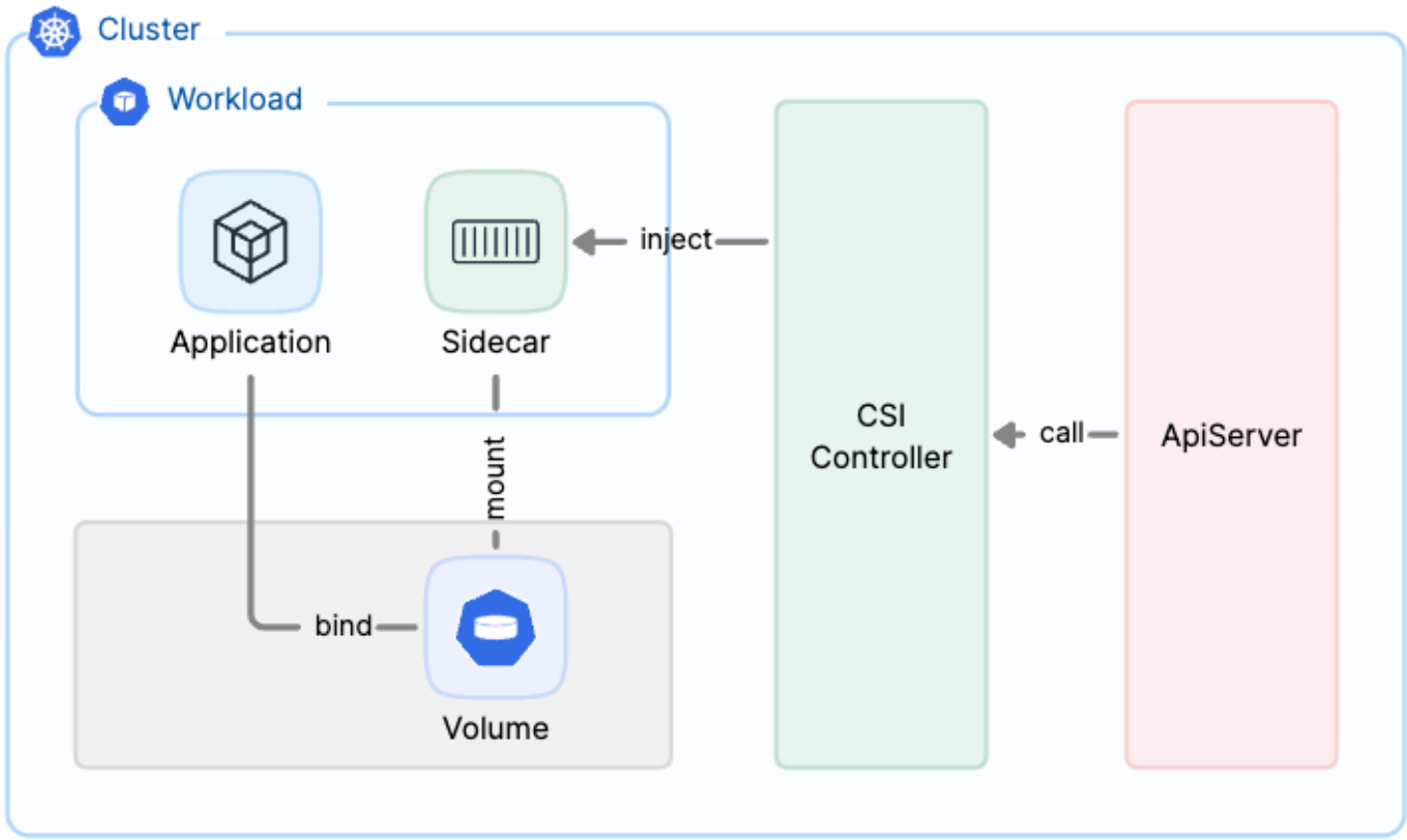
Dynamic Provision



JuiceFS in Kubernetes



JuiceFS in Serverless



Data Security in JuiceFS PV

Data isolation

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

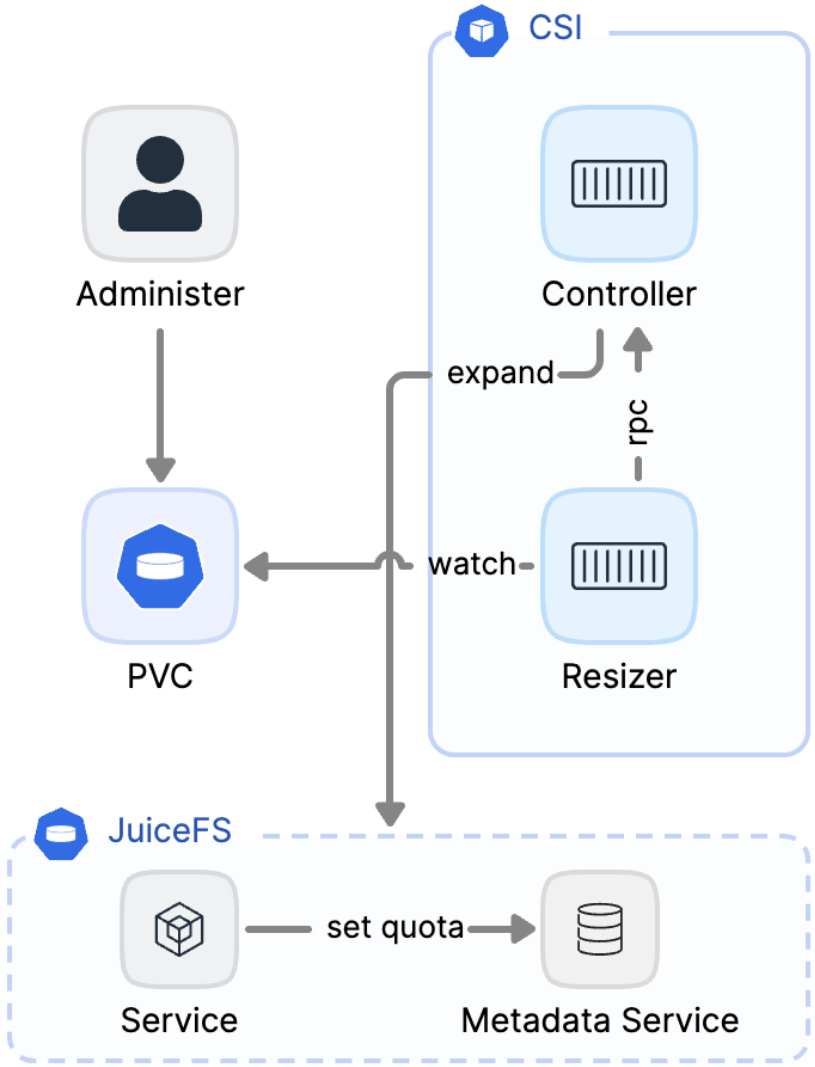
Data encryption

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

Permission control

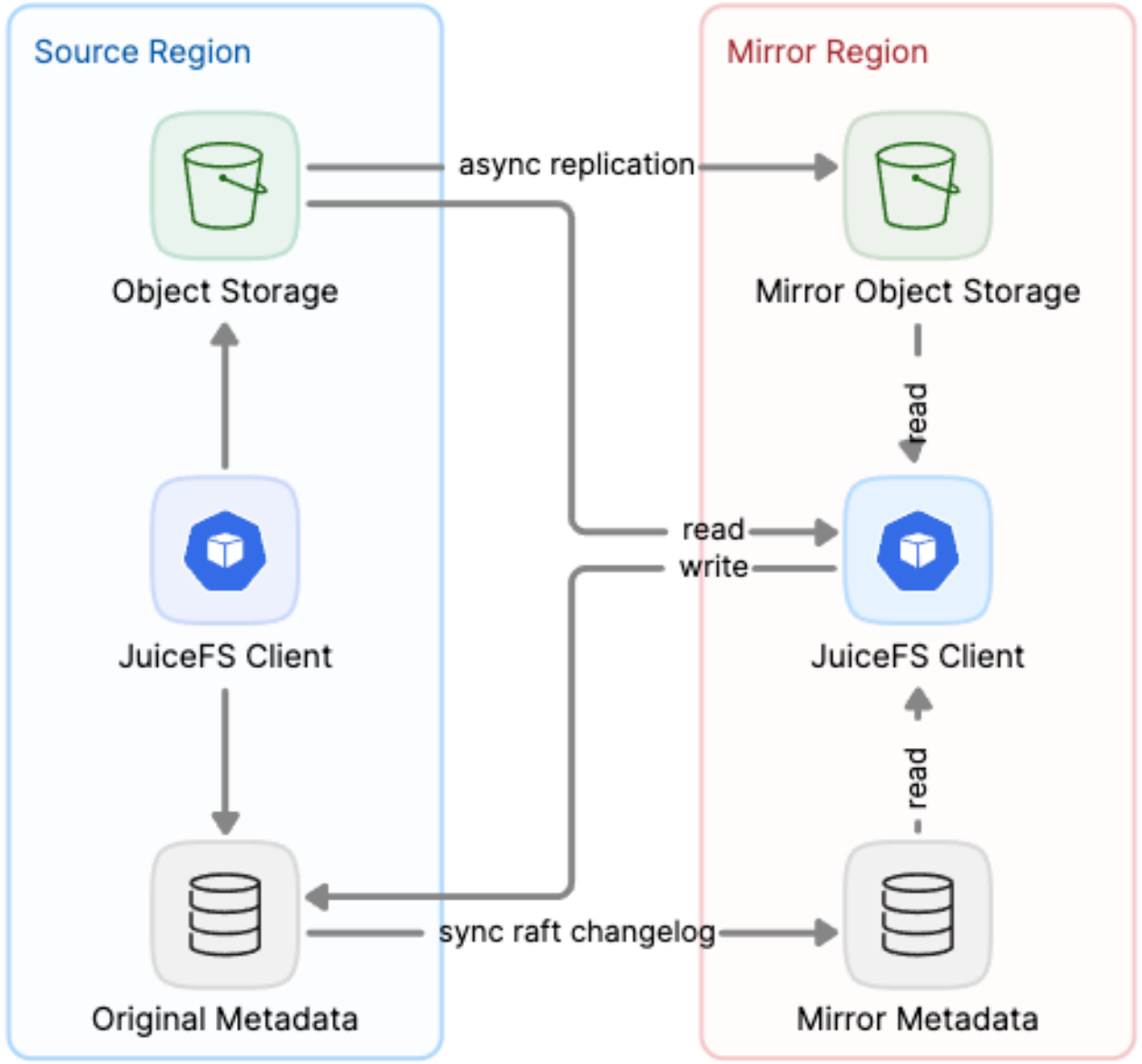
UID and GID of Unix-like systems to manage file permissions
POSIX ACL permissions

Data expansion in JuiceFS PV



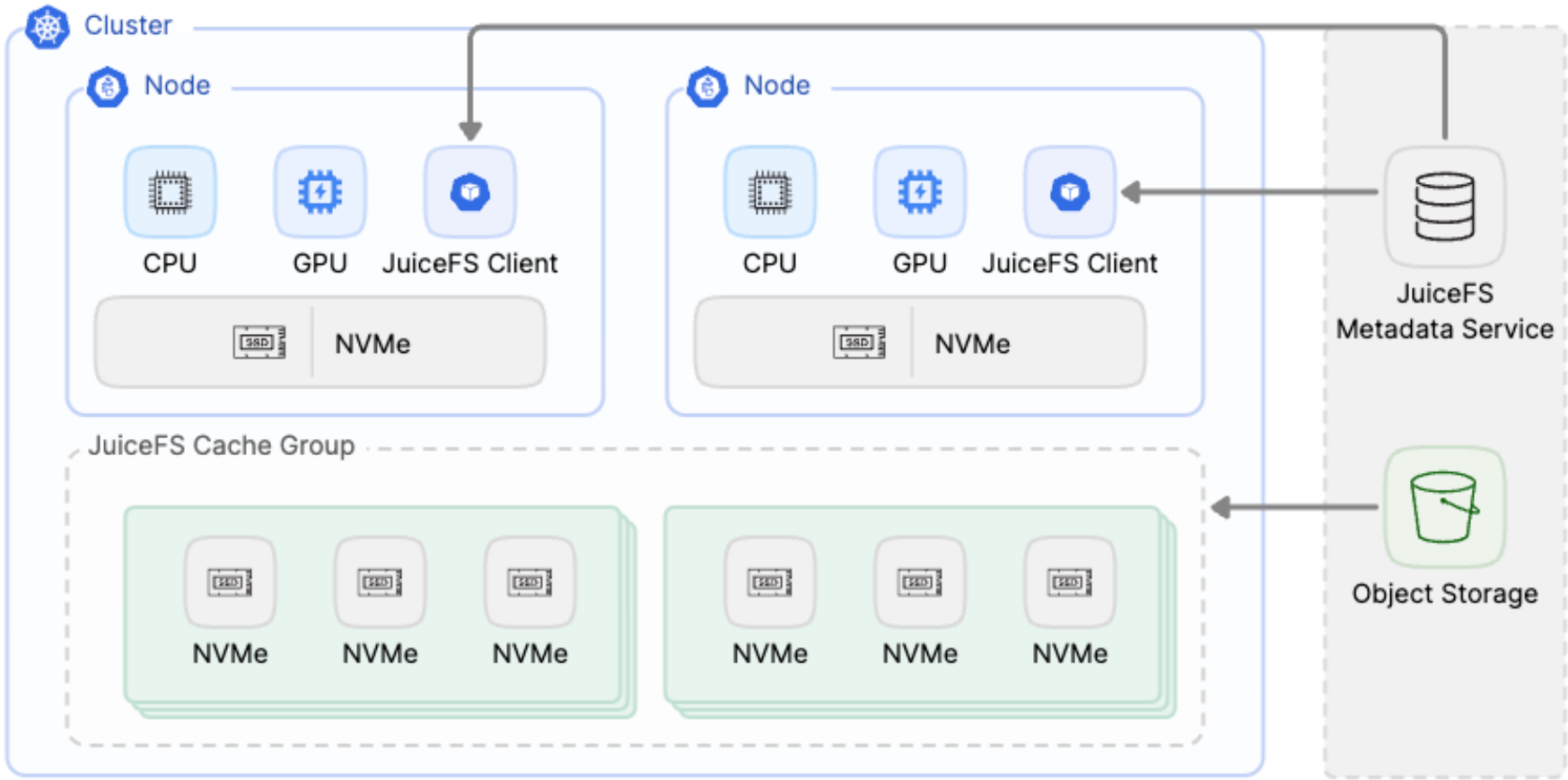

```yaml
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: juicefs-pvc
  namespace: default
spec:
  accessModes:
  - ReadWriteMany
  volumeMode: Filesystem
  storageClassName: juicefs-sc
  resources:
    requests:
      storage: 10Gi
```

Mirror file system

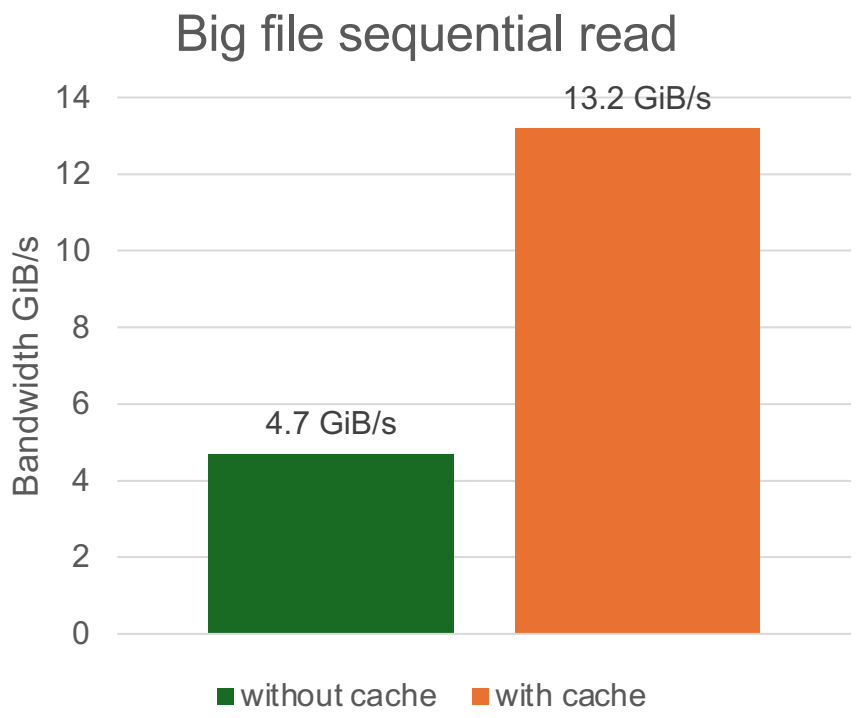


- Data in multi-cloud
- Metadata sync raft changelog to mirror
- Object Storage async to mirror

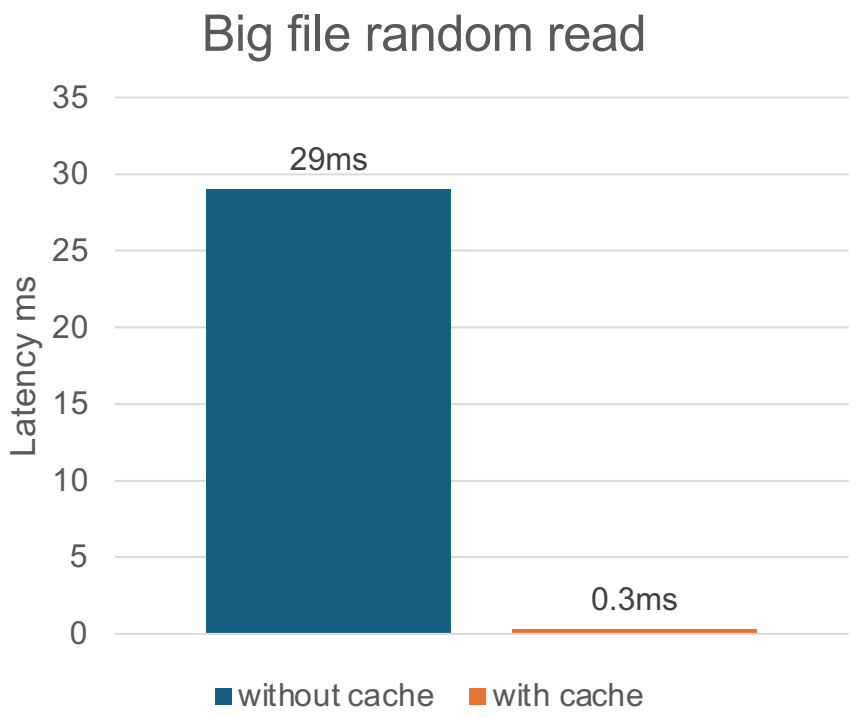
Cache Group of JuiceFS



Benchmark of Cache Group



File: 10 GB
Network speed: 200Gbps



File: 10 GB
Network speed: 200Gbps

3. Optimizations for multi-thousand node clusters

Multi-Thousand Node Clusters

NODE

Too many nodes, difficult to operate and maintain

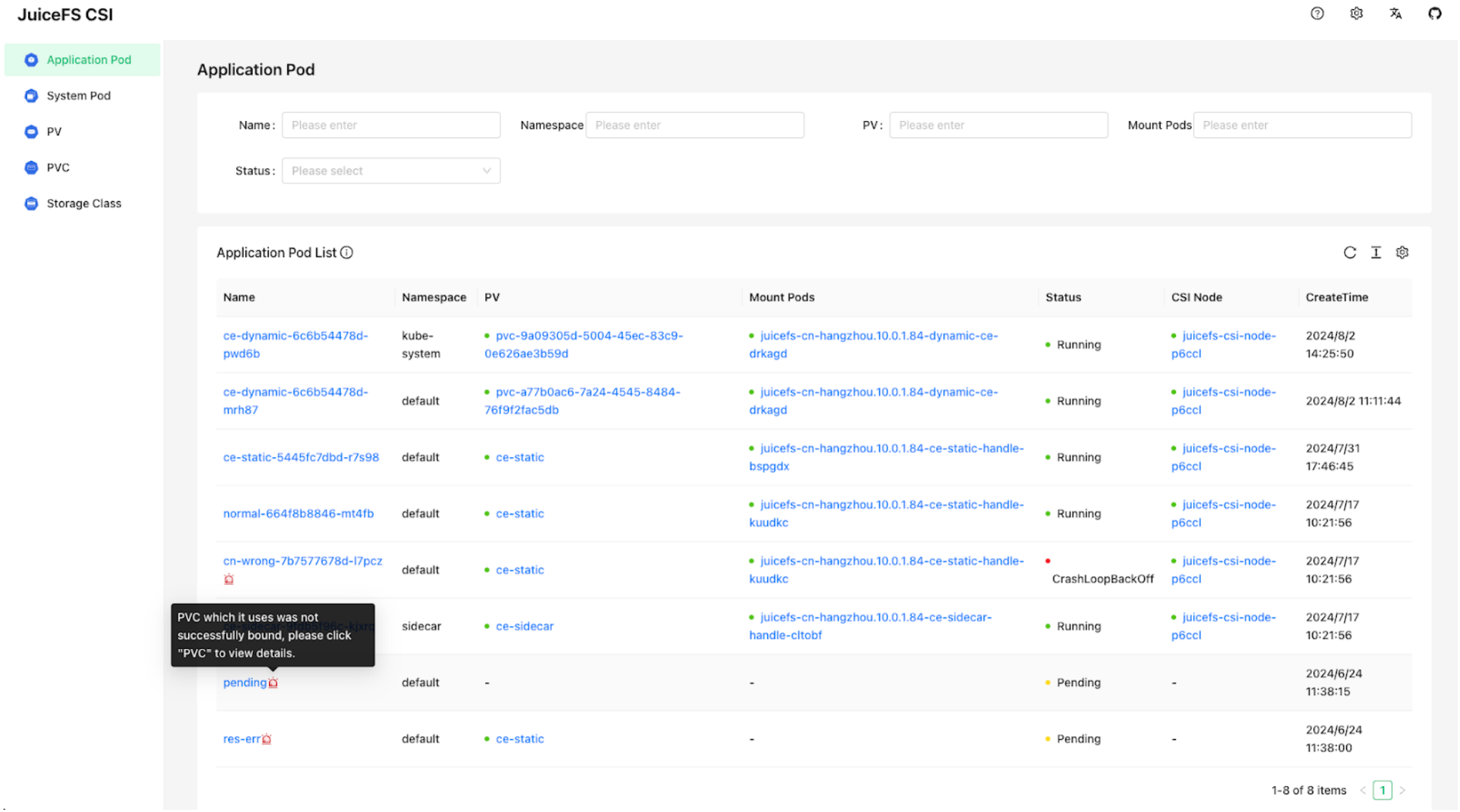
RESOURCES

Too many resources

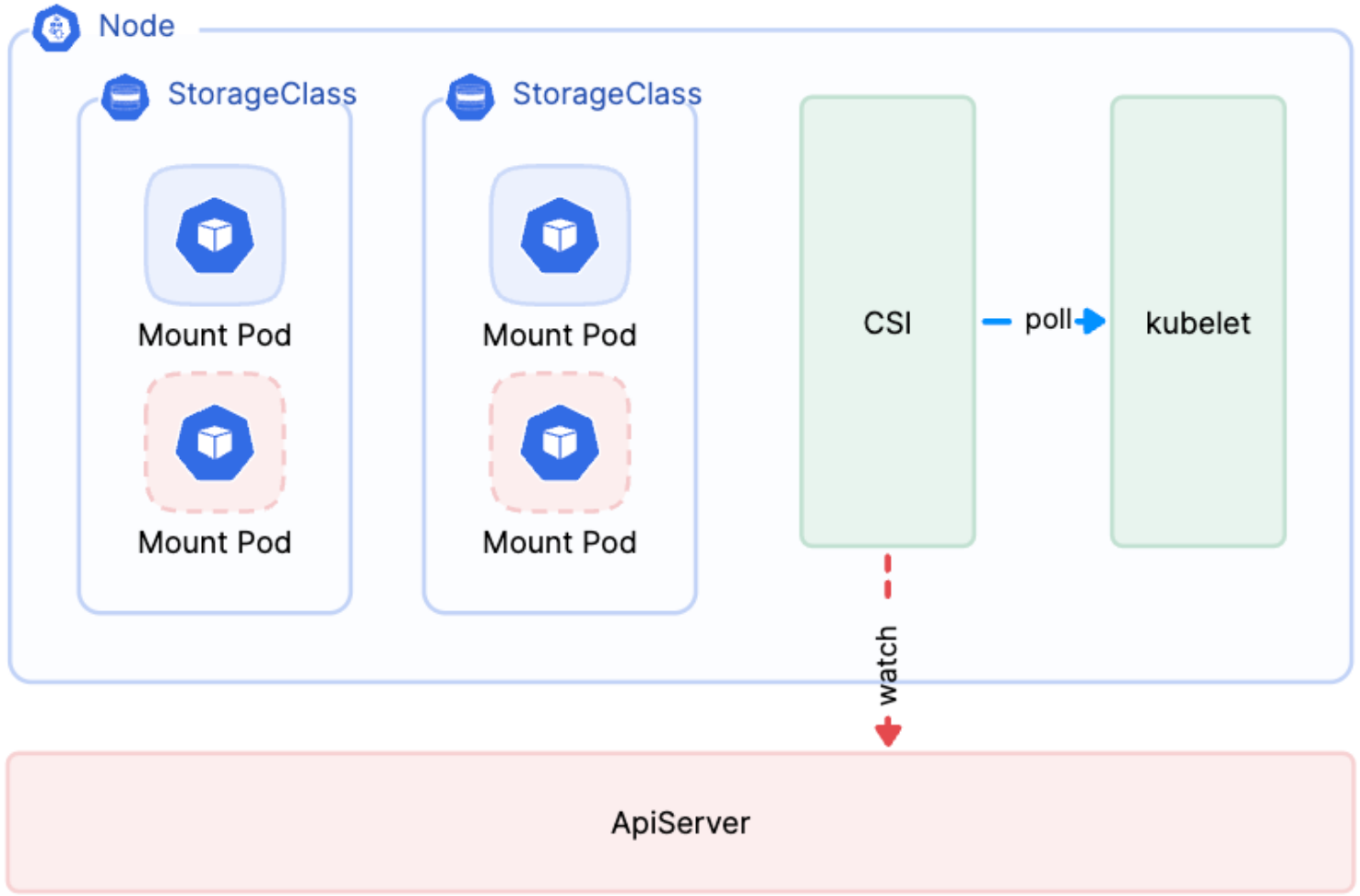
APISERVER

ApiServer is under great pressure

Visualization

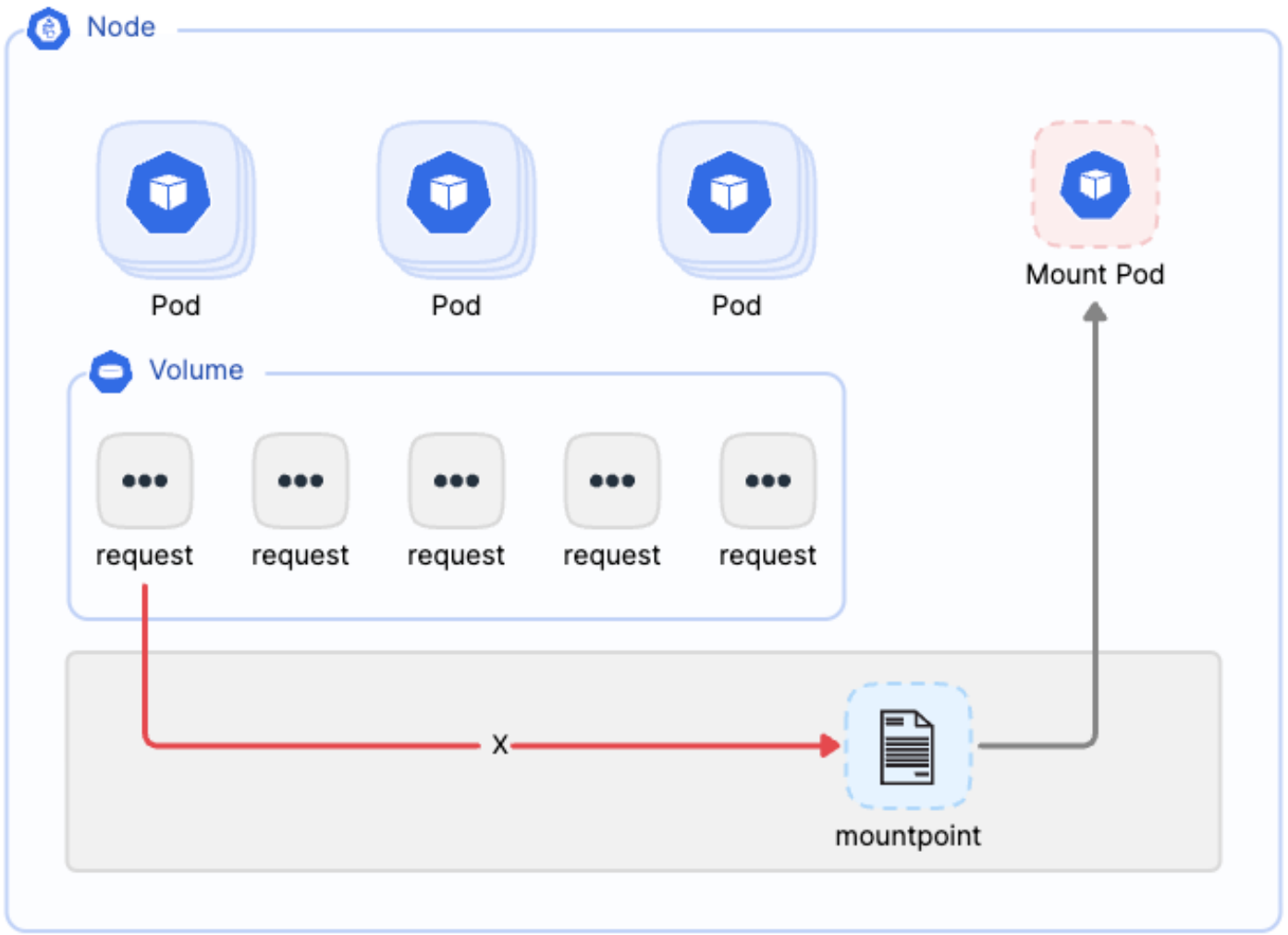


Resource and performance optimization



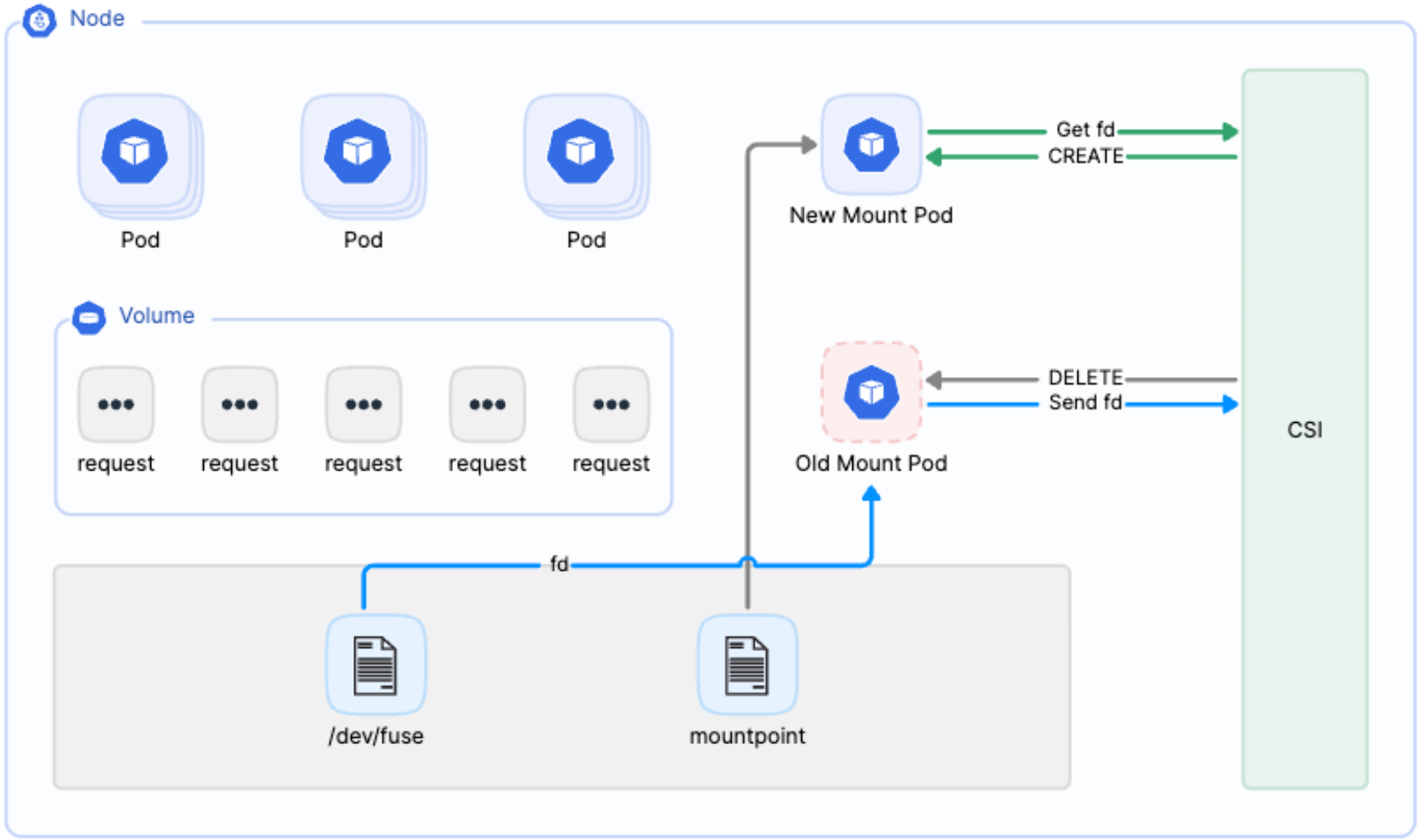
- PVC/StorageClass Shared Mount Pod
- Poll kubelet instead of watching ApiServer

Stability optimization



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

Stability optimization



Mount pod up:

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

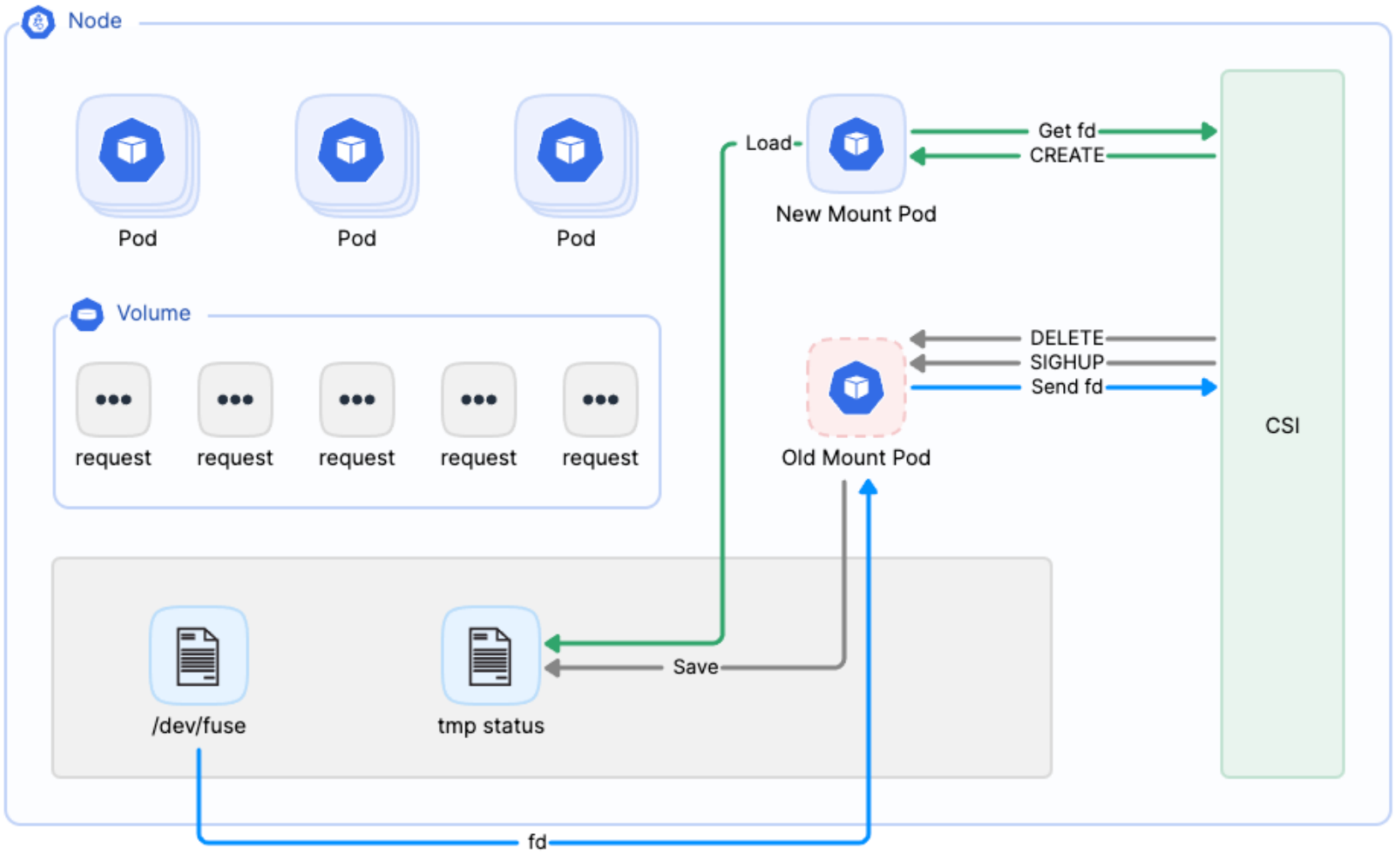
Old mount pod:

- CSI deletes it

New mount pod:

- CSI creates it
- Gets FUSE fd from CSI

Smoothly upgrade



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

4. Demo with JuiceFS

Demo with JuiceFS

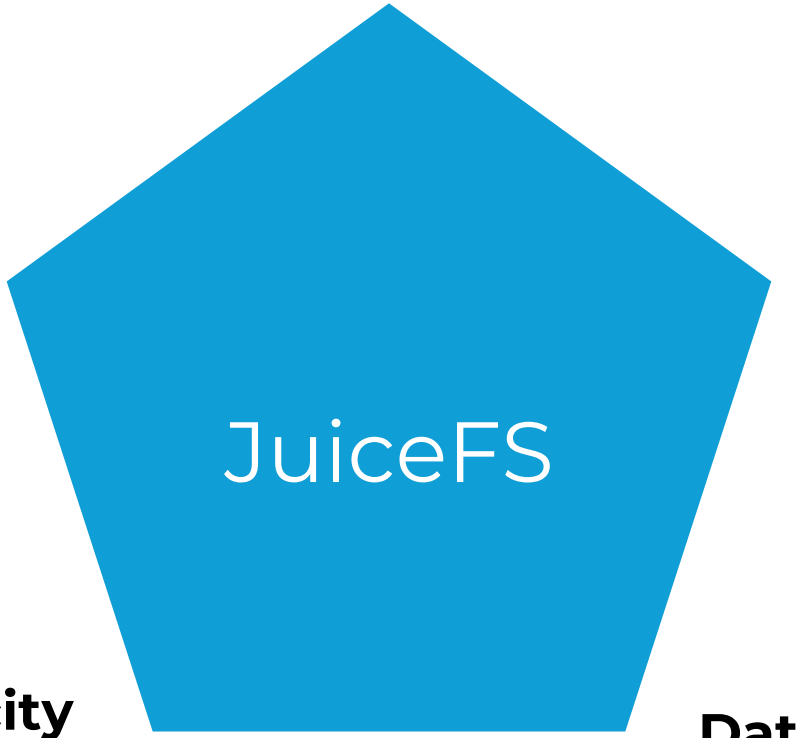
- How to use JuiceFS in Kubernetes
- Impact of JuiceFS on reading speeds for large models


root@iZbp1hizg4bntjyyiu82qlZ:~/workground/juice/deploy/ee/demo
[root@iZbp1hizg4bntjyyiu82qlZ demo]#

Conclusion

POSIX compliance

High performance & Cost control
Distributed cache group



High data security
Isolation / encryption / Permission control

Data elasticity
Sidecar in Serverless / Data expansion

Data consistency
Mirror file system

Thanks

https://juicefs.com

https://github.com/juicedata/juicefs