JuiceFS 在 Kubernetes 环境中数千节点数据集的应用实践

苏锐 - Juicedata 合伙人

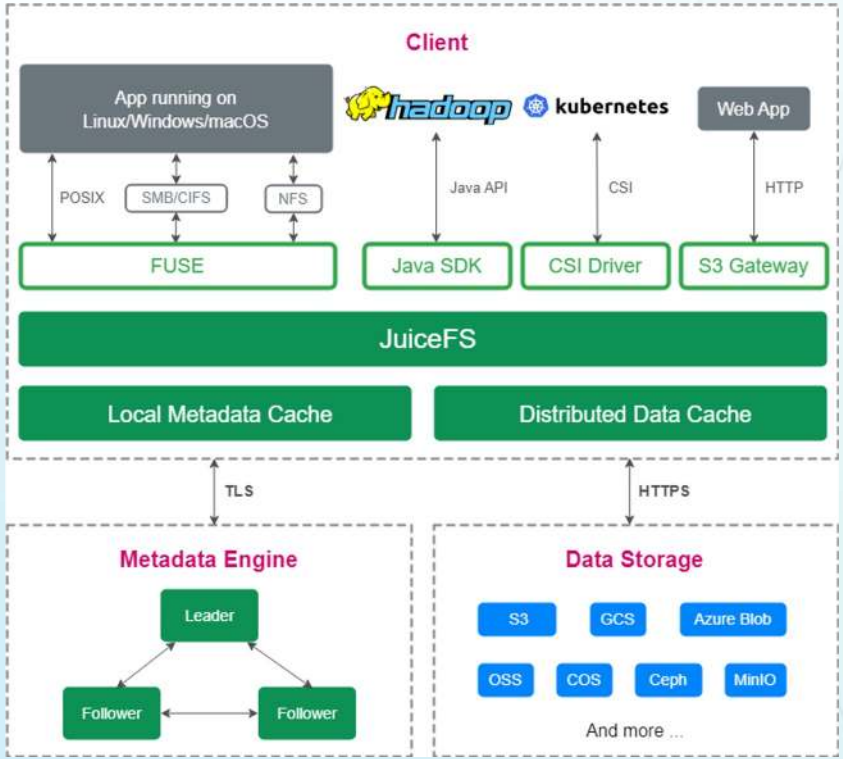
分享大纲

- JuiceFS 是什么
- JuiceFS 在 Kubernetes 上的几种使用姿势
- 数据 在 AI + Kubernetes 中遇到的挑战
- 提升 JuiceFS 在大型 Kubernetes 集群中的体验

苏锐

- 2017 年作为联创开始 JuiceFS 的创业之旅
- 18 年 IT 工作，做过 Tech Lead、PM、CEO
- 西电 三系 校友

JuiceFS 是什么？



- 2017 年发布云服务；
- 支持（几乎）所有全球公有云；
- 生产最大规模单卷近千亿文件，百 PB 容量，聚合吞吐数百 GBps；
- 100% POSIX 兼容。

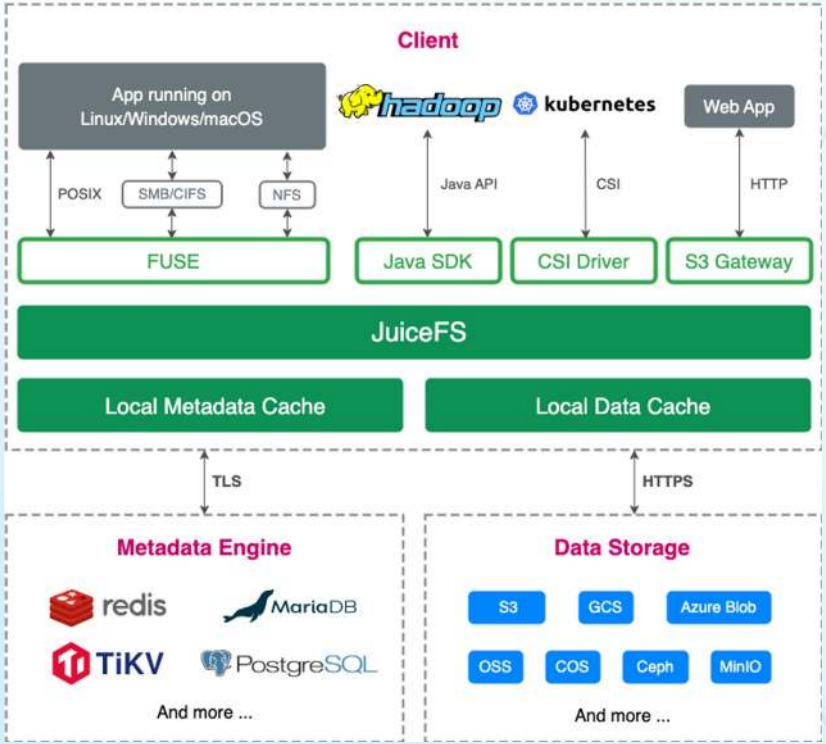
MINIMAX 阶跃星辰 小红书

智谱·AI 面壁智能 知乎

WPS LibLib AI Lepton AI

momenta DP Technology 深势科技

JuiceFS 是什么？



- 2021 年发布；
- GitHub 11.3K ；
- 胖客户端模式，简单上手，简单运维；
- Golang 开发 + CSI 完善，得到云原生开发者支持；
- 使用最多：
  - AI 平台
  - Kubernetes PV
  - 大数据存算分离

JuiceFS 在 Kubernetes 上的几种使用姿势

hostPath

CSI - MountPod

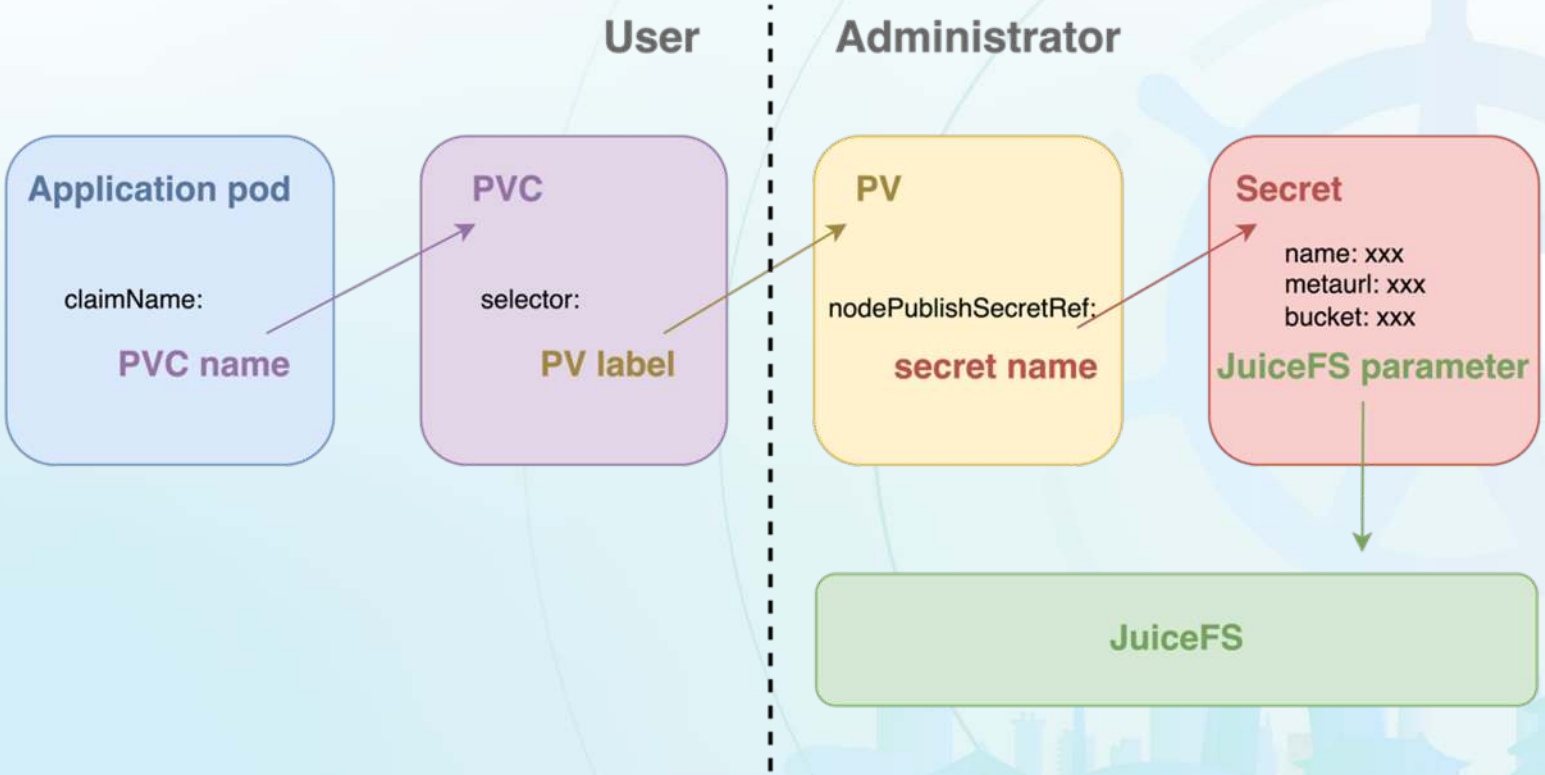
CSI - Sidecar

hostPath

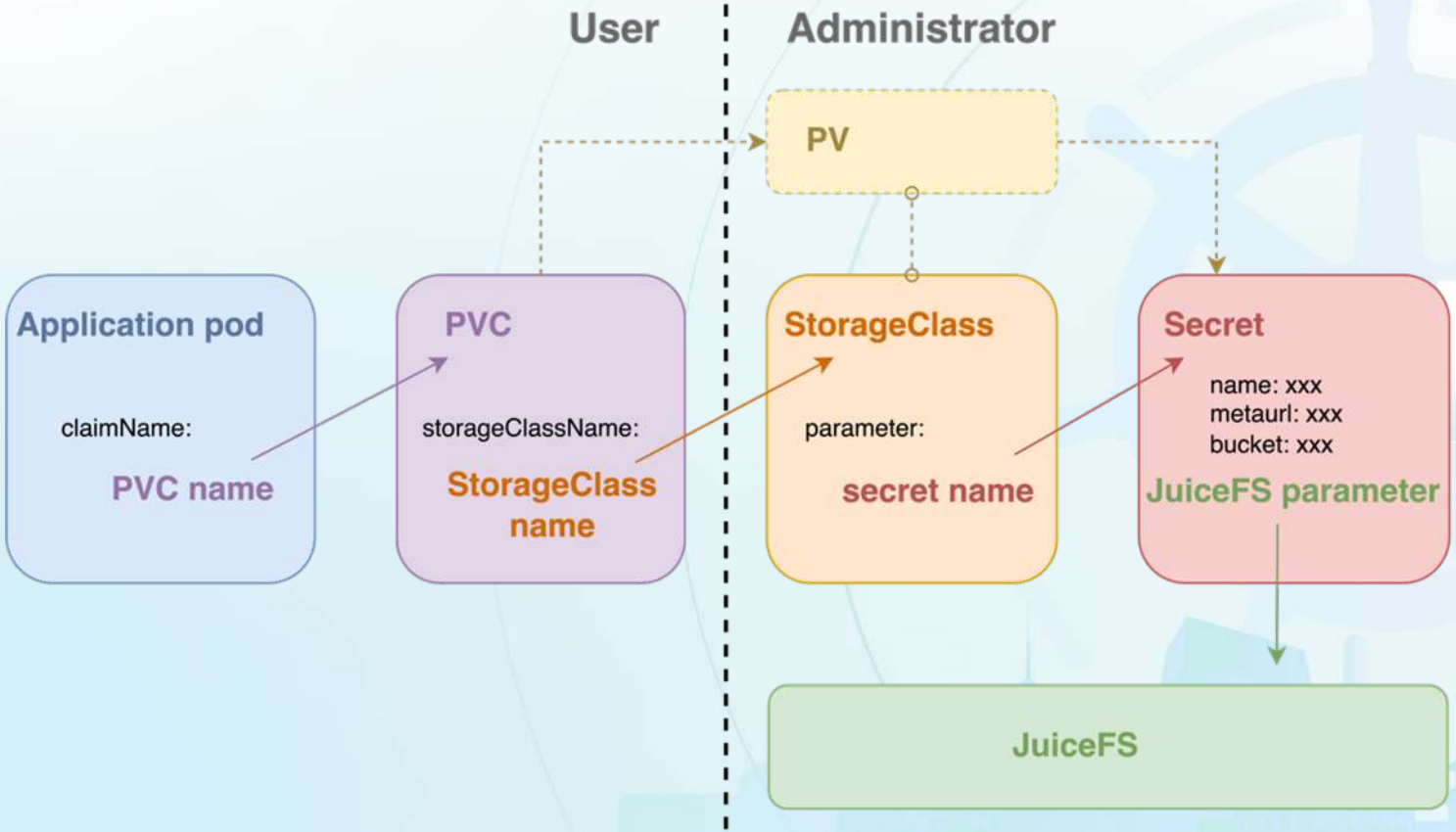
- 将 JuiceFS 挂载到所有宿主机的同一个目录，比如 /mnt/data;
- 挂载参数在宿主机管理；
- Pod 声明 hostPath，指定挂载路径 /mnt/data；
- 灵活性不足；
- 挂载点故障后无法无感恢复。

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: hostpath-pod
spec:
  containers:
  - name: test-container
    image: busybox
    volumeMounts:
    - mountPath: /data
      name: hostpath-volume
  volumes:
  - name: hostpath-volume
    hostPath:
      path: /mnt/data
      type: Directory
```

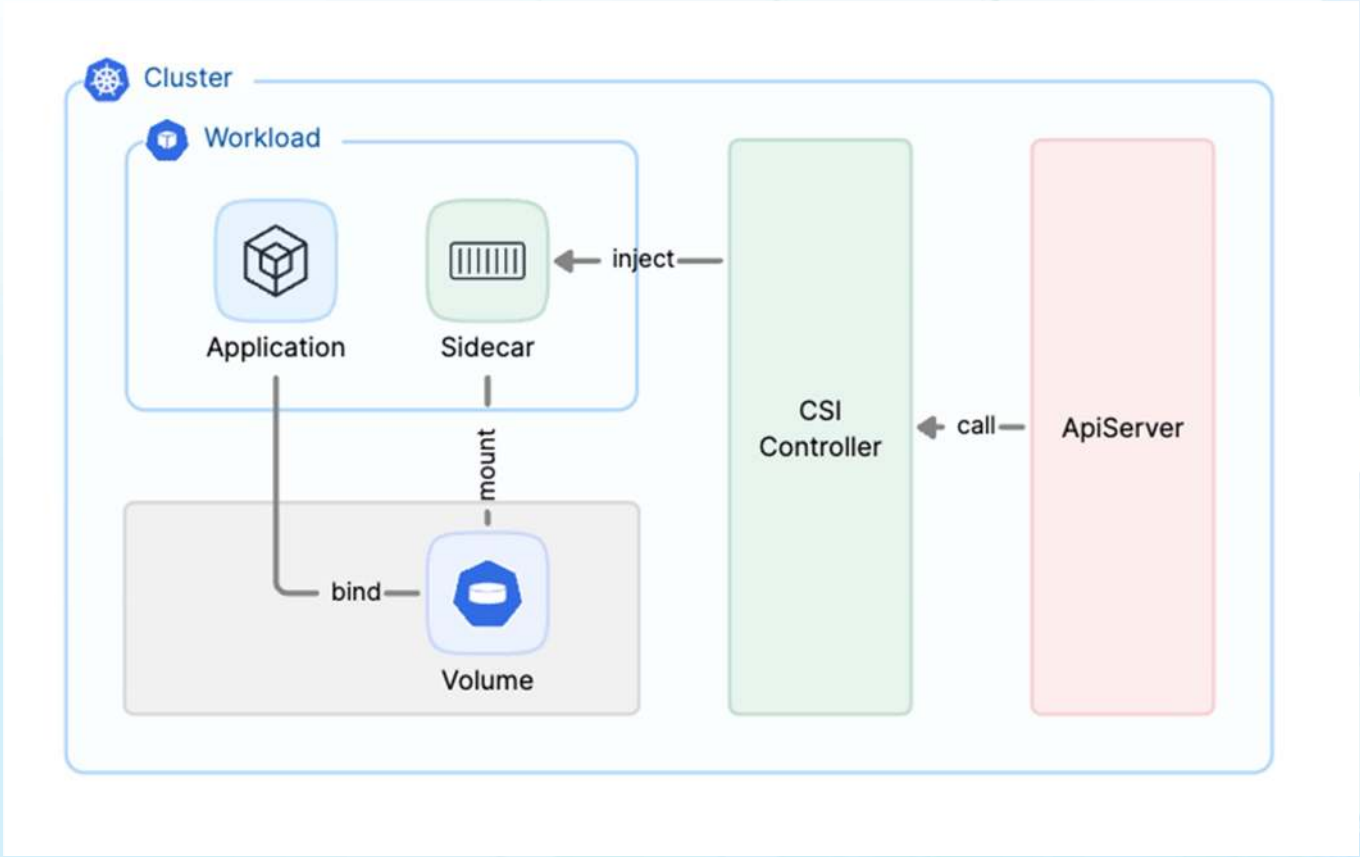
CSI - MountPod - Static Provision



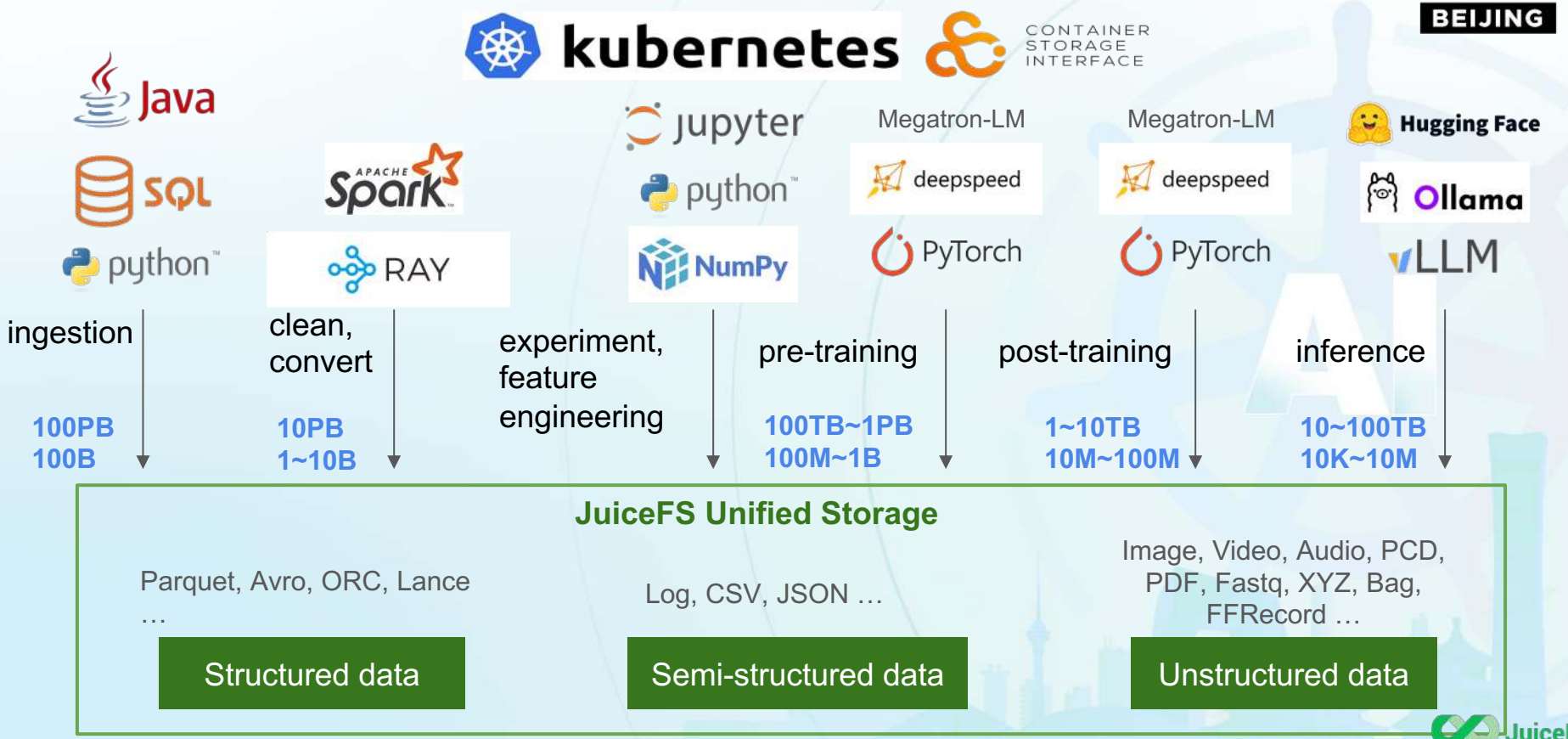
CSI - MountPod - Dynamic Provision



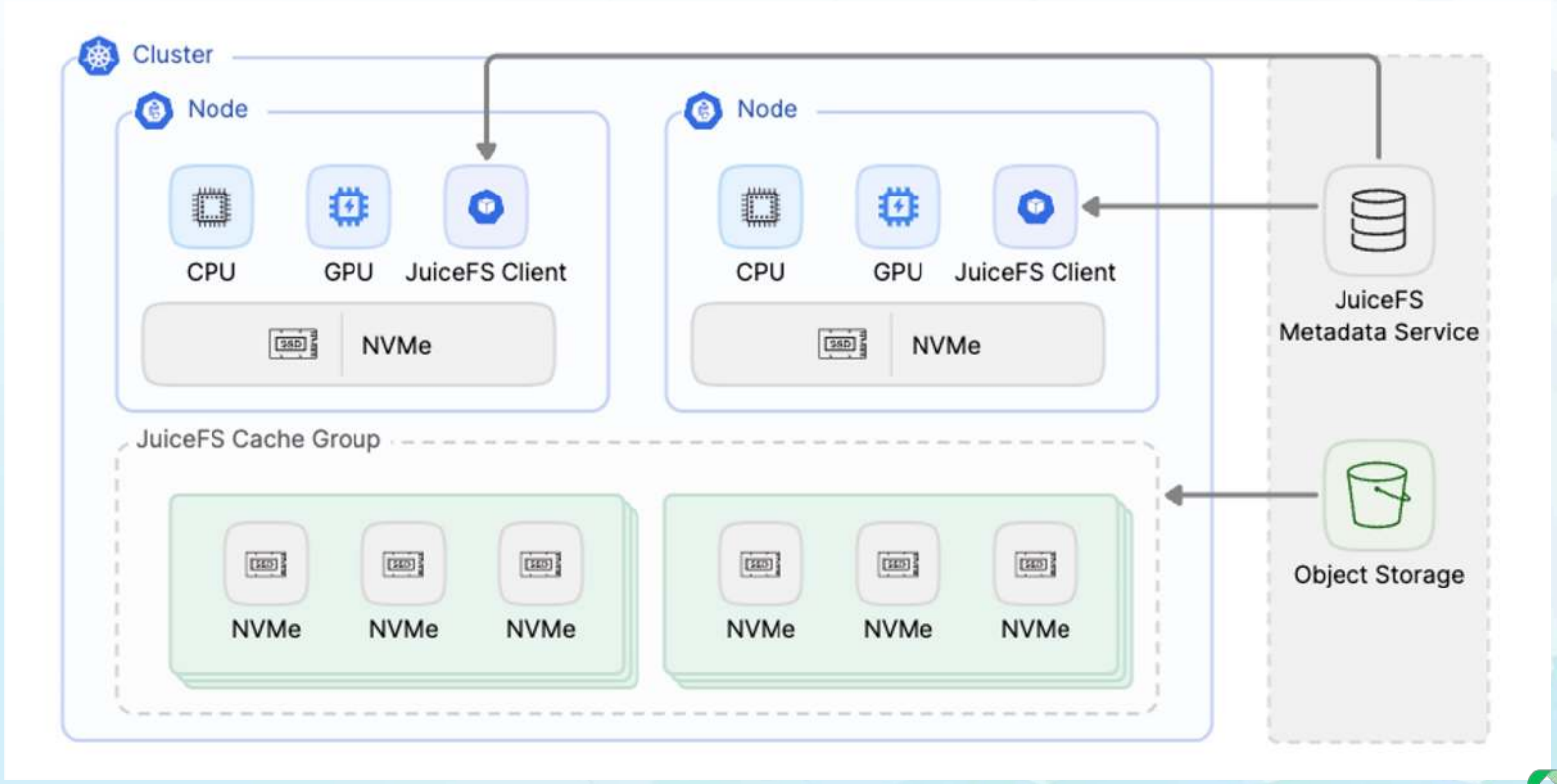
CSI - Sidecar


Cluster
Workload
Application
Sidecar
inject
CSI
Controller
call
ApiServer
mount
bind
Volume

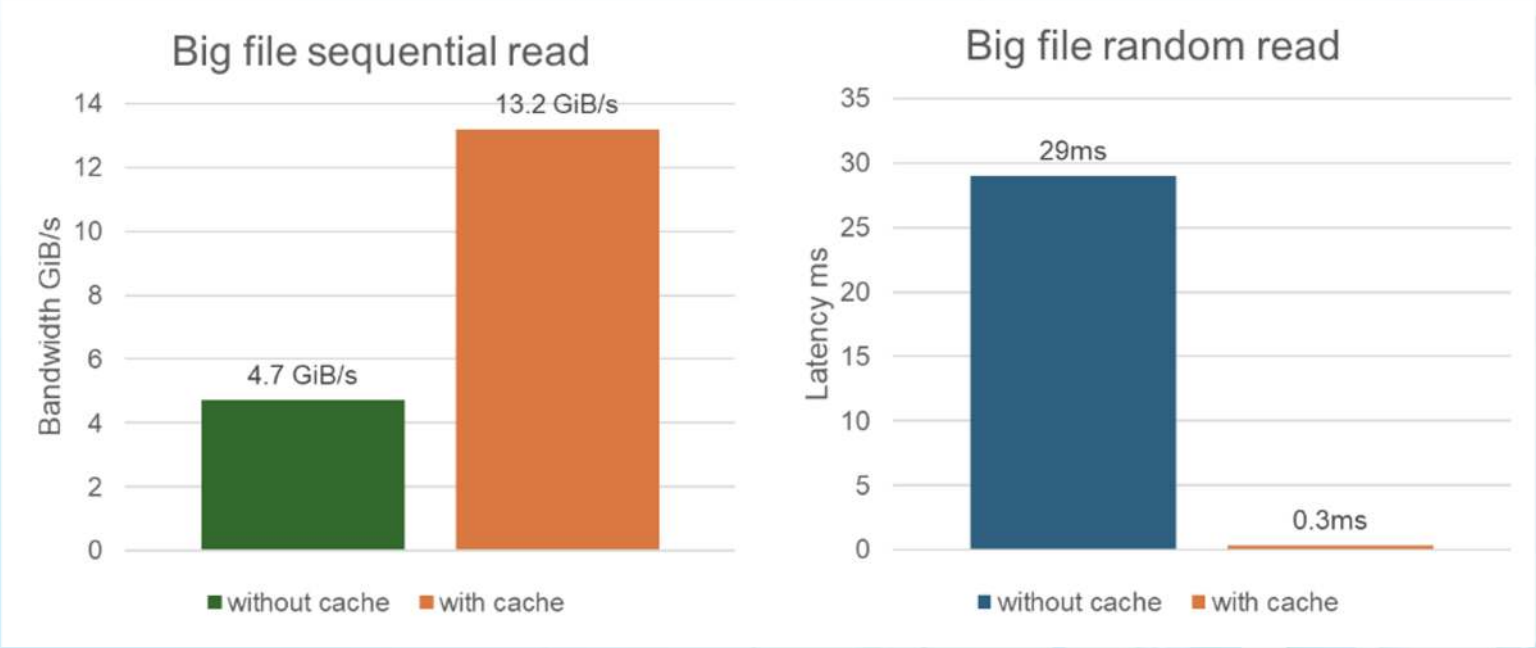
数据在 AI + Kubernetes 中的挑战



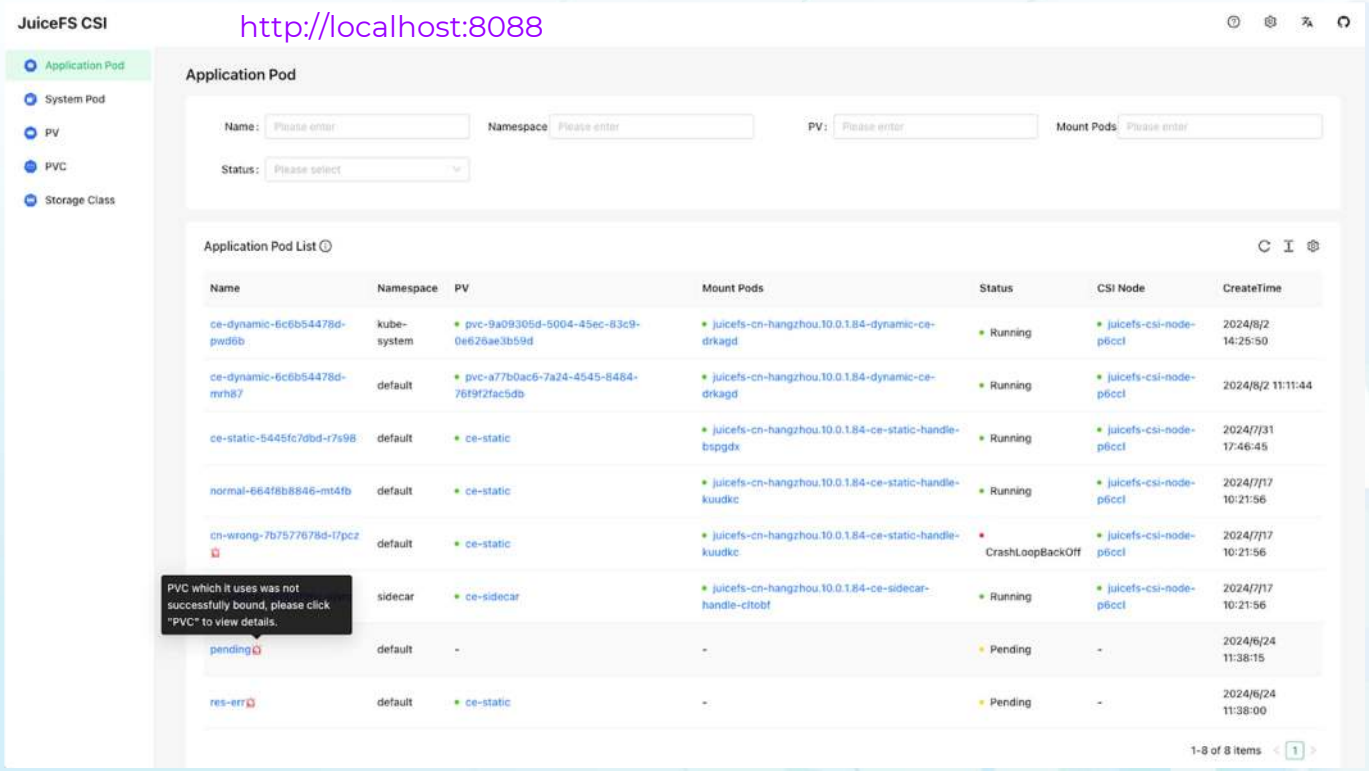
大家都关心的：性能


Cluster
Node
CPU
GPU
JuiceFS Client
NVMe
Node
CPU
GPU
JuiceFS Client
NVMe
JuiceFS Cache Group
NVMe
NVMe
NVMe
NVMe
NVMe
NVMe
JuiceFS
Metadata Service
Object Storage

大家都关心的：性能


Big file sequential read
Bandwidth GiB/s
14
12
10
8
6
4
2
0
4.7 GiB/s
13.2 GiB/s
without cache
with cache
Big file random read
Latency ms
35
30
25
20
15
10
5
0
29ms
0.3ms
without cache
with cache

大家容易忽视的：图形化可观测 - CSI Dashboard



$ helm install juicefs-csi-driver juicefs/juicefs-csi-driver

大家容易忽视的：缓存组配置简单灵活了

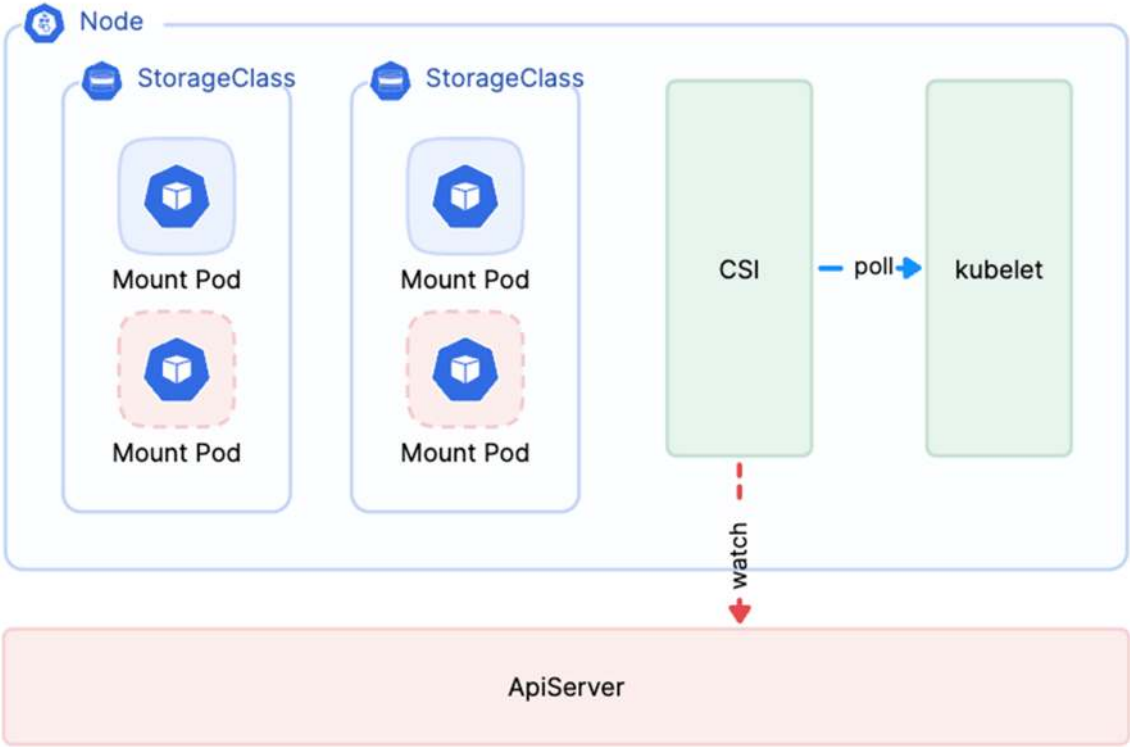
之前，我们要通过 StatefulSet 或 DaemonSet 的方式创建缓存组，但存在以下问题：

- 无法在同一集群内针对不同节点类型或资源（如挂载参数、缓存组权重等）进行单独配置；
- 需要依赖人工监控并手动添加或移除节点，操作繁琐，容易出错；
- 缓存清理需手动执行，不能自动化。

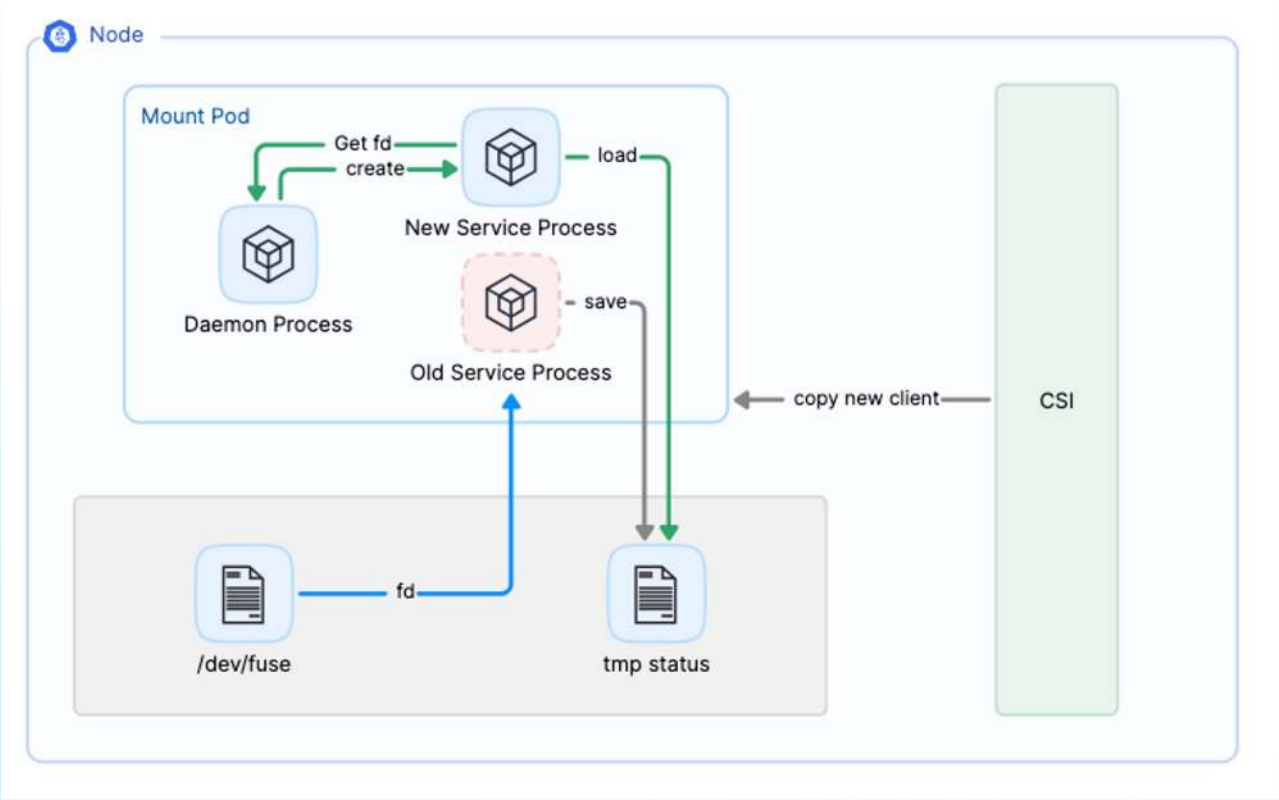
JuiceFS 用 Cache Group Operator 改善了上面的问题：

- 在同一集群中配置不同的节点类型和资源;
- 支持平滑添加或移除节点，尽可能减小加减节点期间缓存命中率波动;
- 缓存自动清理;
- Dashboard 中可以管理缓存组。

大家容易忽视的：给 ApiServer 减负

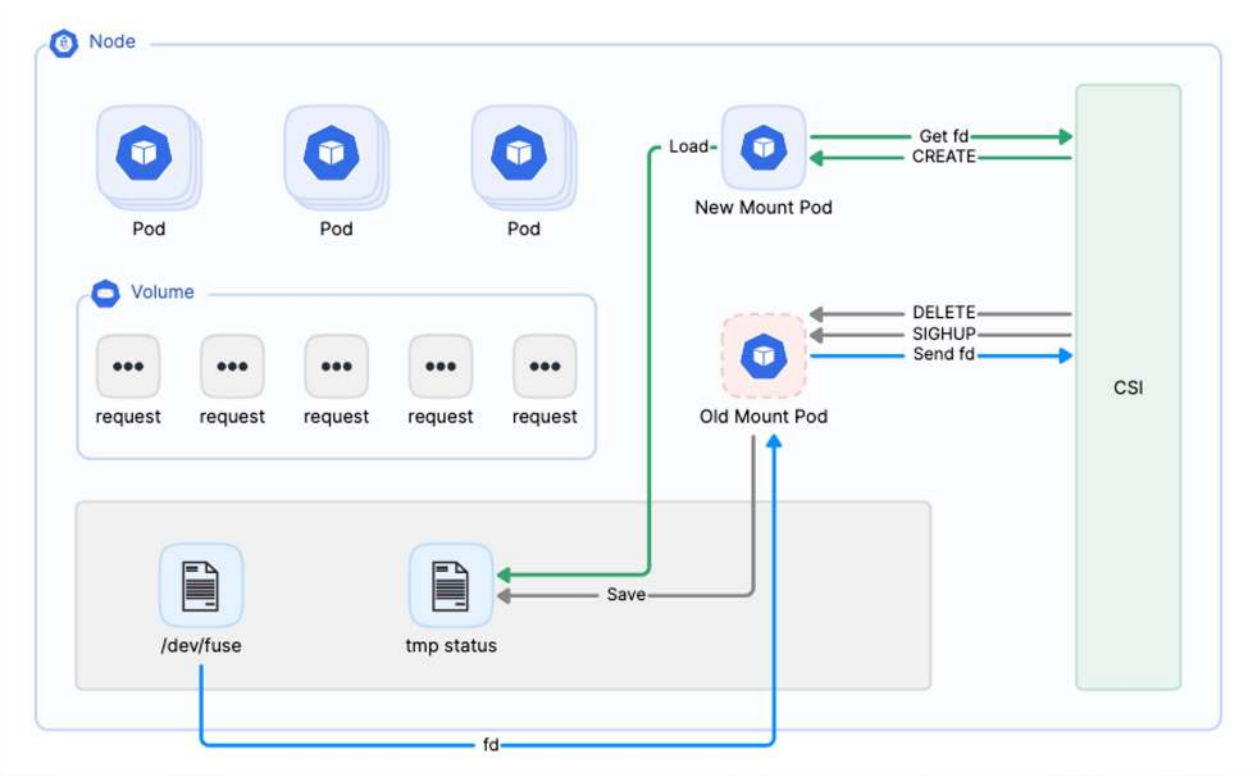


大家容易忽视的：JuiceFS 平滑升级，业务不中断



JuiceFS 客户端二进制 平滑升级

大家容易忽视的：JuiceFS 平滑升级，业务不中断



JuiceFS MountPod 平滑升级

KUBERNETES COMMUNITY DAYS BEIJING 2025

谢谢大家 ❤️

JuiceFS