云原生社区Meetup

第 二 期 · 北 京 站

云原生社区 Cloud Native Community

云原生分布式存储解决方案实践

—— ChubaoFS发展历程回顾

演讲人：刘硕然 OPPO

目录

发展大事件

- Project launch - January 2017
- First application in production at JD.com – June 2018
- Open source from the first commit – March 2019
- Technical presentation to CNCN Storage SIG – June 12 2019
- First External User – Reconova – June 2019
- SIGMOD '19 Presentation (Industrial Paper) – July 4 2019
- Proposed to CNCF Sandbox – August 27 2019
- Enrolled as a CNCF Sandbox project - Dec 2019
- Released S3 compatible interface - Apr 2020
- OPPO joined as a key development organization - July 2020

核心诉求

为大规模容器平台的应用
提供存储服务

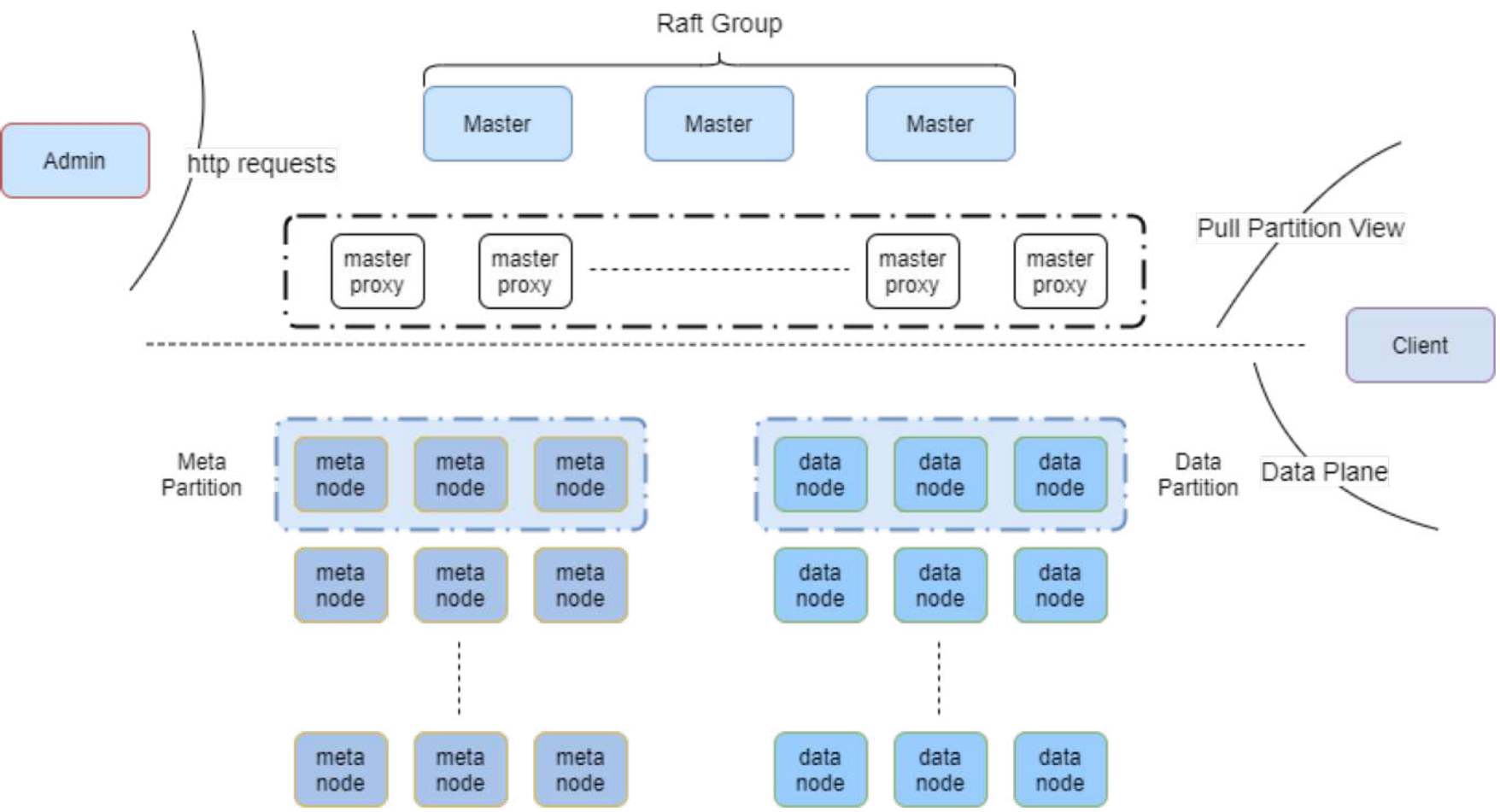
遇到挑战

- 业务方众多
- 弹性扩容需求强烈
- 由本地存储迁移而来
  - 性能
  - POSIX语义遵守
- 文件大小类型复杂
  - 小文件性能及容量横向扩展
- 读写模型复杂
  - 顺序/随机

设计目标

- 服务器规模1k+，客户端实例10k+
- 多租户共享一套集群
- 一键弹性扩容，前期无需预估使用量
- 通过共享提高资源使用率
- 良好的元数据水平扩展性
- 应对各种文件大小类型
- 应对流量洪峰

架构设计



如何减少瓶颈点

- 元数据横向扩展
- 数据面和控制面分离
- 潜在瓶颈点增加代理

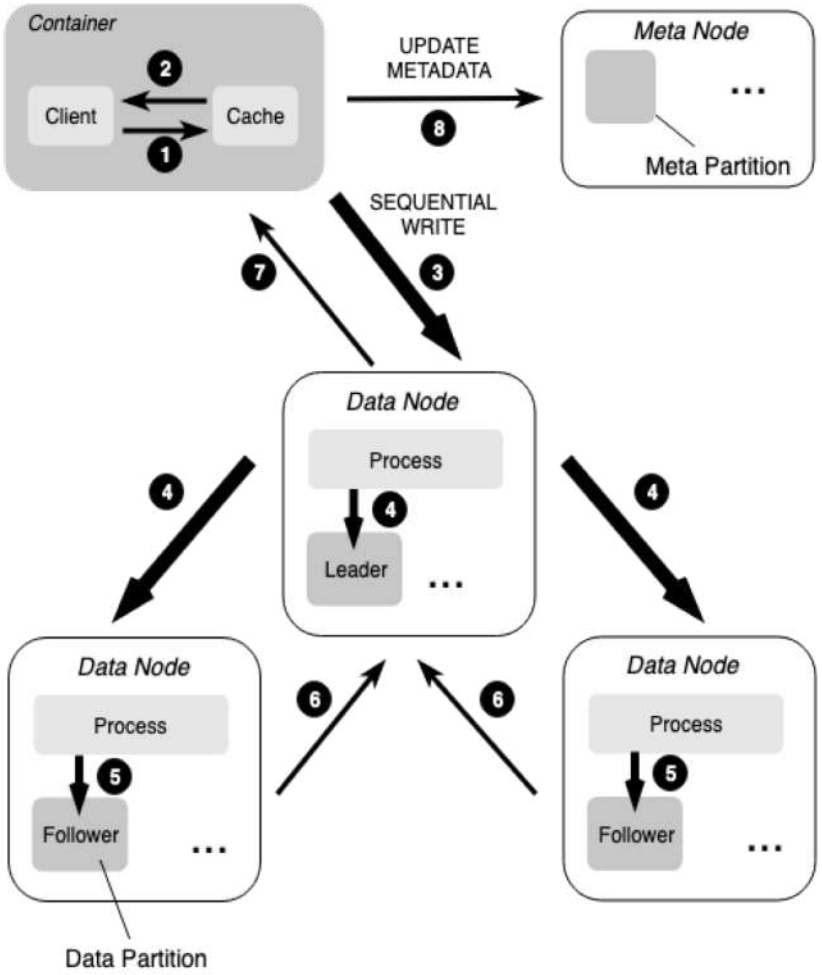
如何提高资源利用率

- 逻辑partition复制组
- 单个partition限制最高物理资源使用量
- Volume由逻辑partition组成，未使用不消耗物理资源

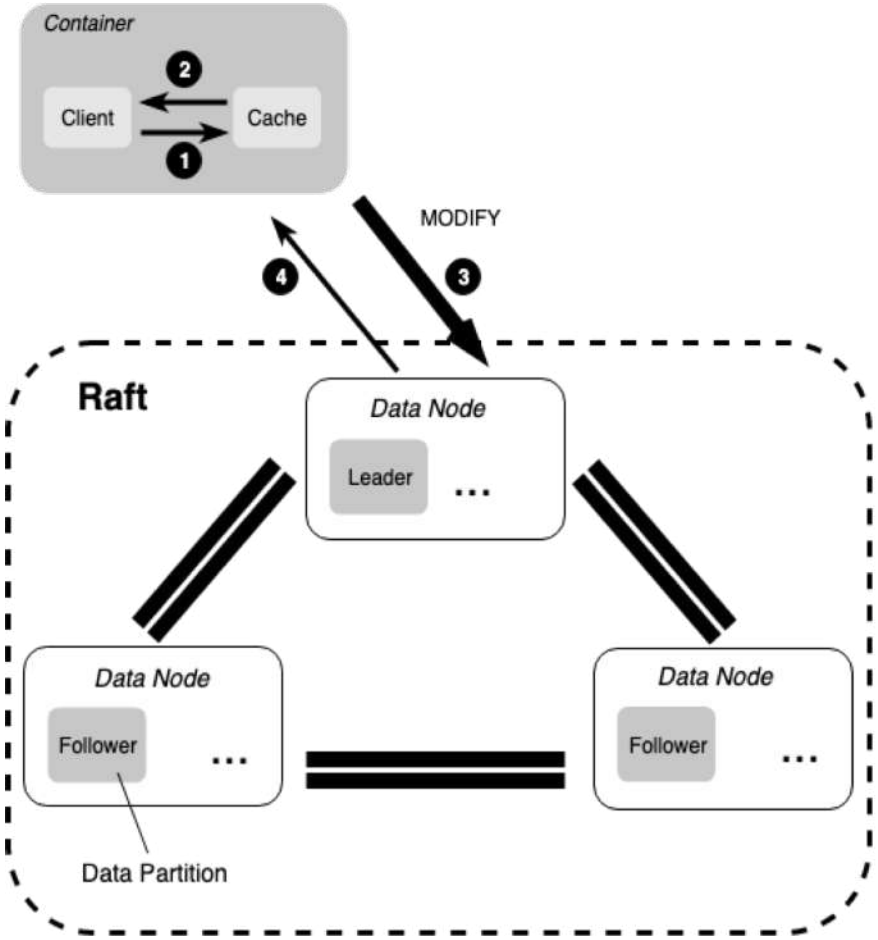
如何应对流量洪峰

- 预分配逻辑partition

架构设计 - 顺序写 & 随机写

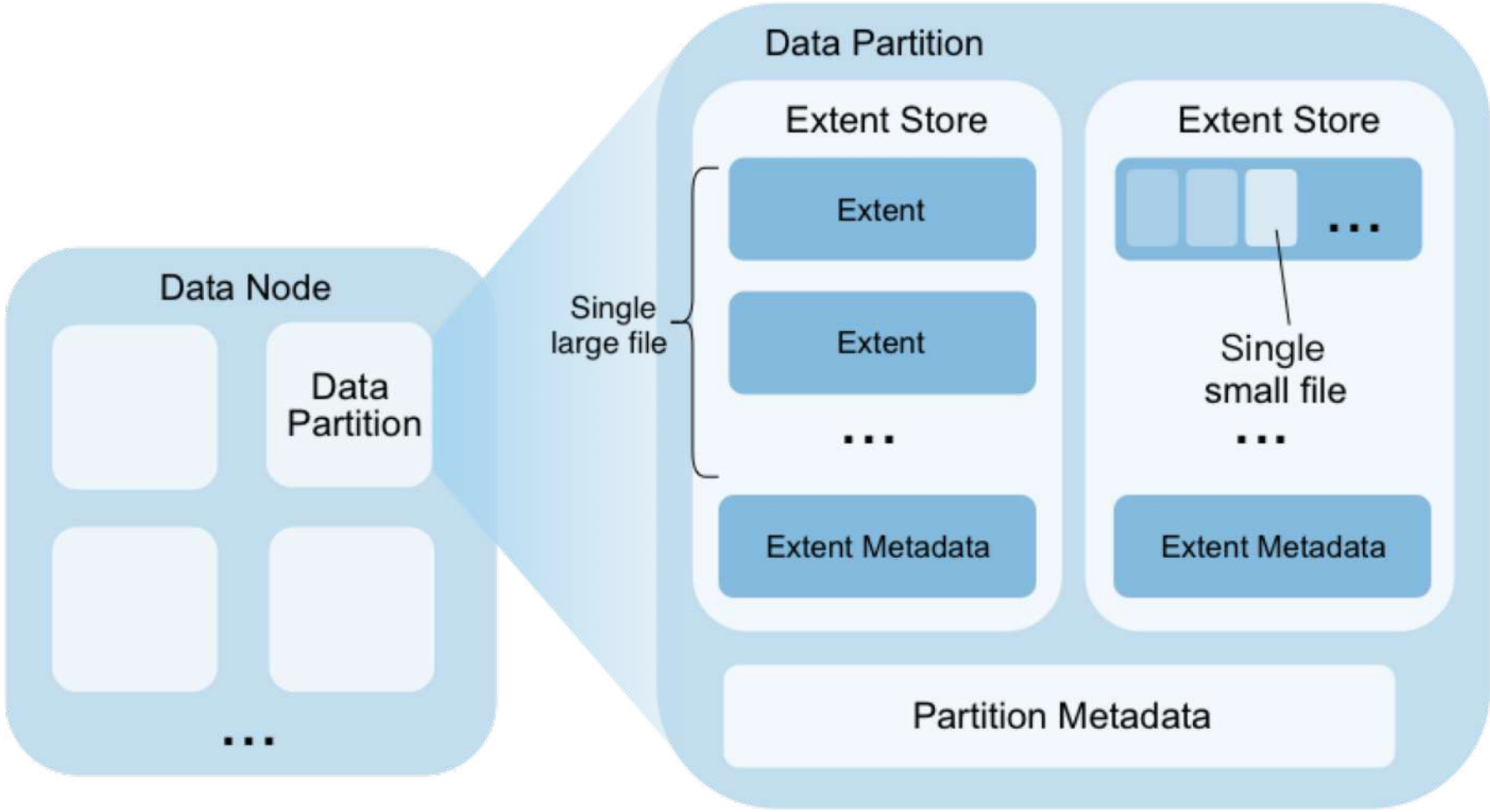


顺序写

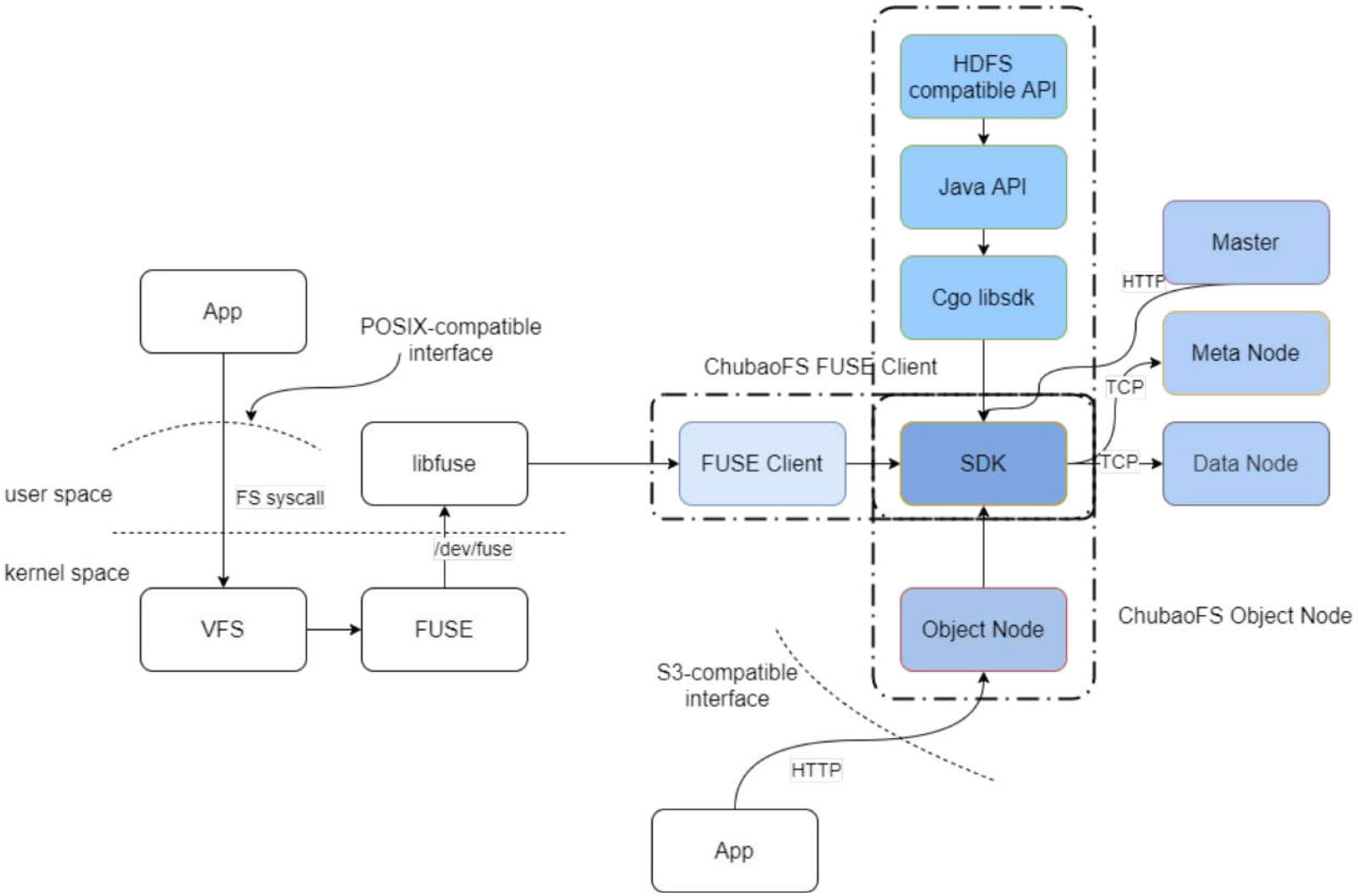


随机写

架构设计 - 小文件聚合


Data Node
Data Partition
...
Data Partition
Extent Store
Extent Store
Single large file
Extent
Extent
...
Extent Metadata
...
Single small file
...
Extent Metadata
Partition Metadata

架构设计 - 文件对象融合存储



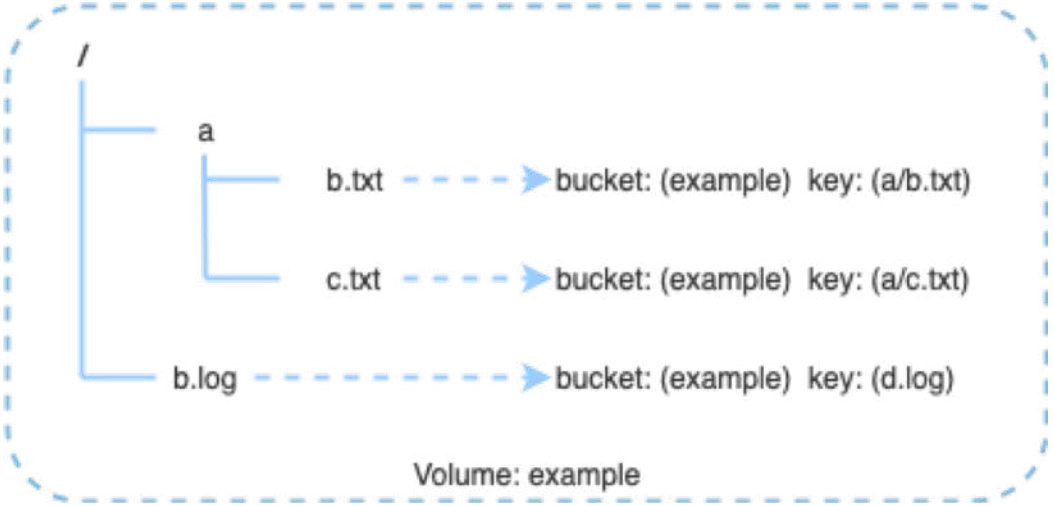
数据多接口访问

- POSIX文件系统
- 对象存储S3
- C API（动态链接库）
- Java API（JNA）
- HDFS compatible API

架构设计 - 文件 vs 对象语义

ChubaoFS	Object Storage
Cluster Name	Region
Volume	Bucket
Path & File	Object

Semantics Matchup



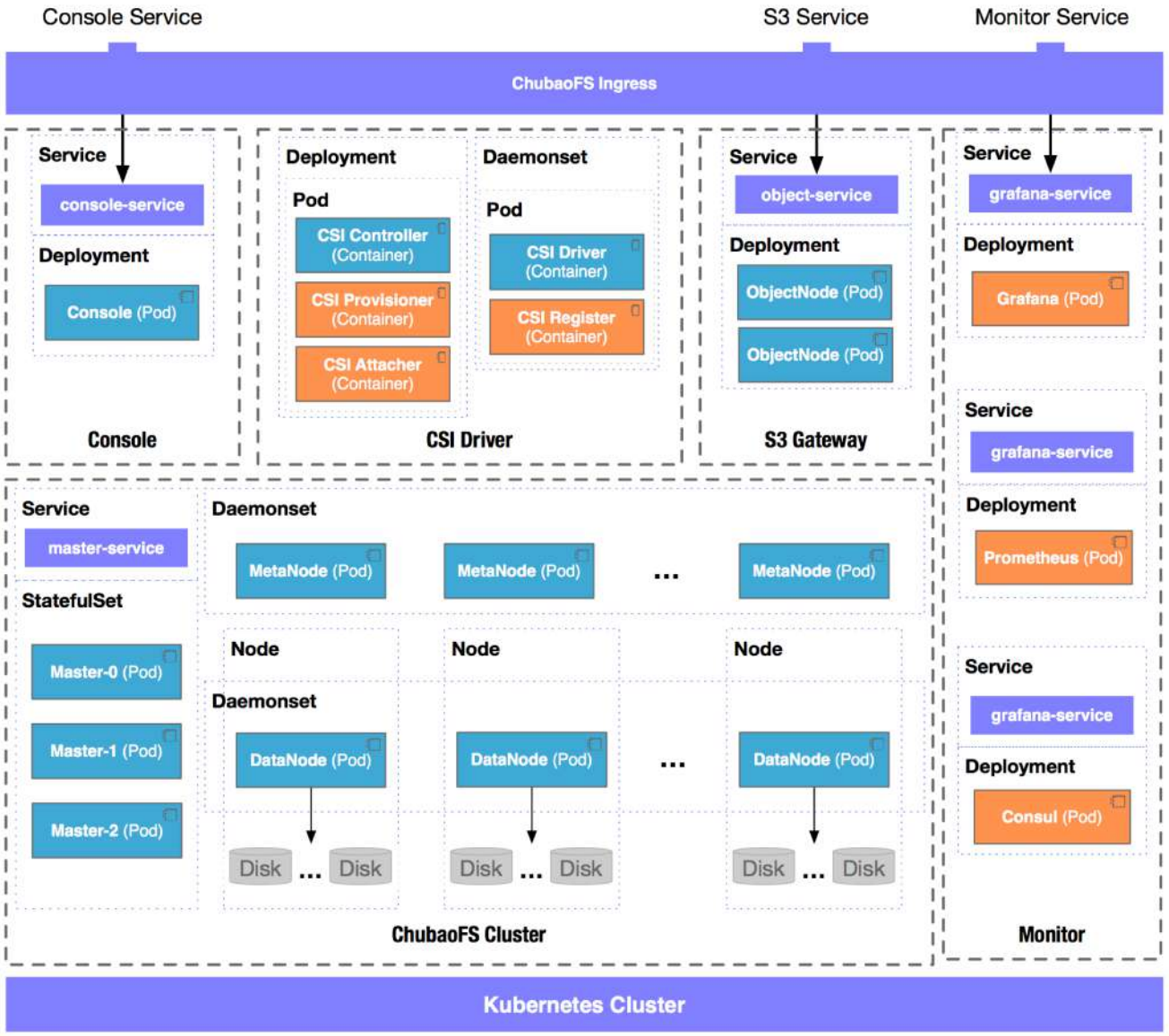
Example

- 优点：元数据操作本质上与文件系统类似，开销小
- 缺点：与S3原生语义有区别，例如：不能同时存在/a和/a/b对象

融入生态 - 云原生

- Kubernetes
  - Most often deployed on Kubernetes
- Harbor
  - Provides backend storage for Harbor Image Center
- Helm
  - Support Helm v2 for deployment
- Prometheus
  - Default monitoring system
- Rook
  - Storage orchestrator for Kubernetes (WIP)

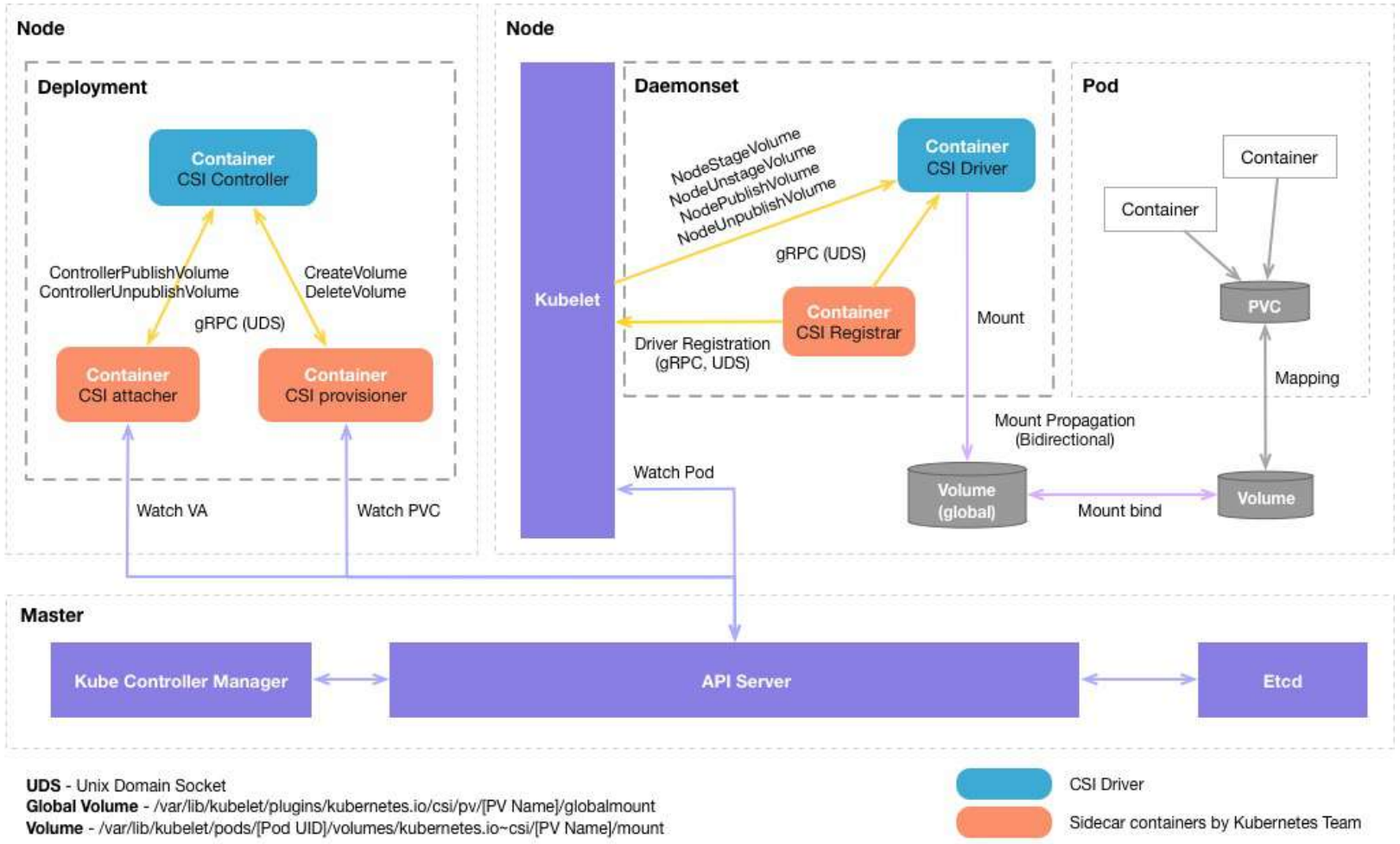
融入生态 - 云原生



云原生社区

- 监控：Prometheus
- **部署：Helm**
- 使用：CSI driver

融入生态 - 云原生



云原生社区

- 监控：Prometheus
- 部署：Helm
- **使用：CSI driver**

社区

云原生社区 Cloud Native Community

- github.com/chubaofs/chubaofs
- https://chubaofs.readthedocs.io/zh_CN/latest/

云原生社区
Cloud Native Community

云原生社区Meetup

第二期·北京站

THANKS