
Architect
SACC
2022 中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2022
激发架构性能 点亮业务活力
云上会议 网络直播 | 2022年10月27-29日
ITPUB

JuiceFS平台构建与海量数据存储实践

携程 - 高级云原生研发工程师 - 张妙成

自我介绍

张妙成 - 携程 - 系统研发部

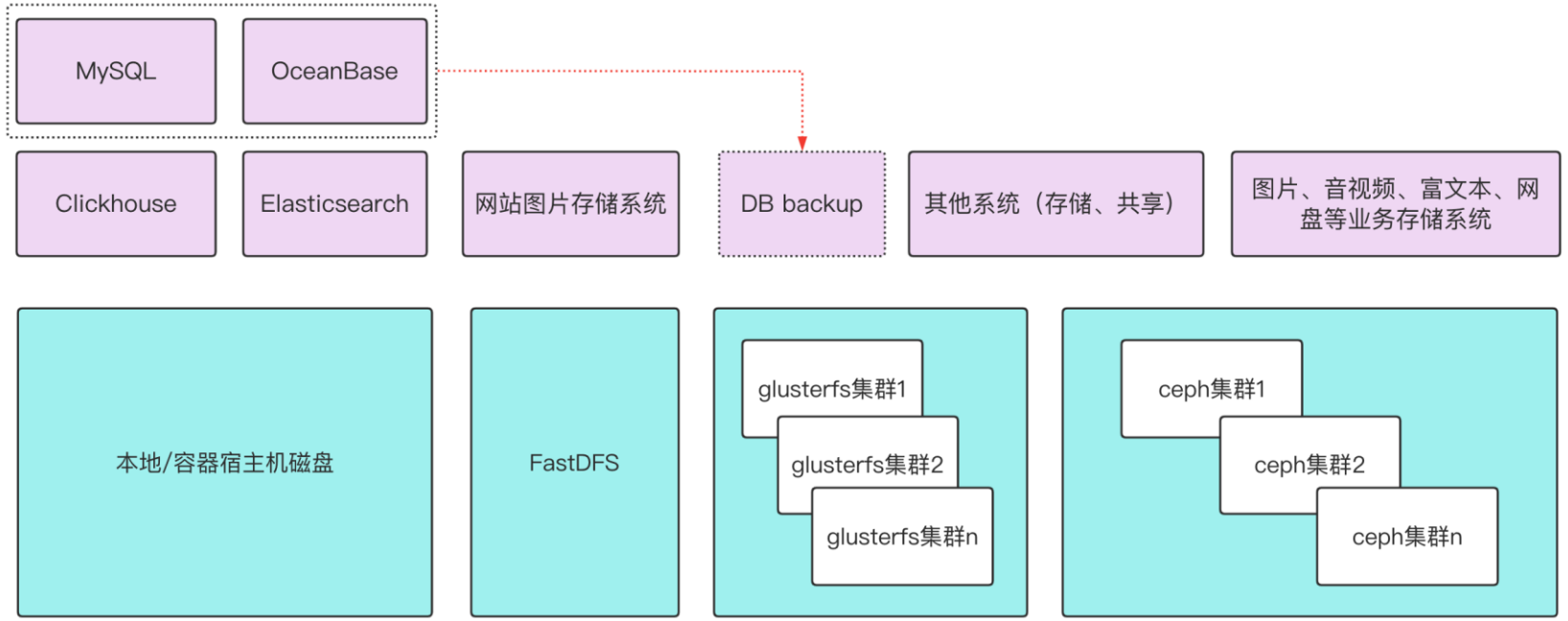
- Elasticsearch PAAS 研发、运维
- JuiceFS 在海量数据场景下的落地

目录 CONTENTS

第一部分
PART 01

需求与技术选型

数据场景、痛点



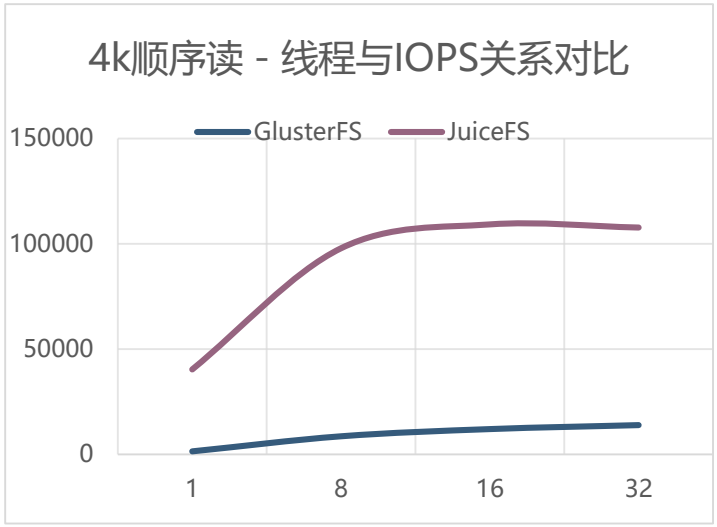
冷数据规模：10PB+

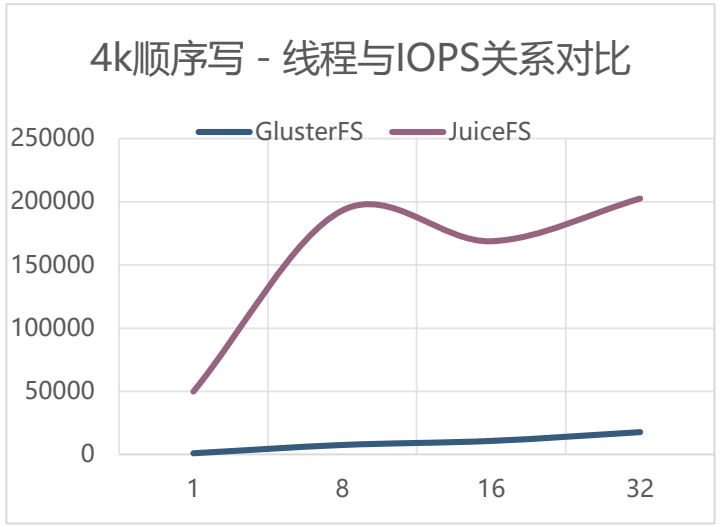
1.弹性：受机器采购周期的制约，无法灵活地弹性扩缩容

2.性能：GlusterFS 顺序读写性能较差，ls、du命令存在性能问题

3.成本：机器替换和扩缩容操作的运维成本、机器使用率影响成本

数据场景、痛点





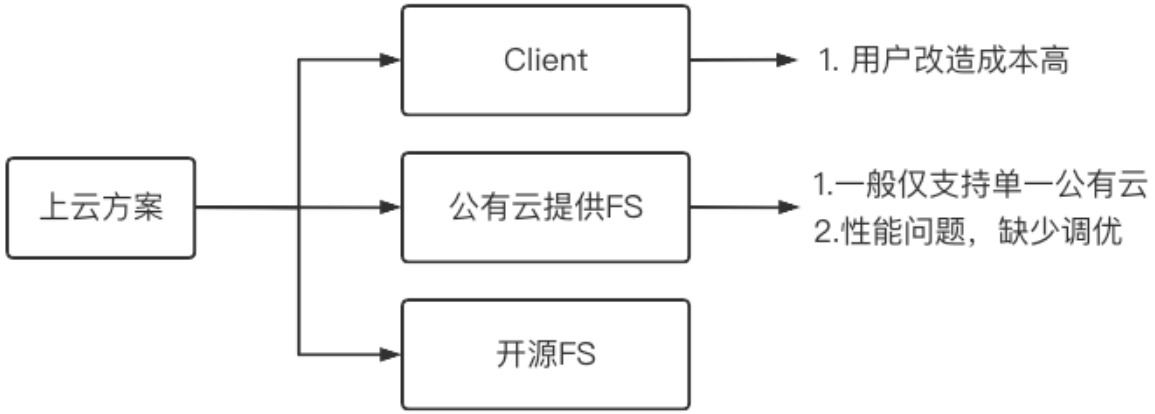
	最高磁盘使用率	最低	平均
GlusterFS	70%	30%	50%
Clickhouse（日志）	90%	30%	55%
Elasticsearch（日志）	80%	40%	60%

1.弹性：受机器采购周期的制约，无法灵活地弹性扩缩容

2.性能：顺序读写性能较差，ls、du命令存在性能问题

3.成本：机器替换和扩缩容操作的运维成本、机器使用率影响成本

文件系统技术选型

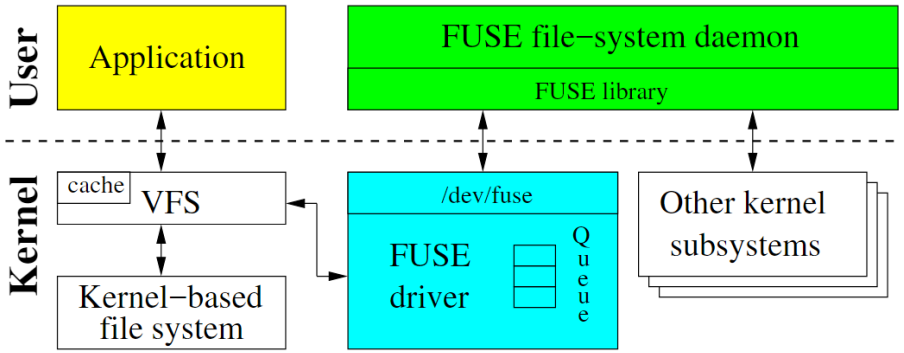


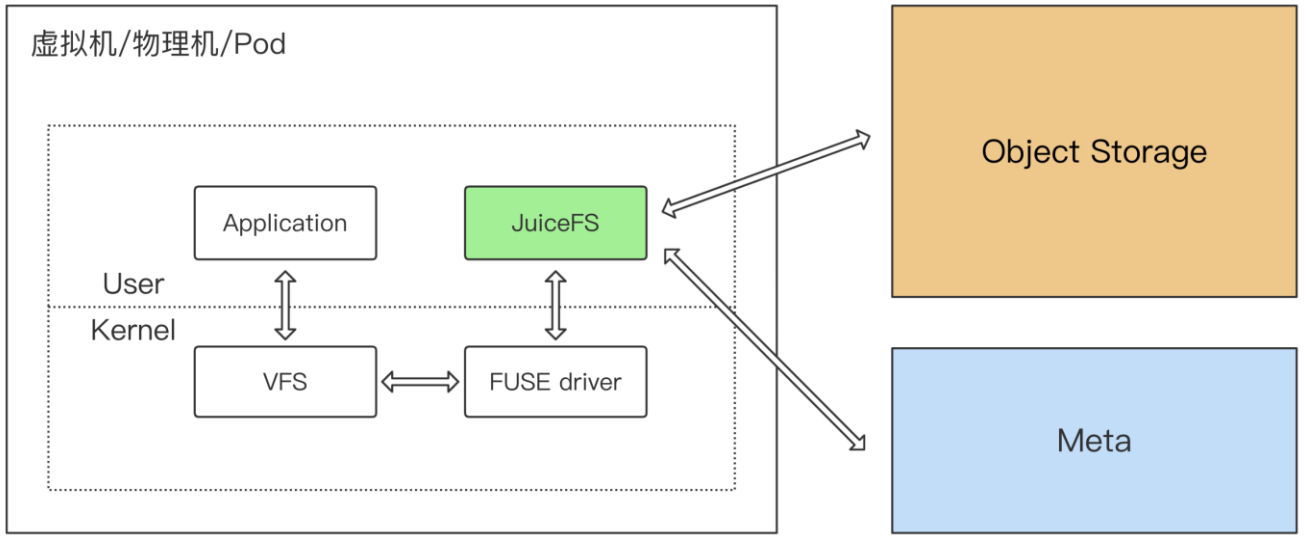
高性能　　强一致

无侵入　　云原生

	完全posix支持	强一致性
JuiceFS	✓	✓
CephFS	✓	✓
CubeFS	✖	✓
Alluxio	✖	✖
S3FS(amazon)	✖	✖

JuiceFS 整体架构

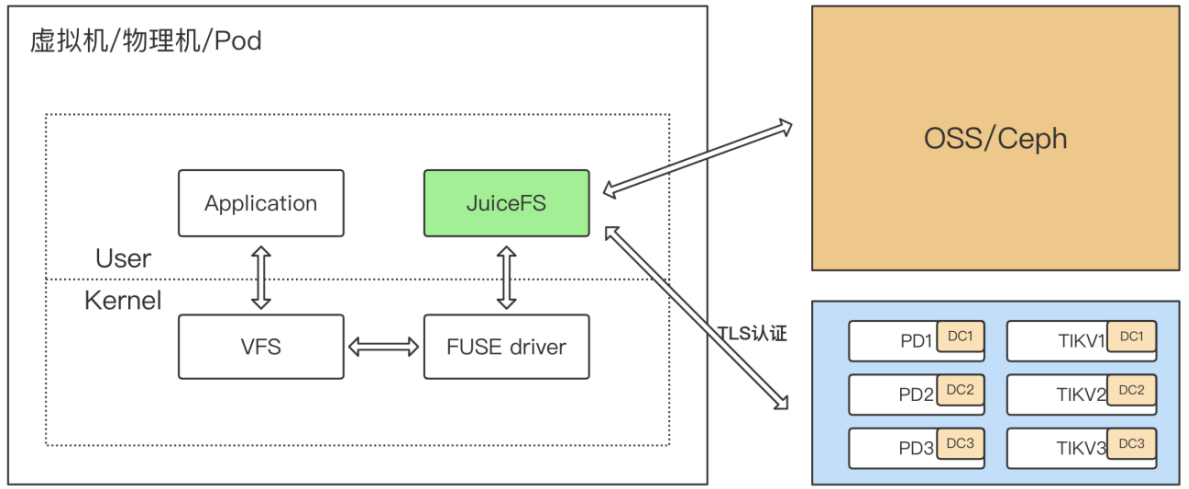




优势：crash不影响内核、更加灵活、开发周期快、易于二次开发

劣势：读写流程长、存在数据copy、吞吐量较差

POC



机器CiCode	CPU	Memory	Storage	Network
Node1	2 Socket / 20 Core / 40 Thread	128G	1.9T SATA SSD	bond0 25G
Node2	2 Socket / 20 Core / 40 Thread	128G	960G SATA SSD	bond0 25G
Node3	2 Socket / 20 Core / 40 Thread	128G	1.2T SATA SSD	bond0 25G

测试：

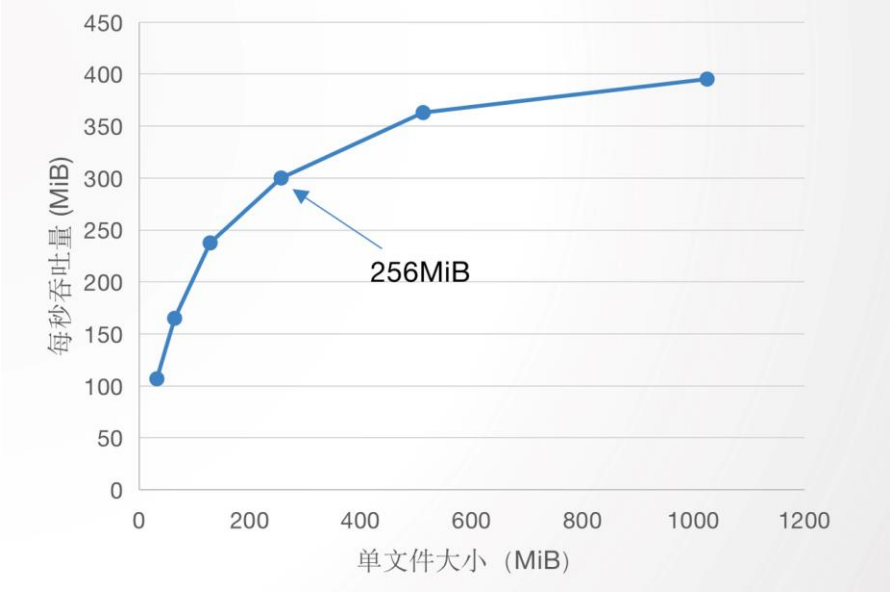
1.整体性能测试

2.TIKV性能测试

3.破坏性测试

POC - 整体性能测试

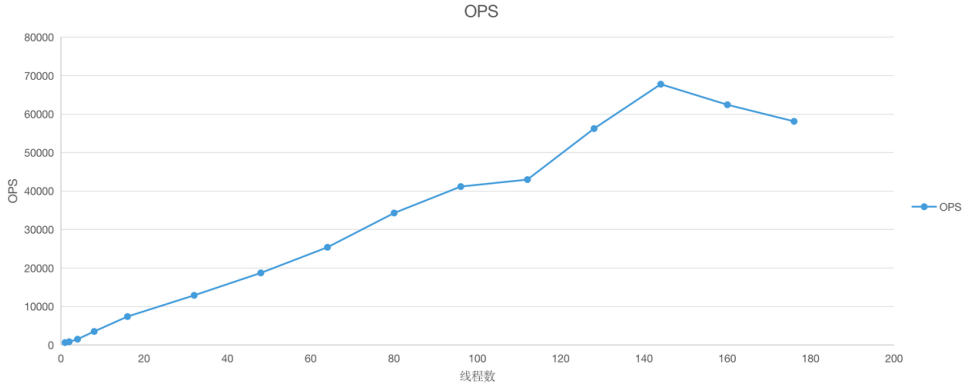
	fio
本地HDD	928MiB/s
JuiceFS	101MiB/s（711MiB/s）



1. 随着文件大小增大，吞吐量也随之增大。
2. 单文件为 128MB~256MB 左右，吞吐量与文件大小的增长曲线明显放缓。

POC - TIKV 性能测试

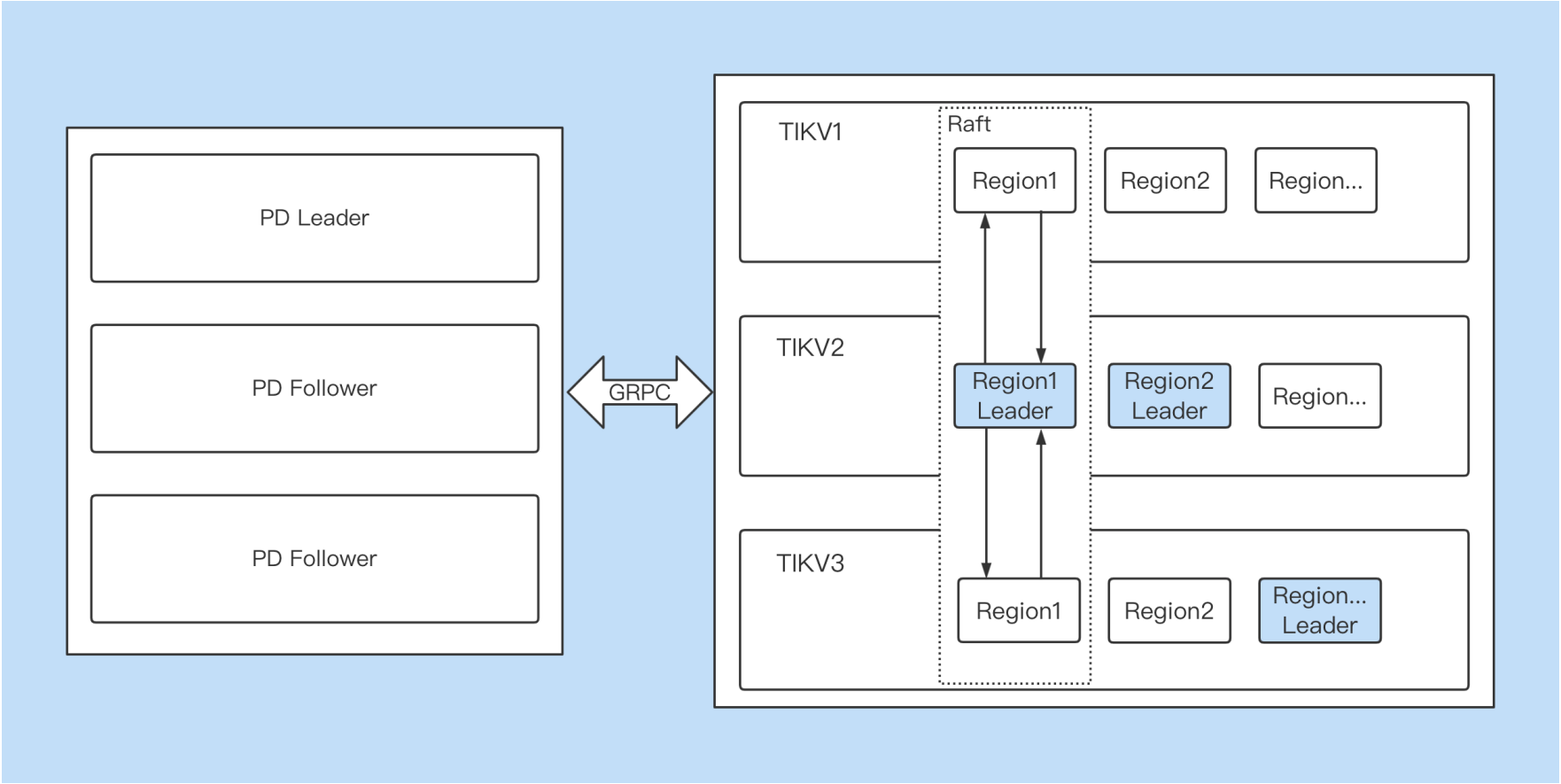
Write 事务写入操作，随着客户端线程数增加，TPS 上升，峰值超过 3w



Get 事务读取操作， 随着客户端线程数增加，QPS 上升，峰值接近 7w

单次操作的响应时间P99<10ms, 相较于对象存储 20-200ms的响应，不会成为瓶颈

POC - 破坏性测试


PD Leader
PD Follower
PD Follower
GRPC
TIKV1
Raft
Region1
Region2
Region...
TIKV2
Region1
Leader
Region2
Leader
Region...
TIKV3
Region1
Region2
Region...
Leader

POC - 破坏性测试

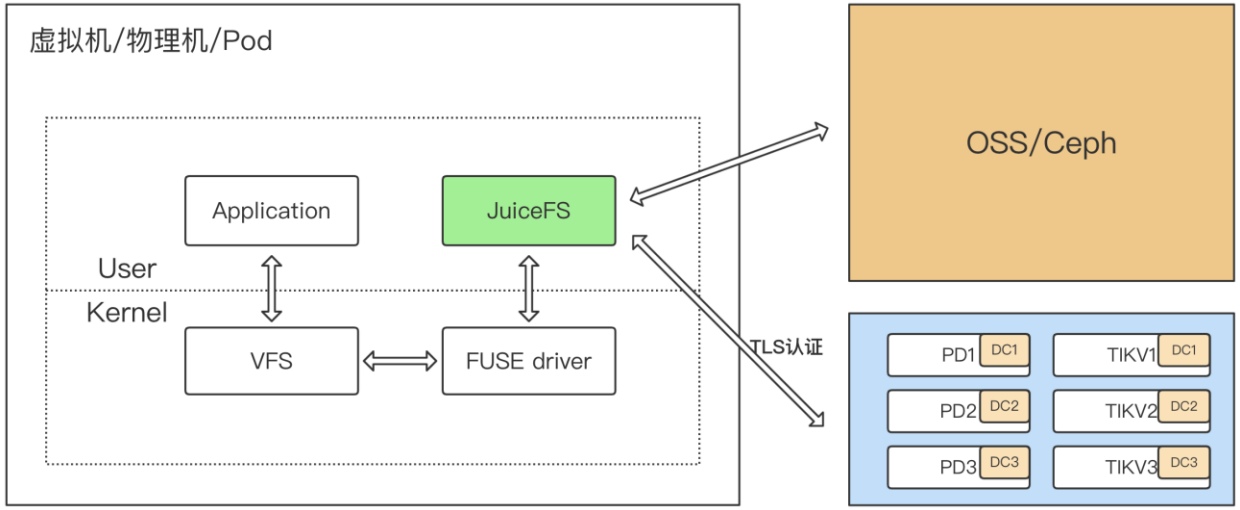
TIKV - Leader region故障

File size/count	正常	异常	Diff（ms）
写 4MiB/1024	237035.52	249333.76	12298.24
读 4MiB/1024	360222.72	362577.92	2355.2

PD - Leader节点故障

File size/count	正常	异常	Diff（ms）
写 4MiB/1024	237035.52	247531.52	10496
读 4MiB/1024	362332.16	362577.92	245.76

POC - 特性总结

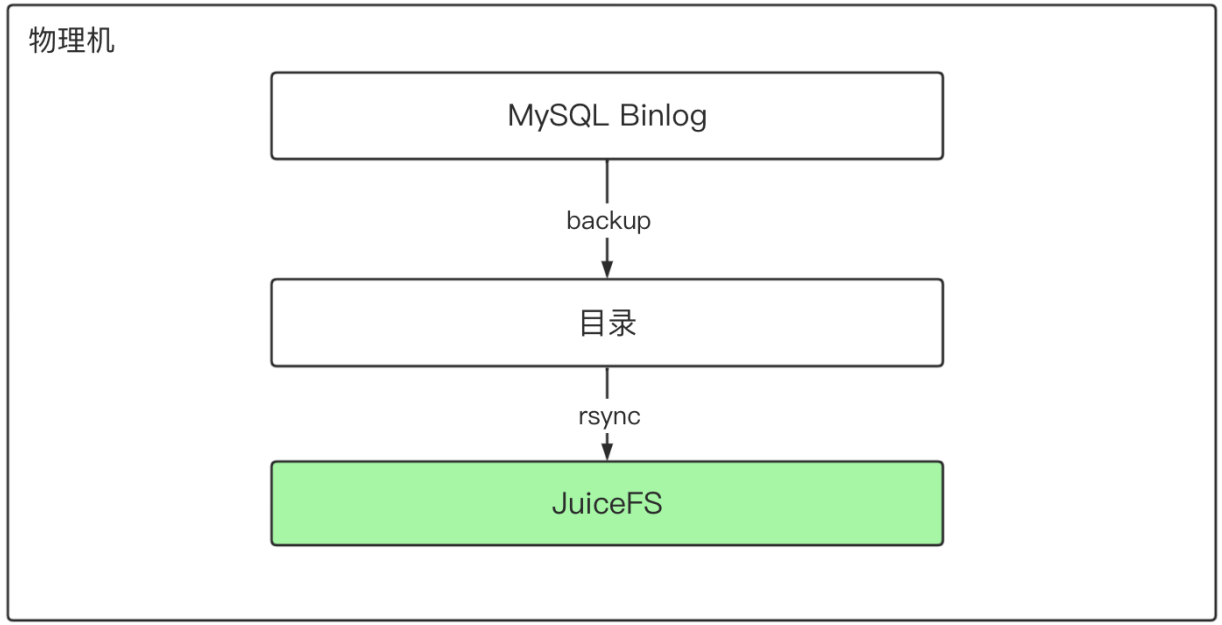


1. 在大文件顺序读写场景下性能较好
2. TIKV 作为元数据库，拥有较好读写性能
3. TIKV 作为元数据库，故障时影响时间在秒级

第二部分
PART 02

典型业务使用方式

数据库备份



单机优化：

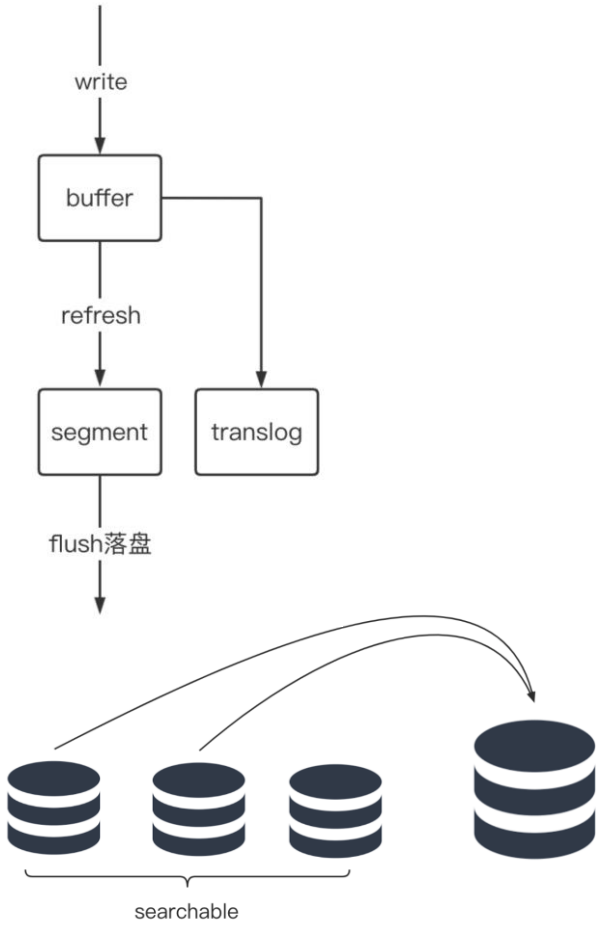
写入场景增大buffer、max-upload线程

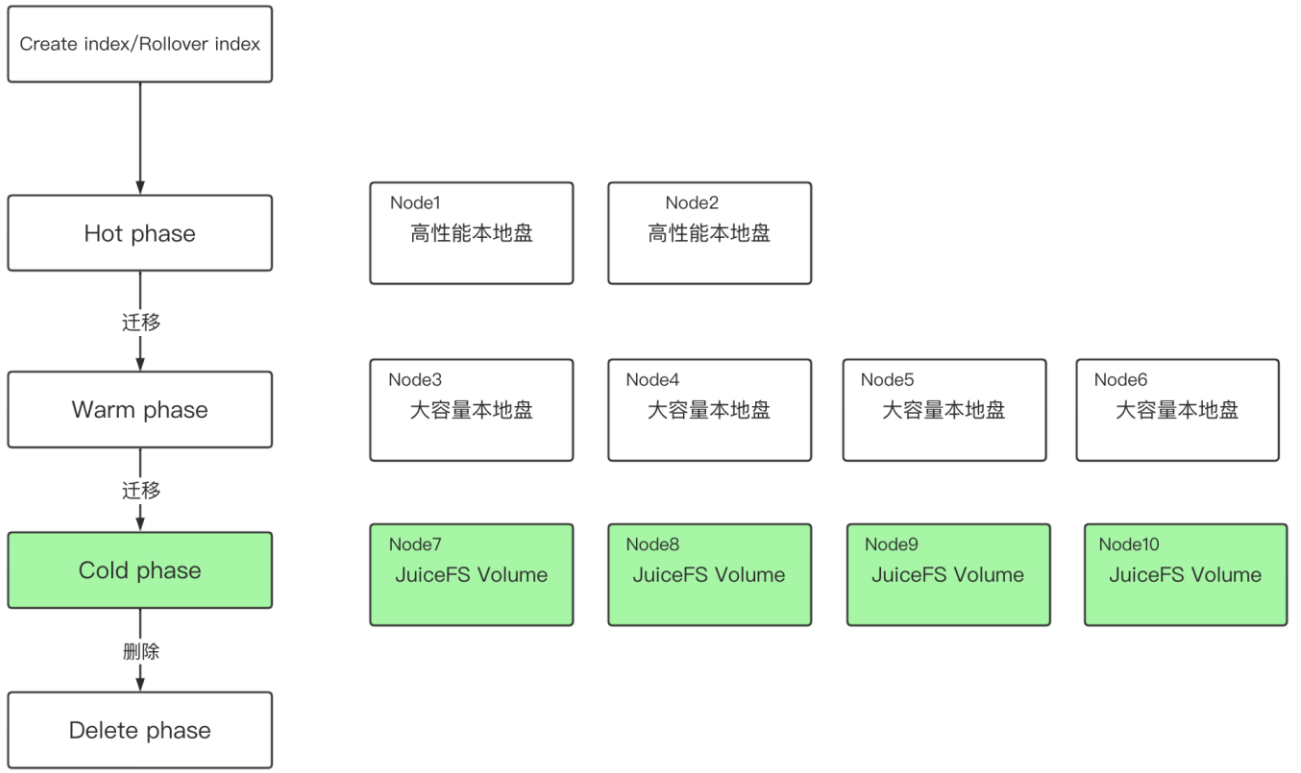
writeback模式

单volume场景：

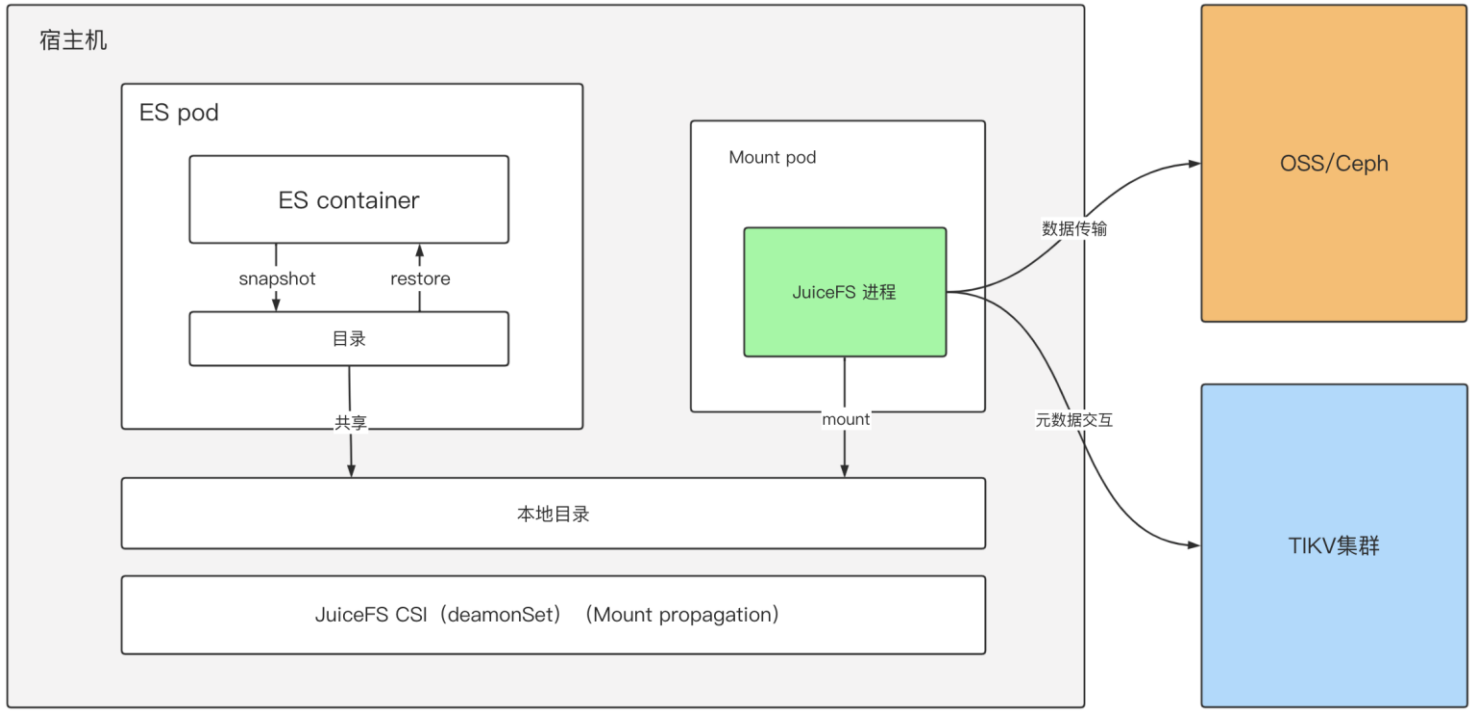
trash、session等bgjob在TIKV中锁竞争

ES 冷热分离


write
buffer
refresh
segment
translog
flush落盘
searchable


Create index/Rollover index
Hot phase
迁移
Warm phase
迁移
Cold phase
删除
Delete phase
Node1
高性能本地盘
Node2
高性能本地盘
Node3
大容量本地盘
Node4
大容量本地盘
Node5
大容量本地盘
Node6
大容量本地盘
Node7
JuiceFS Volume
Node8
JuiceFS Volume
Node9
JuiceFS Volume
Node10
JuiceFS Volume

ES Snapshot & 其他

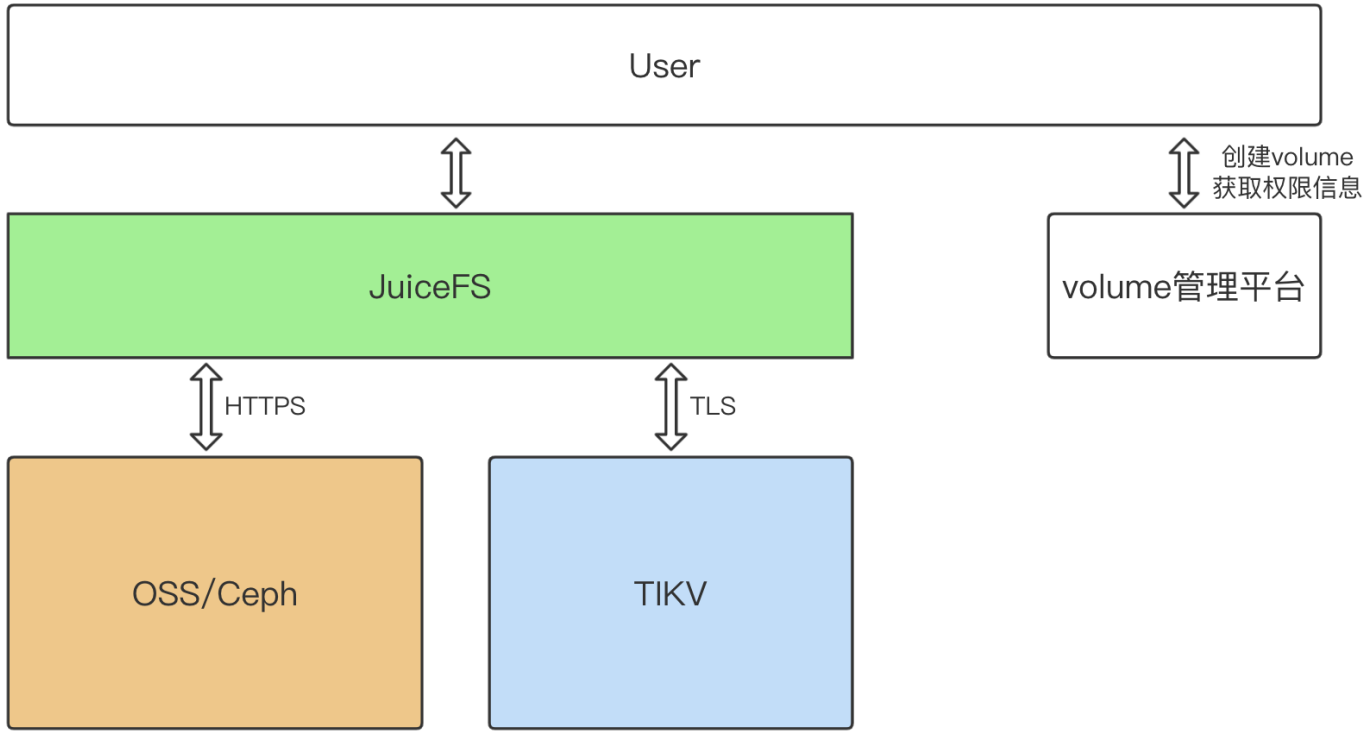


1. clickhouse 冷数据
2. AI 训练数据
3. 图片视频数据存储系统等...

第三部分
PART 03

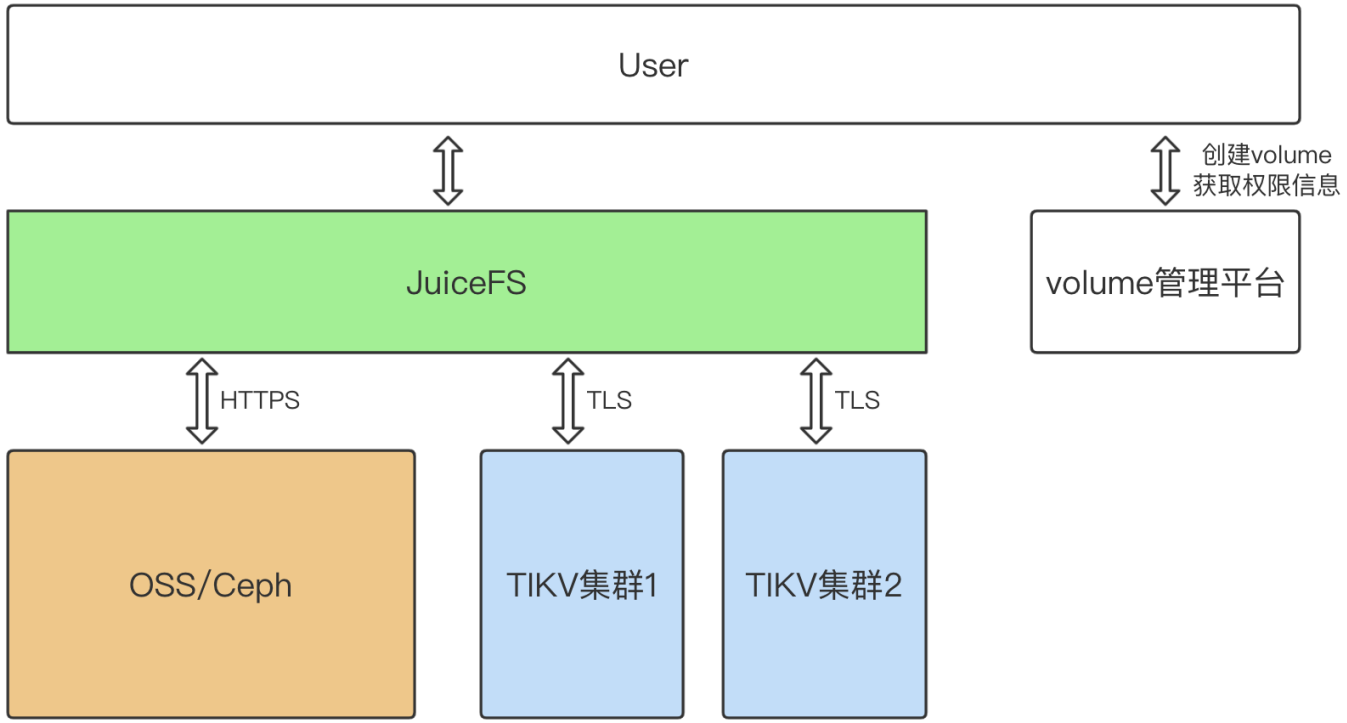
JuiceFS 平台搭建与演进

第一阶段 - 单分布式元数据库



- 权限
- 缓存
- 统计指标

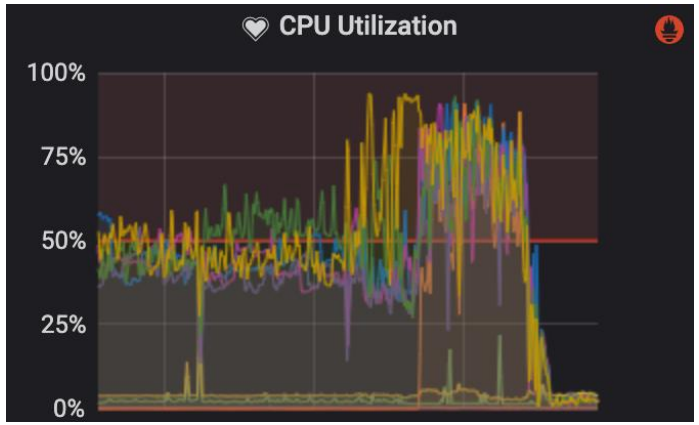
第二阶段 - 元数据隔离

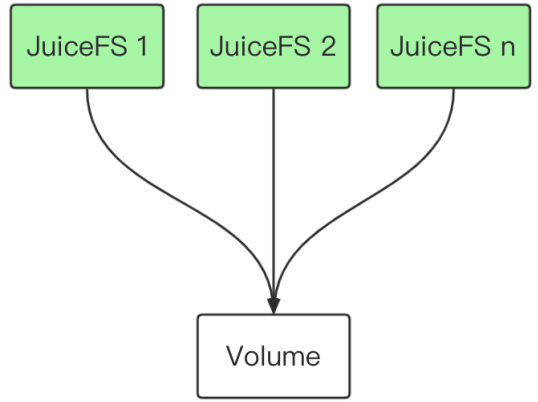


TIKV集群拆分

volume拆分

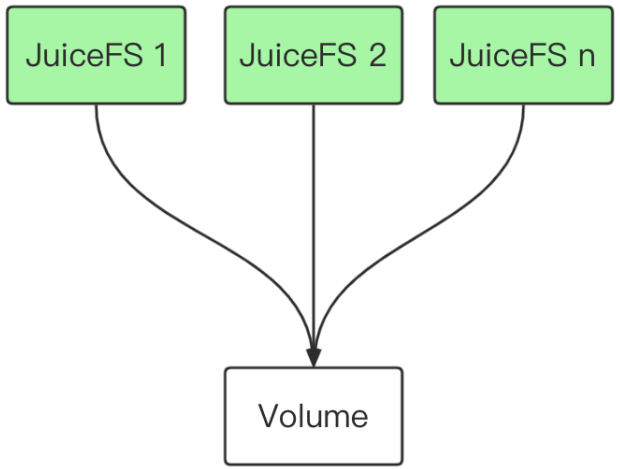
故障 - TIKV CPU 被打满造成雪崩





0. 常规运维 & tikv参数调优回滚
1. 封iptables确定造成故障的业务
2. 500+ JuiceFS同时clean trash

多进程访问volume场景优化



DB备份、共享场景

监控

meta 接口粒度的监控

功能

增加缓存能力

分布式锁

task抽离

增强清理OSS trash数据的能力

熔断

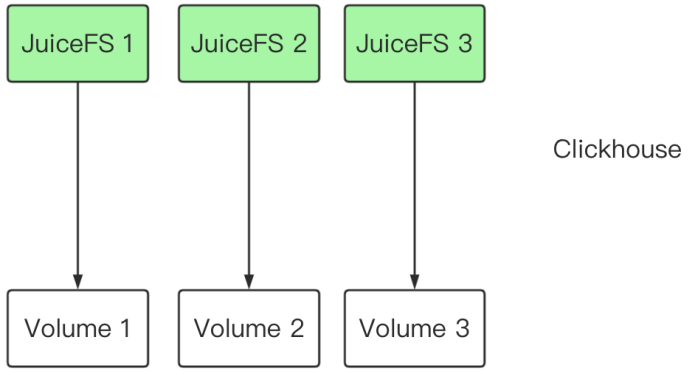
meta接口、OSS接口限流

其他问题与解决方案

遇到的问题?

1. 对象存储中存在僵尸数据、容量不一致

2. 带宽有限，毛刺多



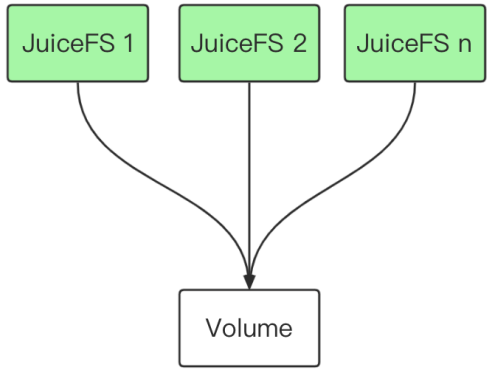
Clickhouse

需要做什么?

1.增强清理OSS trash数据的能力

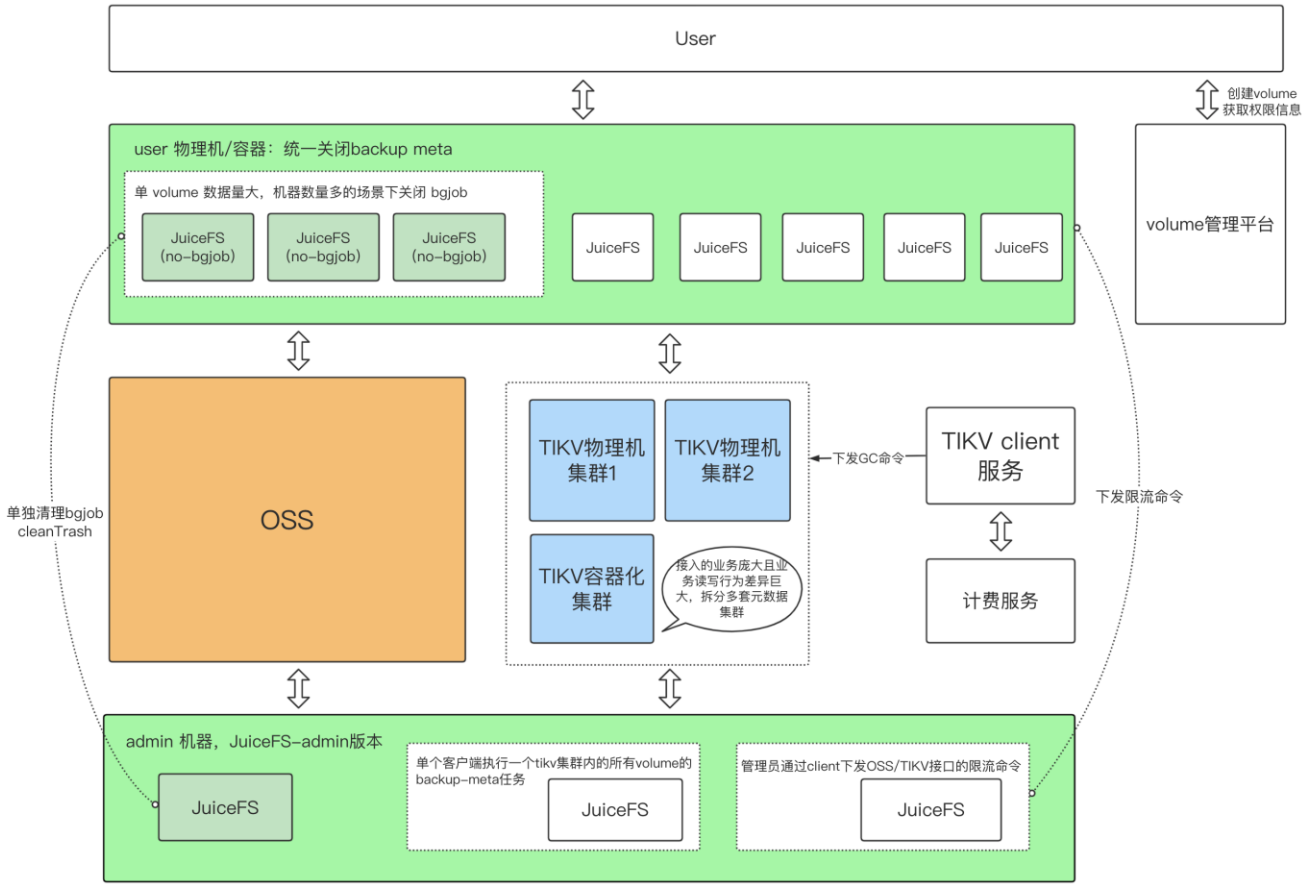
2.CSI的volume单独挂载一个节点

3.meta接口、OSS接口限流



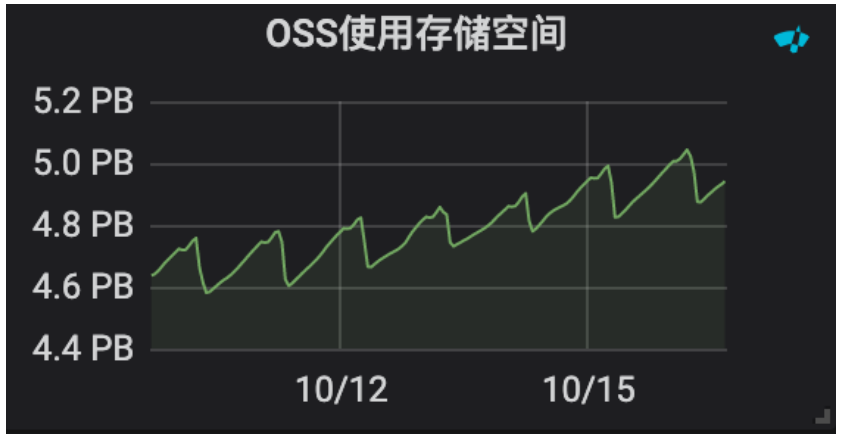
DB备份、共享场景

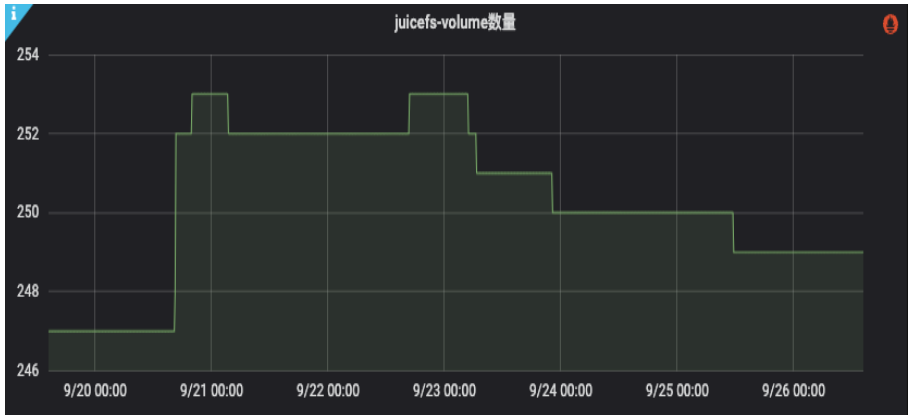
第三阶段 - 多版本、多集群

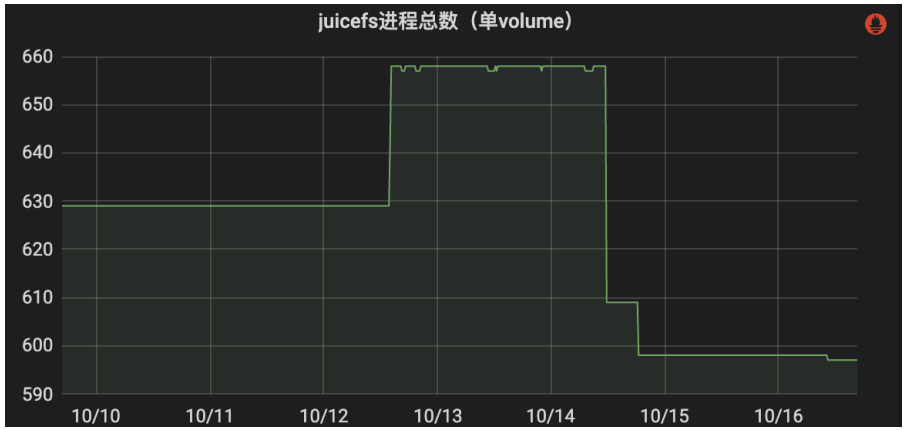


- admin处理bgjob
- 统一元数据备份
- 动态限流
- TIKV容器化
- TIKV GC控制

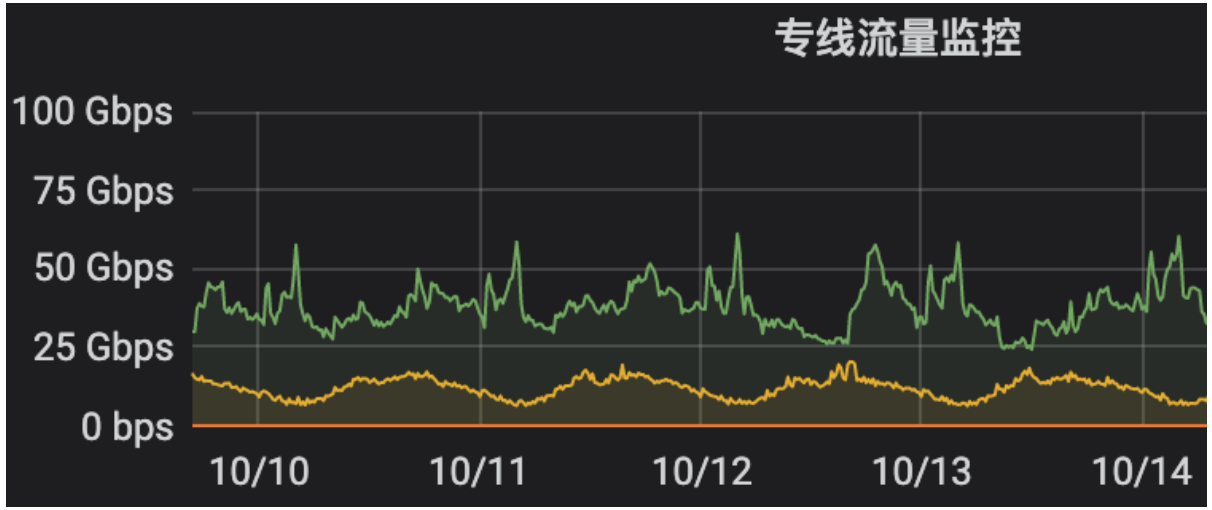
JuiceFS 平台接入量

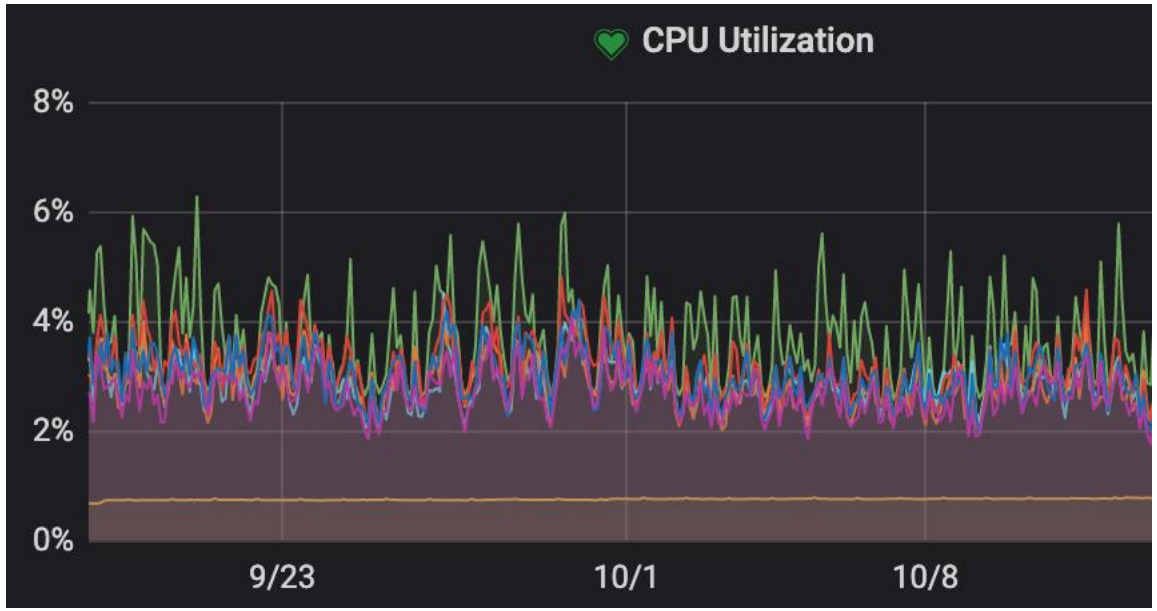






JuiceFS 平台核心指标监控





第四部分
PART 04

展望与总结

总结 - 海量数据上云挑战

性能、一致性

1.缓存、writeback等

2.元数据隔离减少抖动

3.数据清理加速

带宽、元数据压力

1.volume粒度带宽限流

2.元数据接口粒度限流

3.减少访问频率（缓存）

4.减少锁竞争和空转

资源、场景管理

1.volume、权限、计费

2.多版本、多TIKV集群

3.支持虚机、容器场景

高可用、故障恢复

1.定期备份元数据

2.元数据库三中心部署

价值

成本

1. 实际成本节约 50%+
(不预留磁盘空间、单副本)
2. 运维成本大幅降低

弹性

1. 灵活扩缩容（与机器采购周期解耦）
2. 单机无限磁盘容量

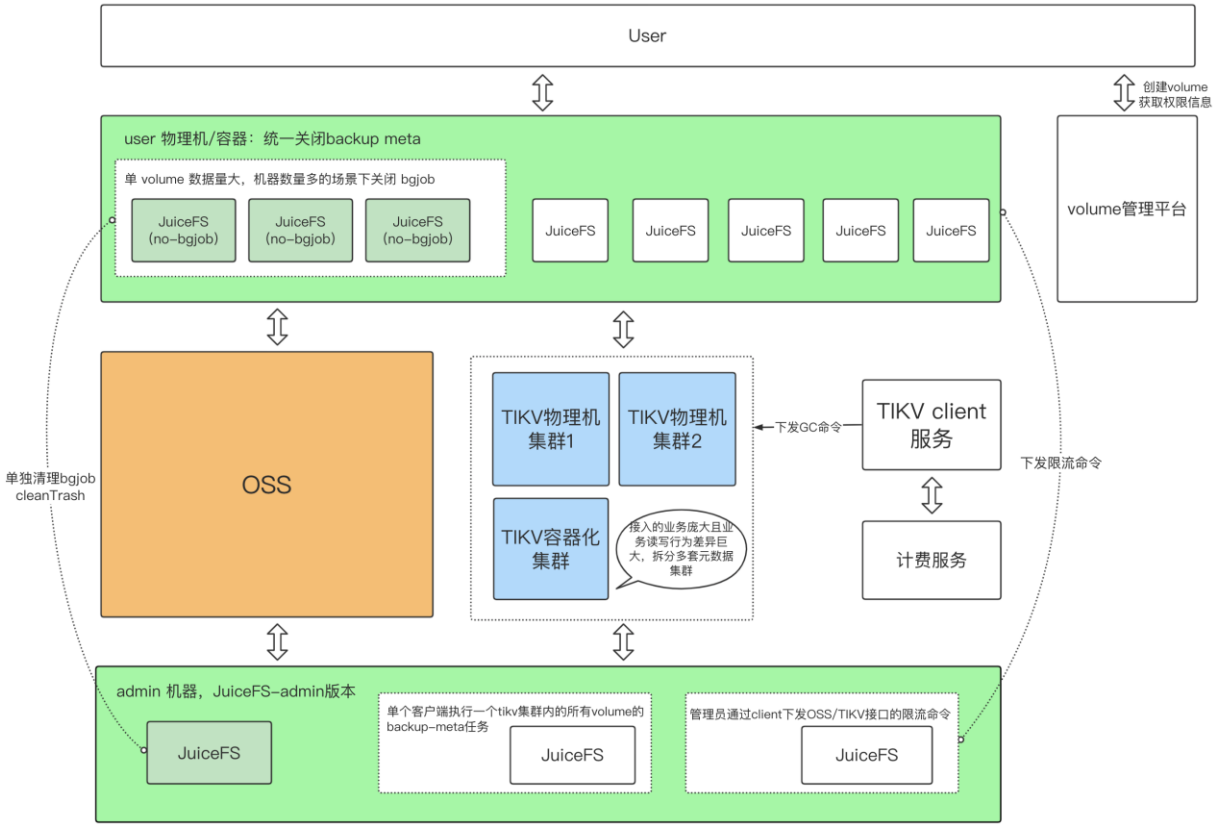
性能

1. 吞吐量高(备份业务无缓存outgoing_traffic 1.5Gbps)
2. ls、du等命令高性能
（12万文件夹共1.4亿文件ls耗时仅2s）

其他

1. 提供统一的数据上云方案（POSIX）

现存问题



1.如何屏蔽对TIKV集群的感知?

2.TIKV如何双活?

3.JuiceFS 如何不停服升级?

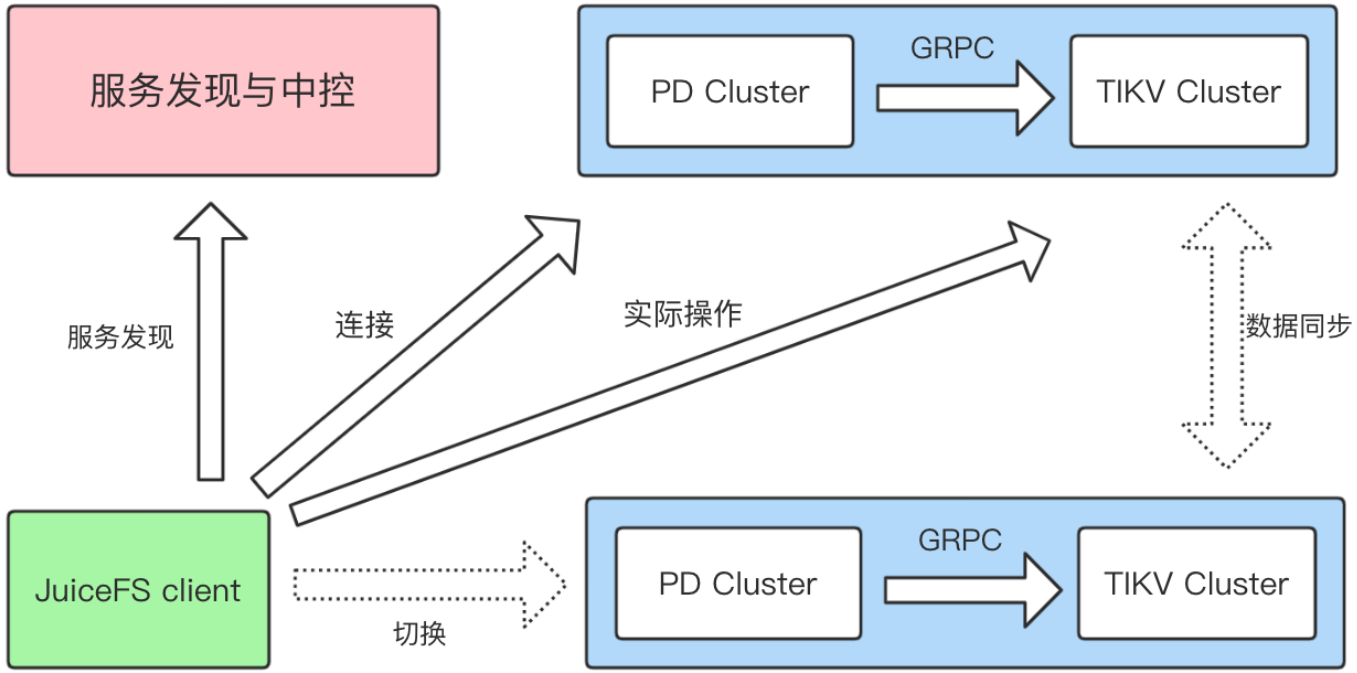
4.如何接入在线应用?

优化方向

1. proxy屏蔽对TIKV集群的感知 && 故障时切换集群

2. TIKV Txn CDC 方式实现数据同步

3. 不停服升级方案：软连接



THANKS