DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

Curve块存储在云原生数据库领域的实践

网易数帆
Curve Maintainer
吴汉卿

目录

- Curve整体介绍
- Curve块存储架构介绍
- Curve块存储云原生数据库实践
- 后续规划

Curve整体介绍

打造云原生、高性能、稳定易运维的开源分布式存储系统

- 支持私有云、公有云、混合云上部署
- 支持CSI插件
- 支持容器化部署(curveadm)
- 支持K8s部署
- ...

云原生

- 支持RDMA
- 支持SPDK
- 支持多级缓存
- 数据/元数据性能可水平扩展
- ...

高性能

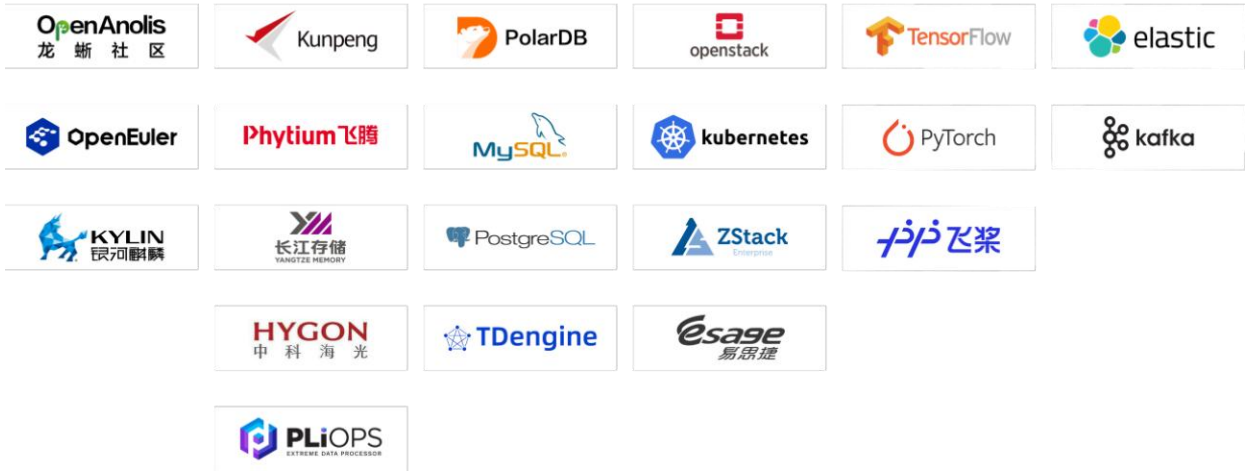
- 一键部署、一键升级、一键扩容
- 全局无单点故障
- 数据/元数据多副本高可靠
- 常规故障及日常运维IO时延不抖动
- ...

易运维

Curve整体介绍

开源生态

操作系统 | 芯片 | 数据库 | 云原生 | AI 训练 | 大数据



官方认证

- ✓ **信创认证**：国家工业信息安全发展研究中心测试结果显示，Curve 在文件存储与块存储**通过全部49个测试用例**
- ✓ Curve进入CNCF **沙箱**，意味着全球顶级开源基金会对Curve存储系统及开源社区的认可

社区发展

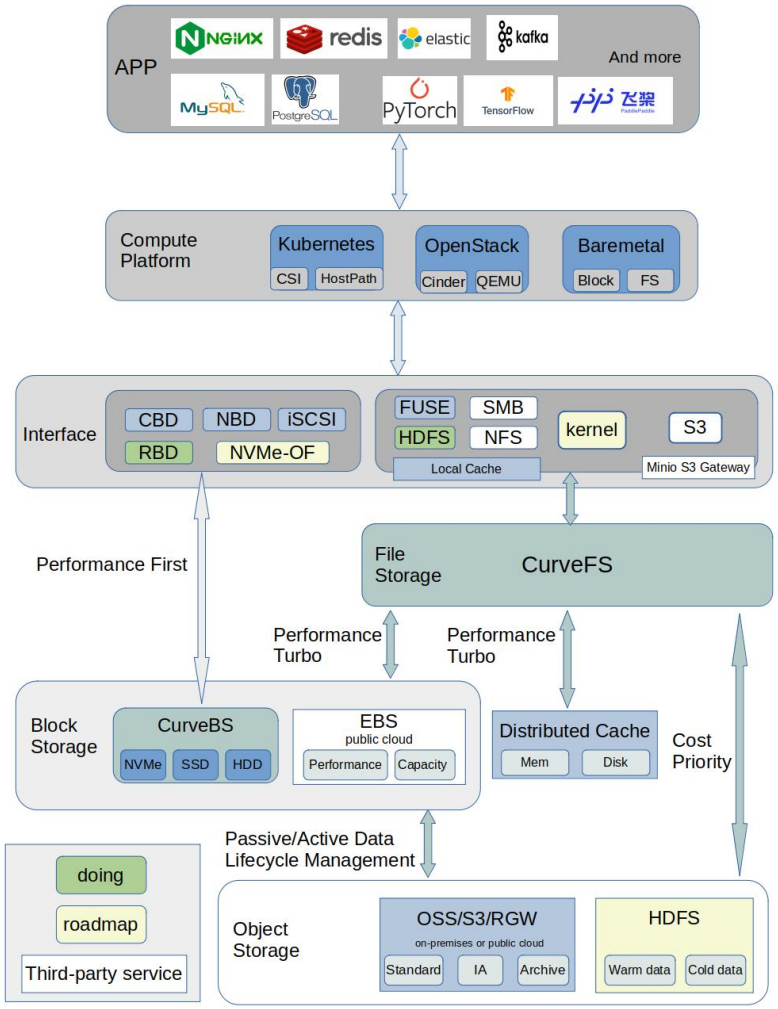
- ✓ 主页: https://opencurve.io/
- ✓ Github: https://github.com/opencurve/curve 2K+ star
- ✓ 50+ 外部开发者
- ✓ 公众号：OpenCurve
- ✓ 用户群：添加微信号OpenCurve_bot可邀请加群

Curve整体介绍

- 应用场景
  - ✓ 对接OpenStack平台为云主机提供高性能块存储服务
  - ✓ 对接Kubernetes为其提供RWO、RWX等类型的持久化存储卷
  - ✓ 对接PolarFS作为云原生数据库的高性能存储底座，完美支持云原生数据库的存算分离架构
  - ✓ Curve作为云存储中间件使用S3兼容的对象存储作为数据存储引擎，为公有云用户提供高性价比的共享文件存储
  - ✓ 支持在物理机上挂载使用块设备或FUSE文件系统

- 生产用户

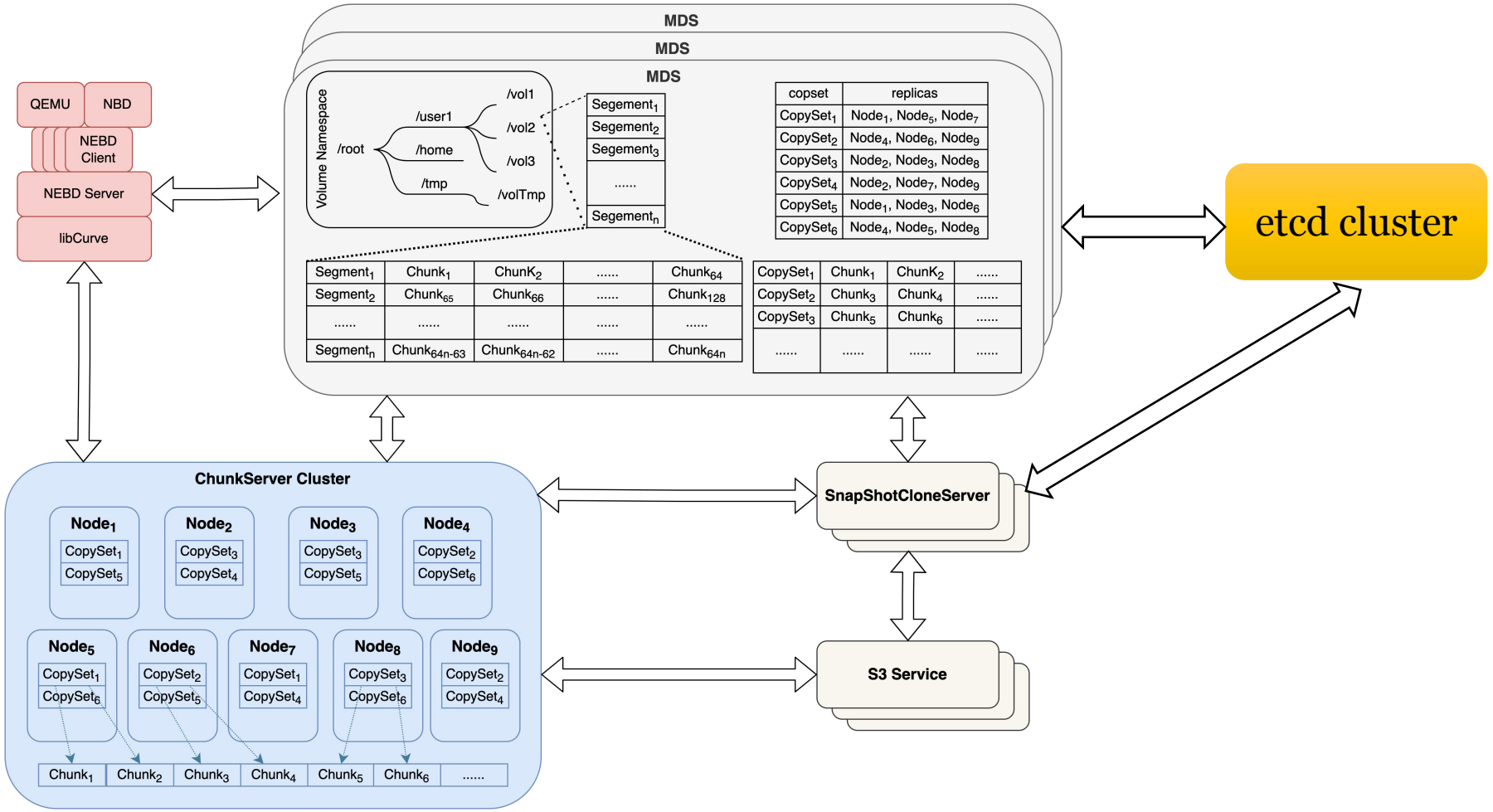
网易严选 网易云音乐 有道 youdao LOFTER 网易云信
网易游戏 翼信蓝鲸 SmartMore 思谋 万方电子 创云融达



目录

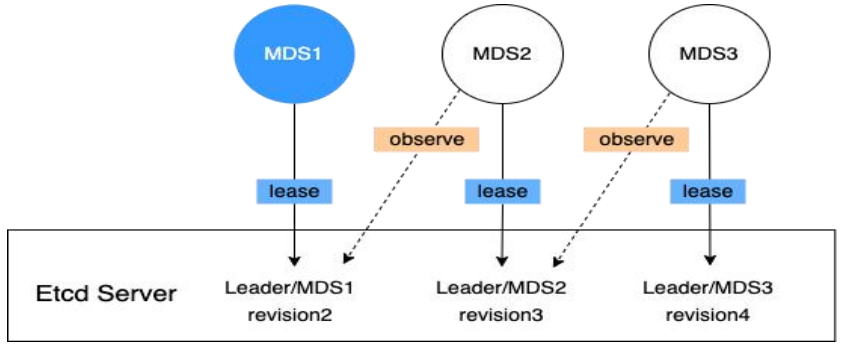
• Curve整体介绍
• Curve块存储架构介绍
• Curve块存储云原生数据库实践
• 后续规划

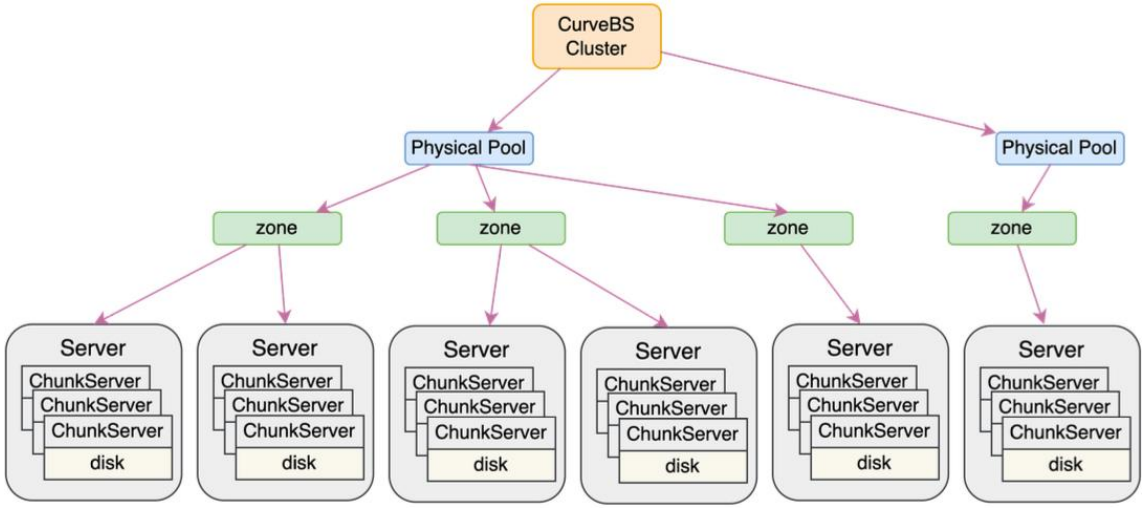
Curve块存储架构介绍



元数据管理 MDS+ETCD

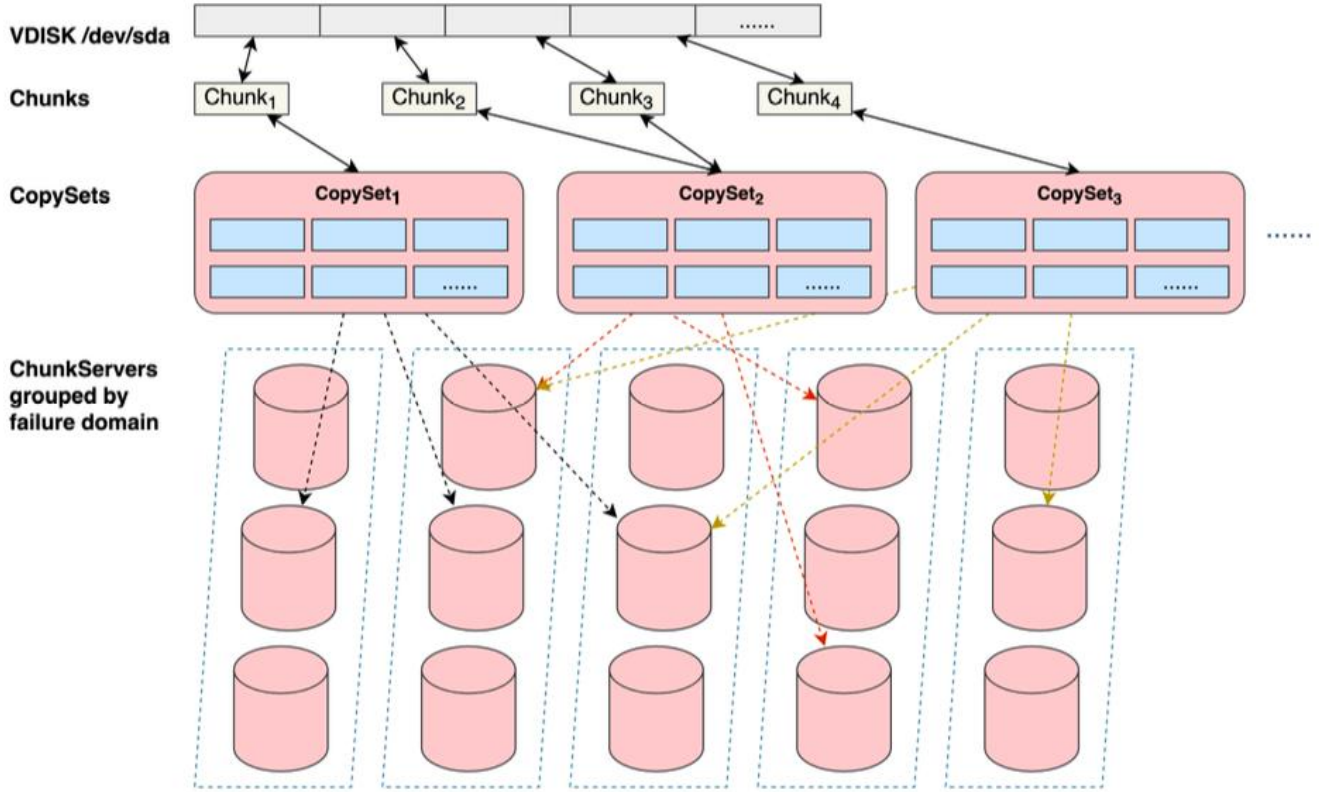
- MDS主备模式，通过etcd选主，元数据存储在etcd
- 管理集群的拓扑结构
- 存储节点Copyset的分配和管理
- 集群的均衡、异常调度





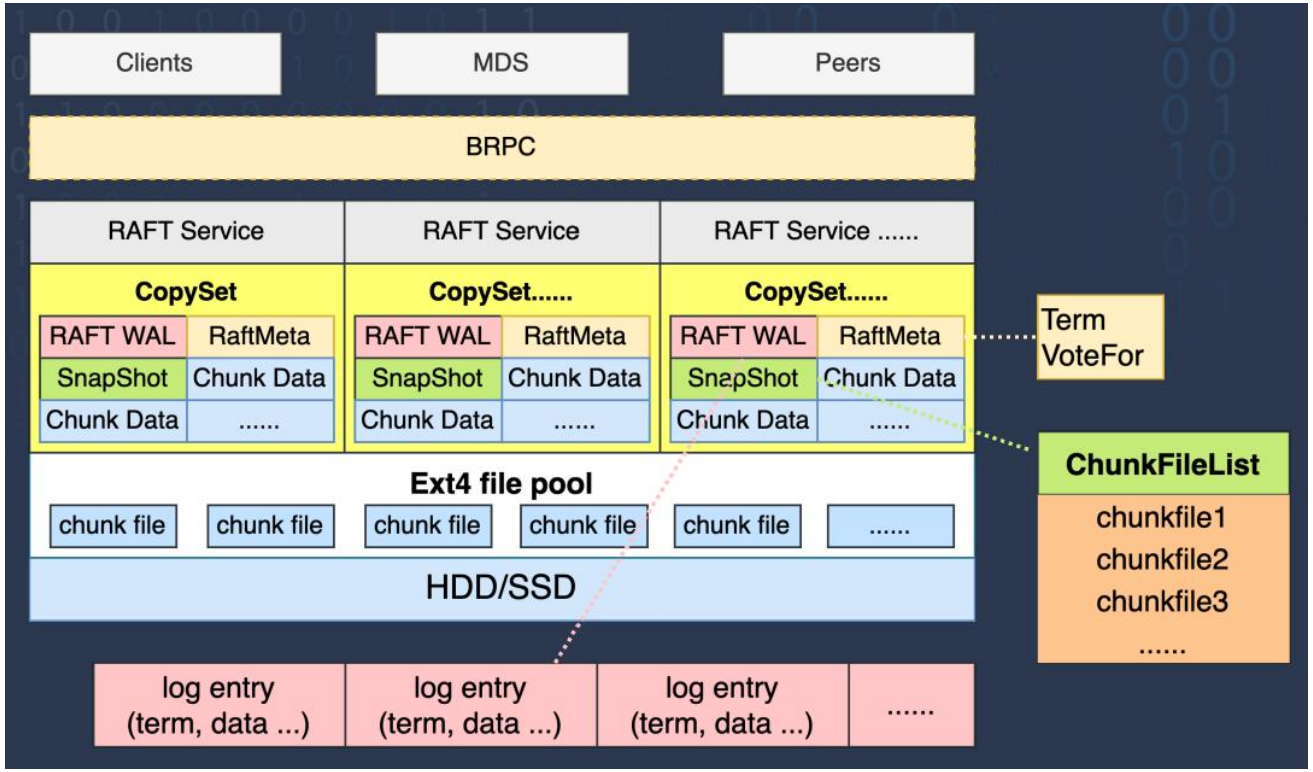
元数据管理 MDS+ETCD

- 卷的空间分配和卷元数据索引
- 空间分配在实际写入时触发
- 卷元数据索引
  - 卷 -> 逻辑chunk
  - 逻辑chunk -> 物理chunk



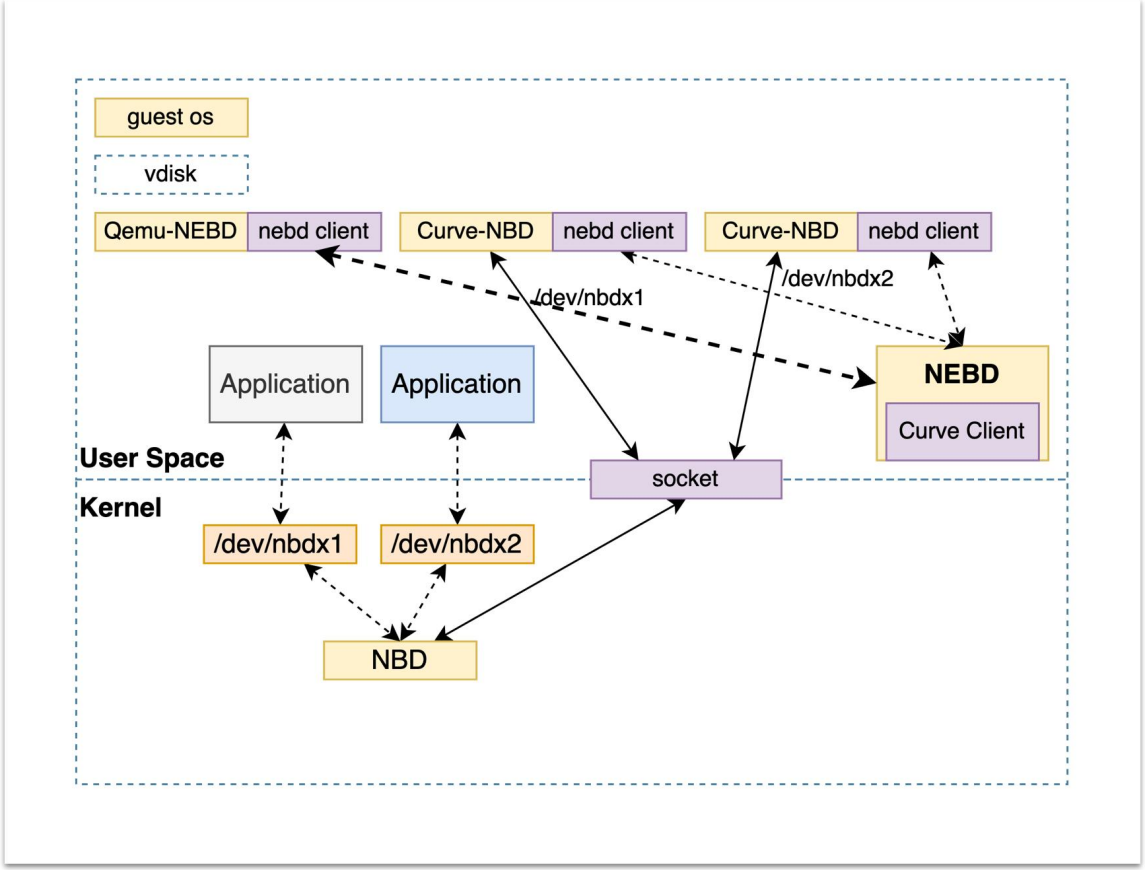
数据节点 Chunkserver

- 实际管理卷的数据存储
- 通过raft协议保证三副本的数据一致性
- 基于本地文件系统(ext4)实现，为了减少文件系统的写放大，通过预分配chunkfilepool的方式，对盘进行预写
- 读bypass raft
  - 单挂载，applied index
  - 多挂载，leader lease
- apply 读写分离 / fast apply
- 轻量级raft快照



客户端

- 向上层提供块设备接口
- 支持NBD、iSCSI、NVMe-oF、QEMU
- 支持K8s CSI
- librbd协议兼容
- 缓存元数据信息，加速IO处理
- 支持客户端热升级（NEBD服务，可选）
- 支持QoS，多挂载
- 支持条带卷，优化大IO顺序读写性能

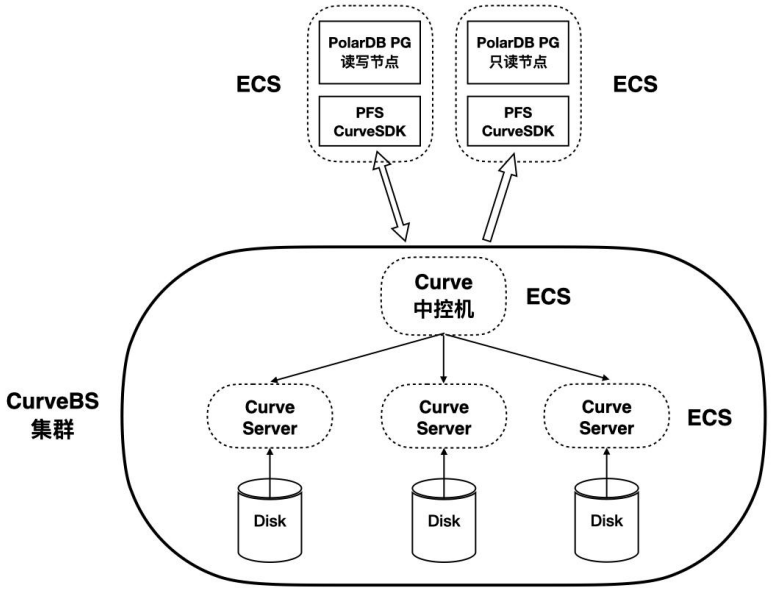


目录

• Curve整体介绍
• Curve块存储架构介绍
• Curve块存储云原生数据库实践
• 后续规划

Why 存算分离

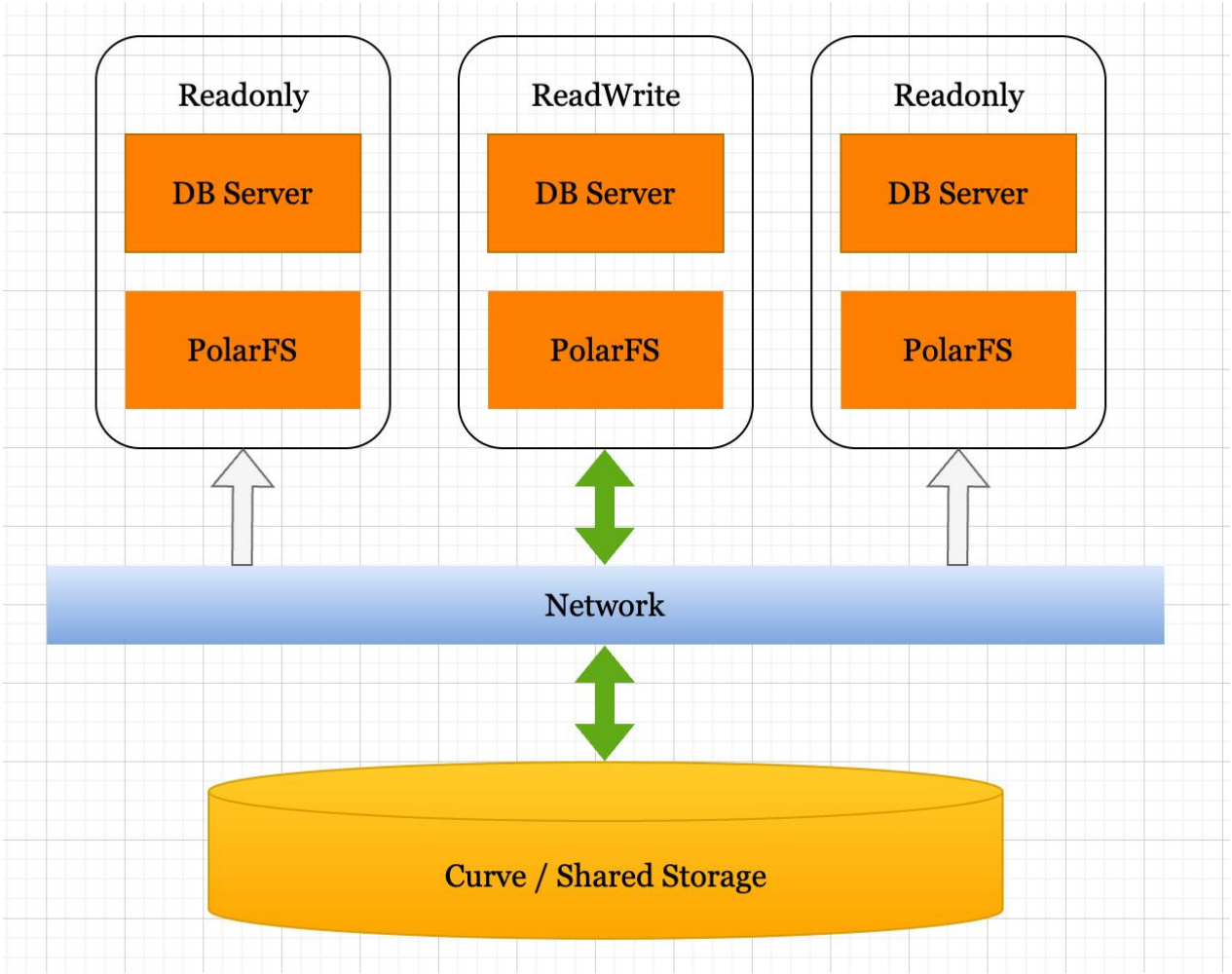
- 存算一体架构下，资源扩展不够灵活、资源争抢、故障耦合等
- “分库分表”中间件无法很好的适应互联网大规模和复杂的业务需求
- 云上的存算分离数据库无法云下部署（如各大公有云厂商的RDS产品）
- 通过存算分离架构解决传统数据库痛点问题，如：主从复制延迟大、数据备份代价大、节点重建时间长、节点扩容/扩展弹性小、资源无法高效使用、客户端驱动能力弱等



存算分离对存储系统的要求

✓极致的性能（低延迟）、最大程度发挥硬件的性能
✓数据一致性、可靠性高
✓高可扩展，性能和容量可线性扩展
✓各种故障场景下表现稳定，IO时延波动小
✓部署简单，运维操作门槛低，常见故障自愈能力强，监控指标丰富易用
✓开源存储项目可满足上述要求的极少，Curve块存储是其中之一

PolarDB-FileSystem

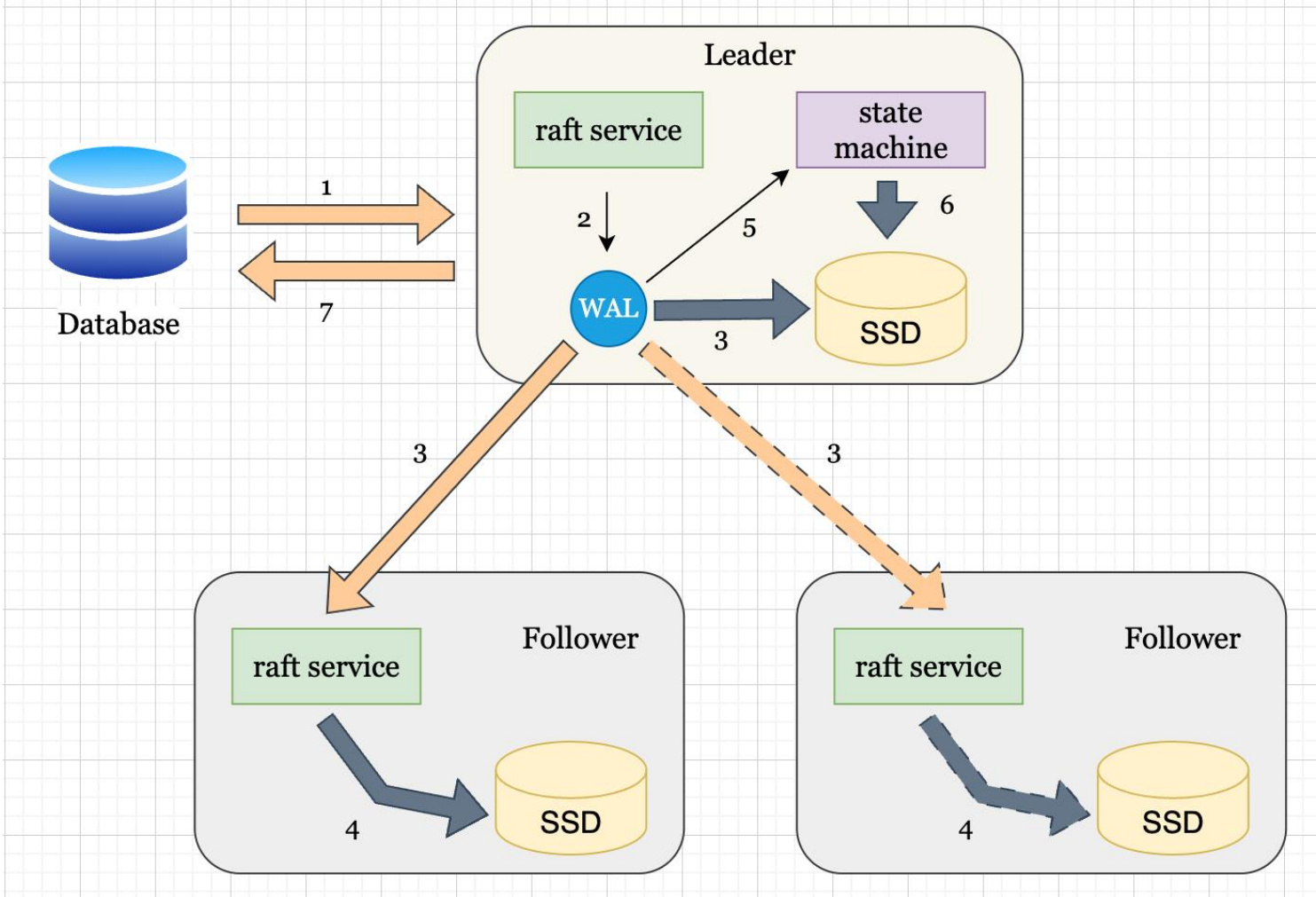


PFS for CurveBS

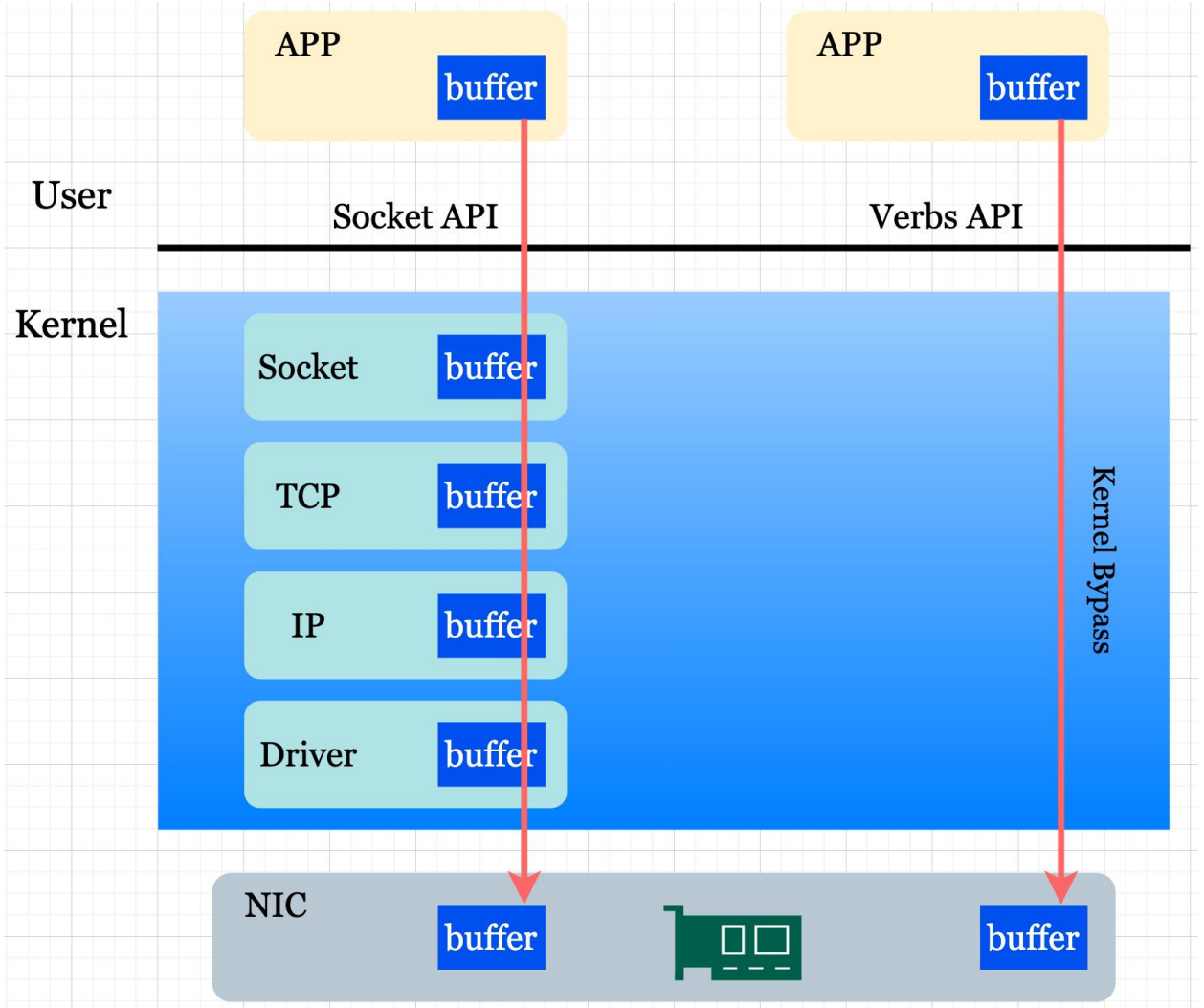
- PFS适配&优化
  - ~2000 LoC
  - PFS使用无锁工作队列，降低shared memory轮询线程开销
  - 使用unix socket作任务通知
  - 优化pfs_lseek锁，对多线程更加友好；支持大于4M的读写；减少pfs journal的补偿读
- Curve优化
  - 支持共享挂载
  - IO fence
    - 数据库异常情况下的主从切换是独立于存储的，为了保证数据的一致性，通过IO fence来隔离旧的主节点的inflight IO请求
    - 具体的实现为，在数据库只读节点切换为读写节点时，会重新挂载一次PFS，在挂载的过程中，升级一次版本号

https://github.com/opencurve/PolarDB-FileSystem/tree/curvebs_sdk_devio

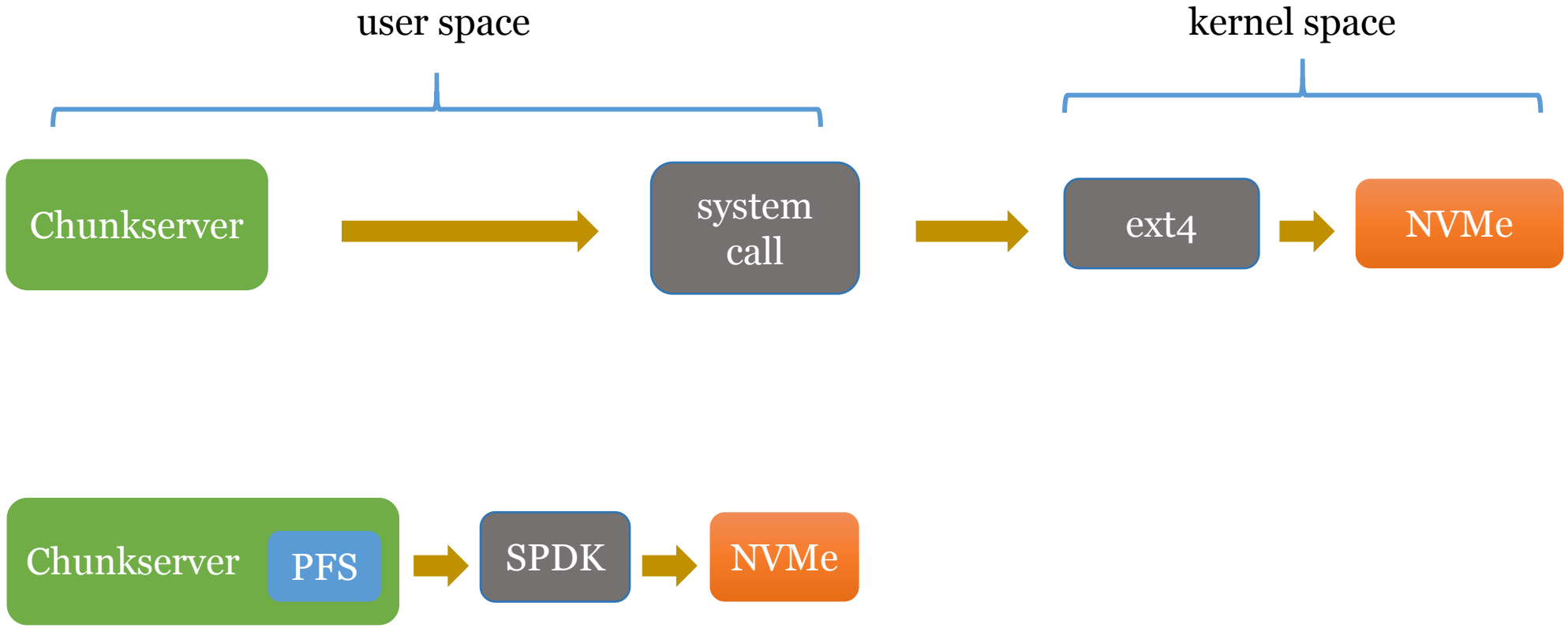
极致性能优化


Leader
raft service
state machine
WAL
SSD
Database
1
2
3
4
5
6
7
Follower
raft service
SSD
Follower
raft service
SSD

极致性能优化 - RDMA


APP
buffer
APP
buffer
User
Socket API
Verbs API
Kernel
Socket
buffer
TCP
buffer
IP
buffer
Driver
buffer
Kernel Bypass
NIC
buffer
buffer

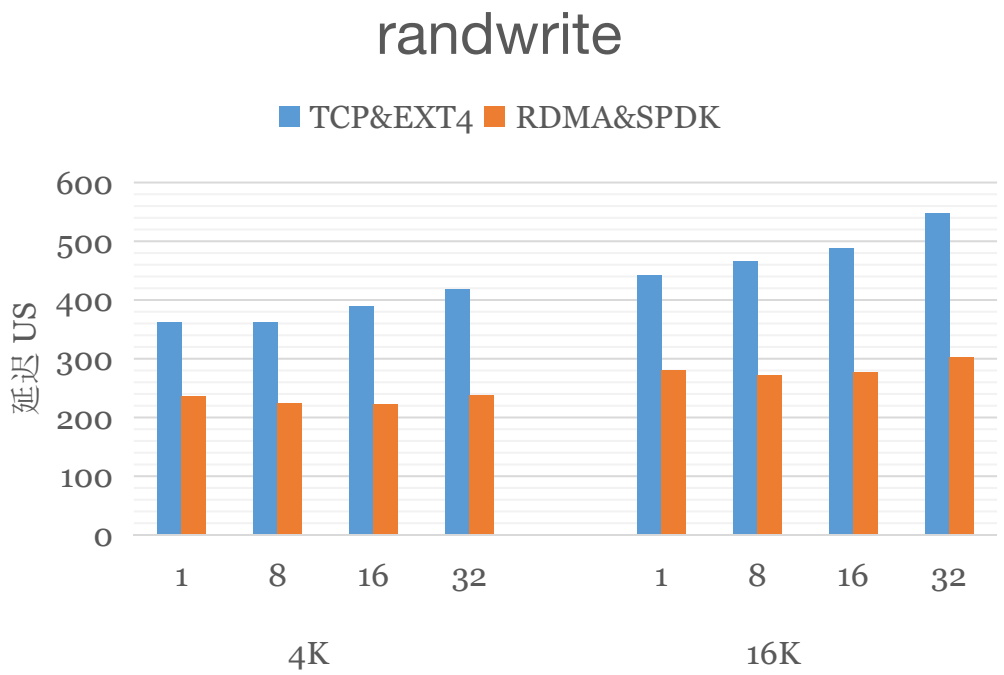
极致性能优化 - NVMe & SPDK

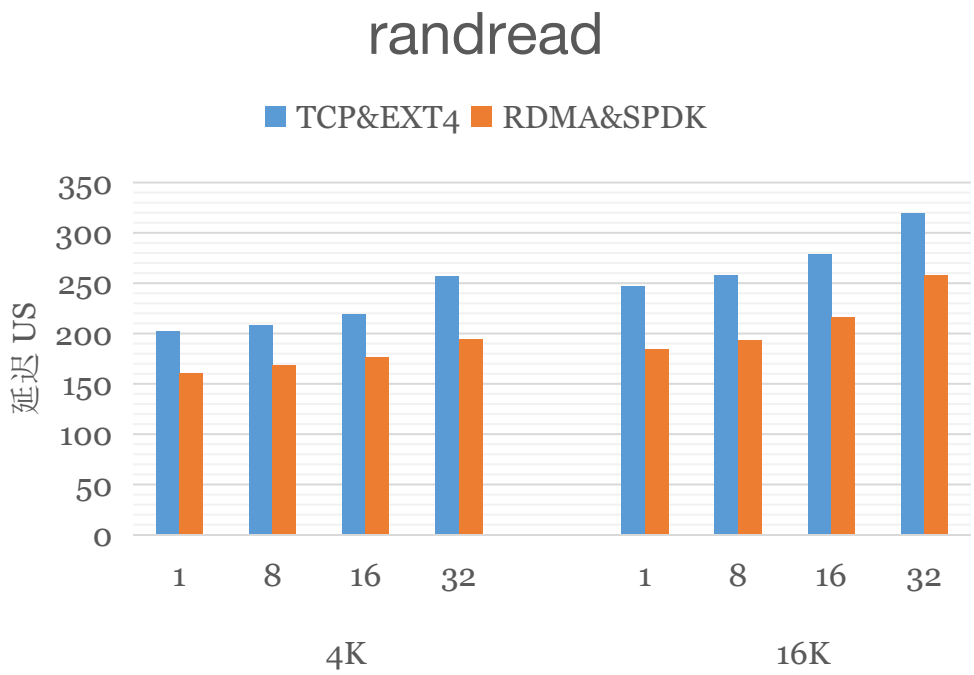


极致性能优化 - RDMA & SPDK

- RDMA基于UCX对BRPC进行适配，利用DCQCN算法，避免拥塞，提高网络性能
- PFS增加SPDK驱动，管理NVMe
- PFS移除daemon，直接使用pfs core api，减少 ~100us 的开销
- 增加pfs_readv/writev接口，适配BRPC IOBuf
- 增加pfs_readv/writev_dma接口
- RDMA内存从DPDK大页内存，实现整条链路的zero copy

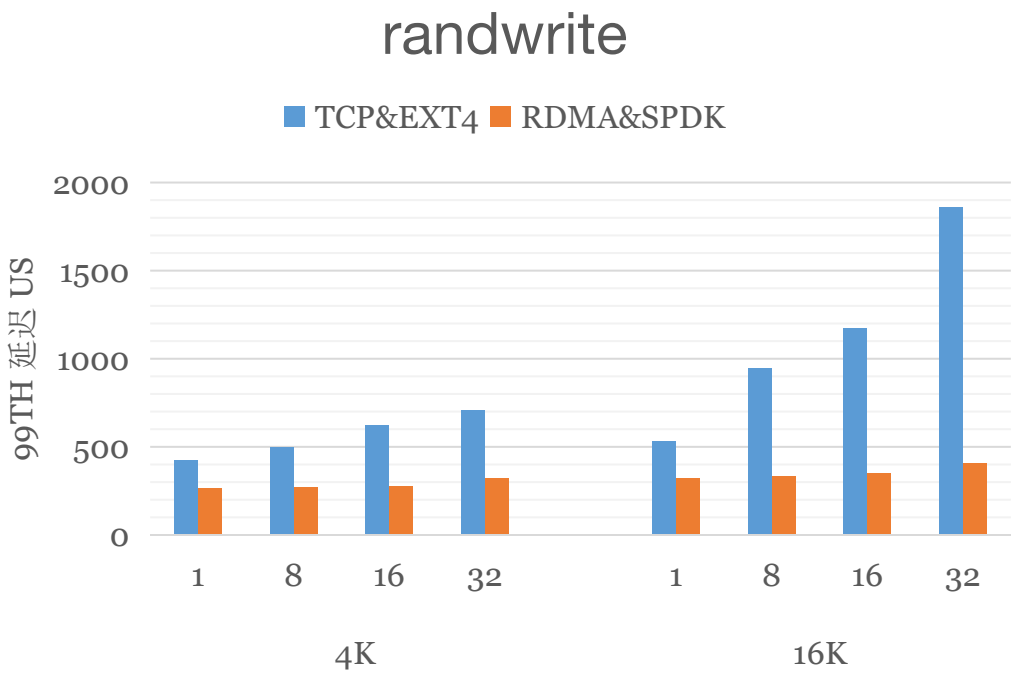
极致性能优化 - RDMA & SPDK

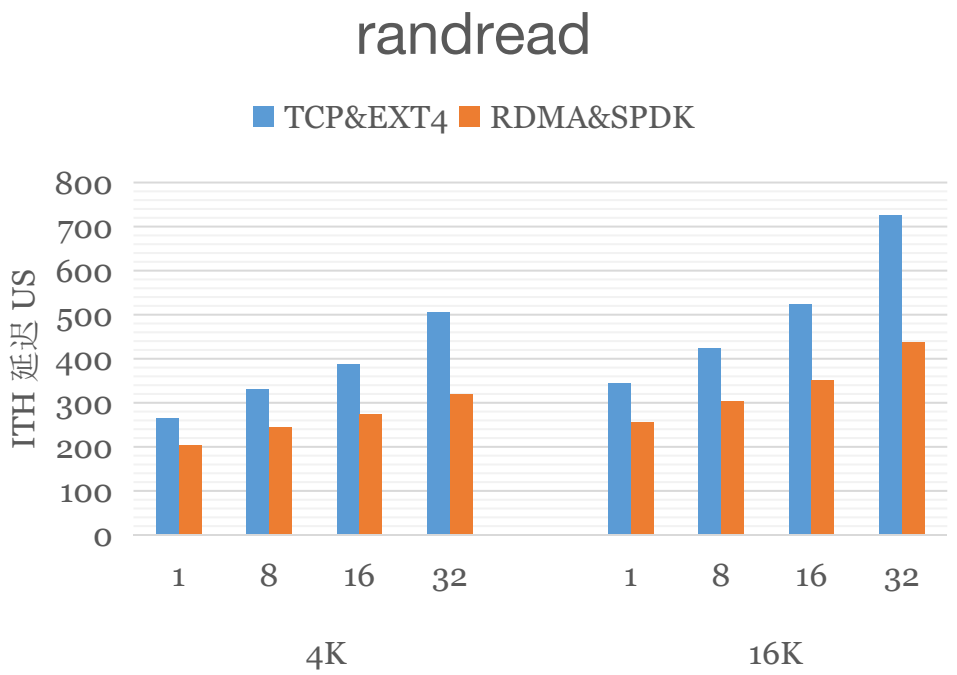




平均延迟

极致性能优化 - RDMA & SPDK

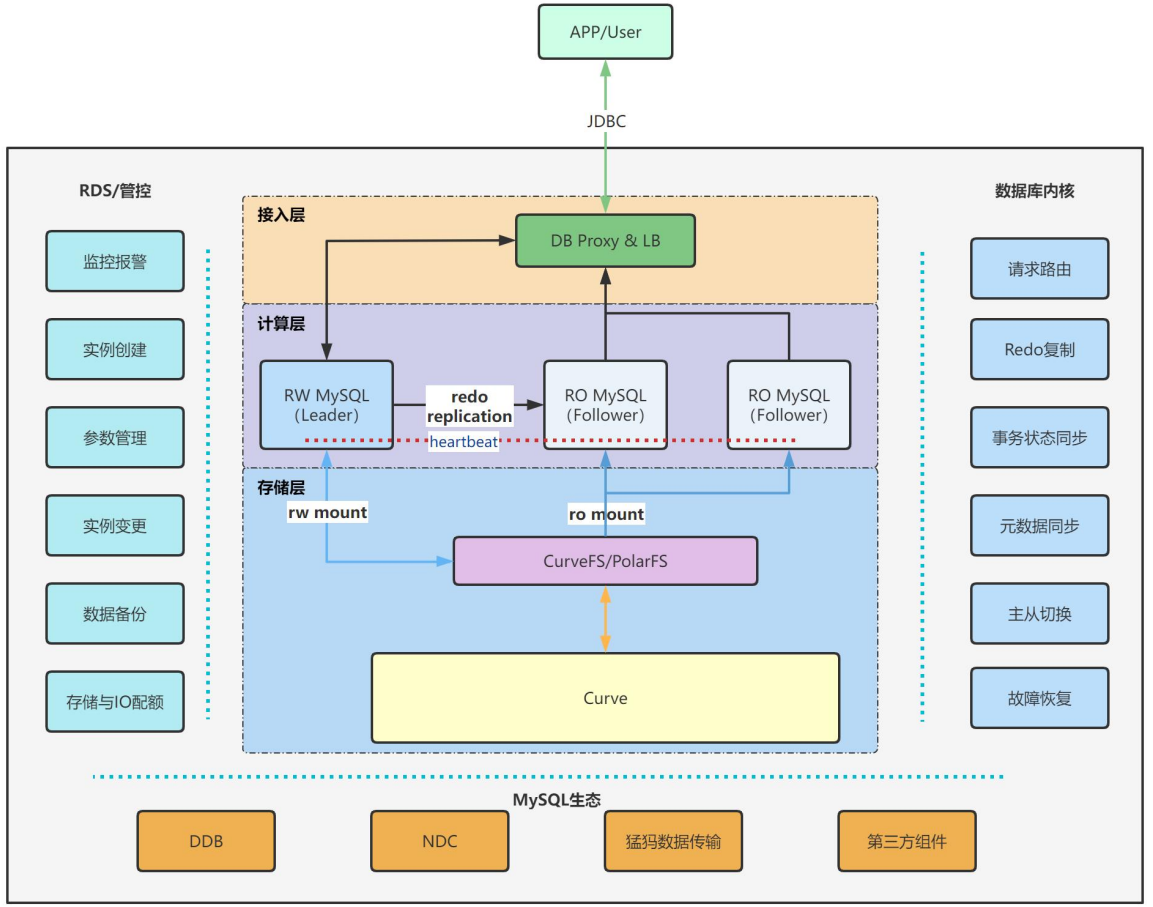




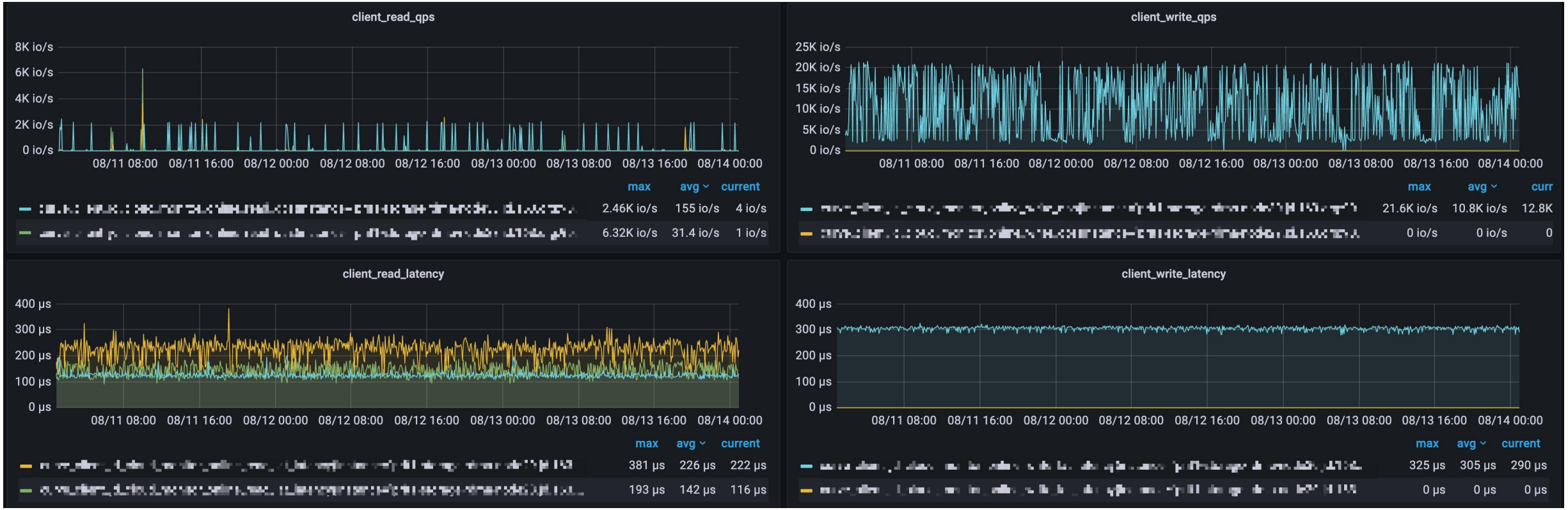
99th延迟

MySQL + PFS + Curve

- 基于共享存储且完全兼容MySQL生态的云原生数据库产品
- 实现基于共享文件系统（PolarFS）的redo主从复制，从库MVCC读等核心能力
- 基于raft实现了集群管理，支持计算节点动态添加、删除和主从自动切换等
- 通过redo拆分和redo io异步化改造数倍提升了在共享块存储情况下的事务提交性能
- 生产环境上线

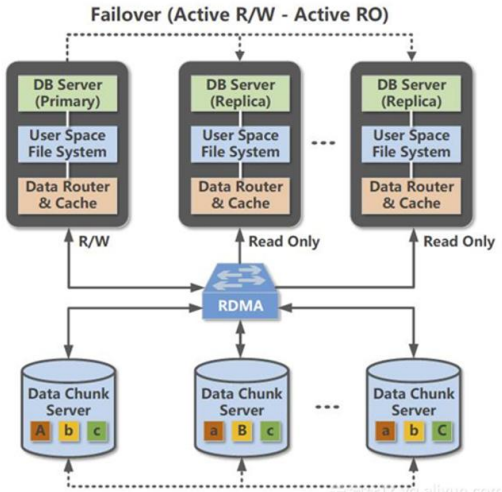


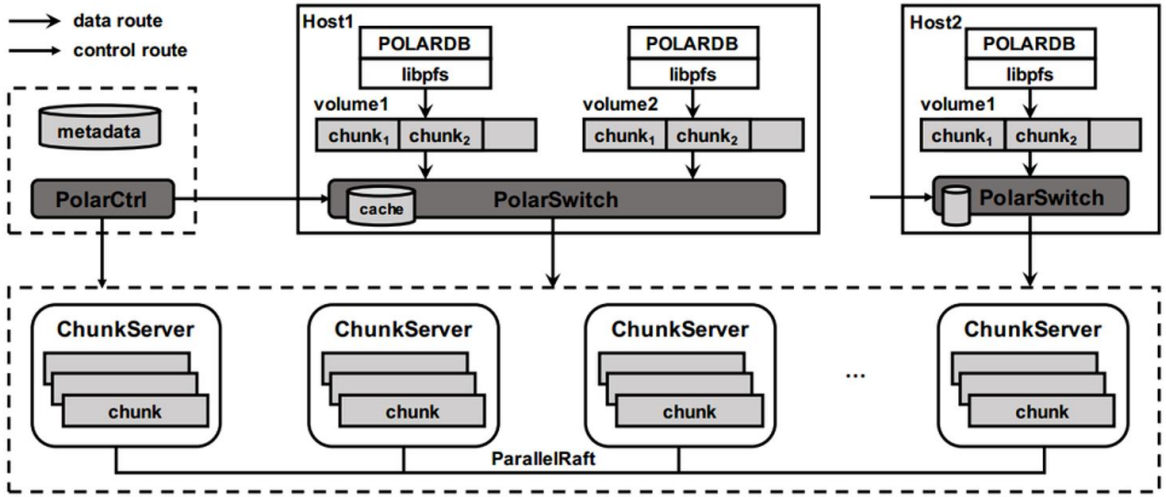
MySQL + PFS + Curve



PolarDB for PG + PFS + Curve

- Curve是目前PolarDB开源社区唯一原生适配的Shared-Storage方案，也是PolarDB社区的生态合作伙伴
- Curve相比其他开源存储系统，如Ceph具备较好的性能优势和时延稳定性
- RDMA & SPDK 版本对比TCP+EXT4
  - benchmarkSQL每分钟事务数提升20%
  - pgbench延时降低29%，TPS提升40%





https://apsaradb.github.io/PolarDB-for-PostgreSQL/zh/deploying/storage-curvebs.html

目录

- Curve整体介绍
- Curve块存储架构介绍
- Curve块存储云原生数据库实践
- 后续规划

CurveBS Roadmap

- 内部MySQL云原生数据库持续优化，大压力、超大容量场景验证
- PolarDB完善RDMA&SPDK版本适配及性能调优
- RDMA优化，根据网络质量自动fallback TCP；读性能优化(incast)
- 自研本地存储引擎，避免双写对大IO的影响
- 整体架构优化，RPC、Raft、2副本 ...

THANKS

扫码关注Curve公众号