
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

开源分布式存储系统Curve在云原生数据库领域的实践

网易杭州研究院

Curve社区PMC

王盼

目录

- Curve项目介绍
- 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项目介绍

- 项目愿景

打造云原生、高性能、稳定易运维的开源分布式存储系统

云原生	高性能	易运维
• 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（开发中）	• 支持RDMA • 支持SPDK • 支持多级缓存 • 数据/元数据性能可水平扩展 • …	• 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 • …

Curve项目介绍

- 社区情况

开源生态

操作系统 芯片 数据库 云原生 AI 训练 大数据

OpenAnolis 龙蜥社区 | Kunpeng | PolarDB | openstack | TensorFlow | elastic
OpenEuler | Phytium飞腾 | MySQL | kubernetes | PyTorch | kafka
KYLIN 银河麒麟 | 长江存储 YANGTZE MEMORY | PostgreSQL | ZStack | 飞桨
HYGON 中科海光 | TDengine | esage 易思捷
PLiOPS

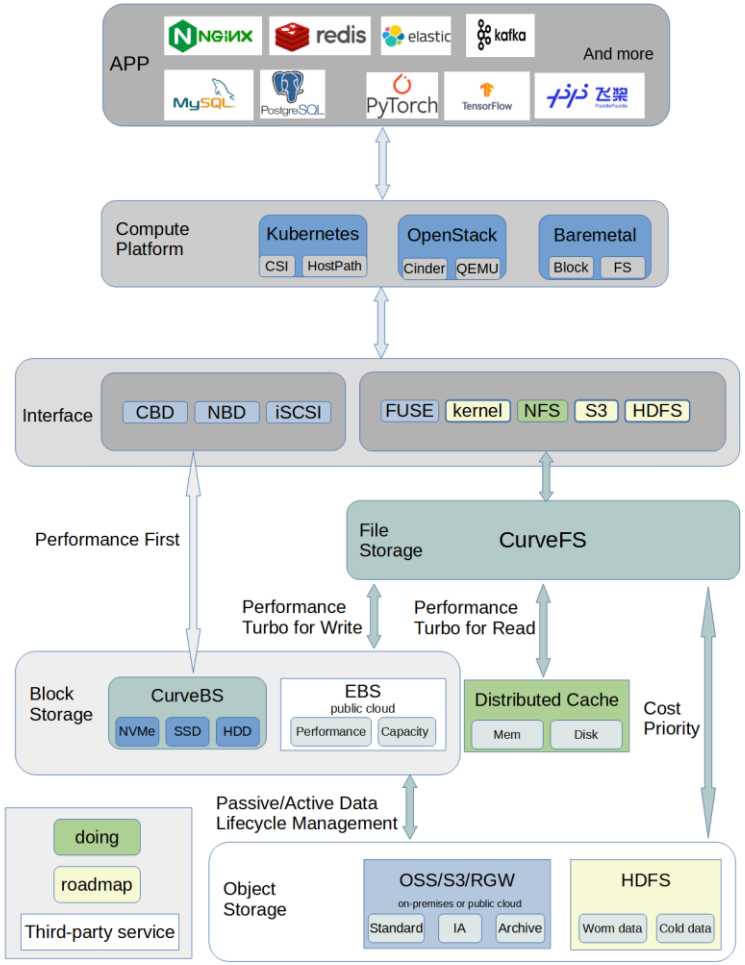
官方认证

- ✓ **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** 沙箱项目，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

- 主页：https://opencurve.io/
- 论坛：https://ask.opencurve.io/
- Github：https://github.com/opencurve/curve
- 公众号：OpenCurve
- 用户群：添加微信号OpenCurve_bot可邀请加群
- Slack：workspace cloud-native.slack.com，channel #project_curve

Curve项目介绍

- 主要应用场景
  - 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

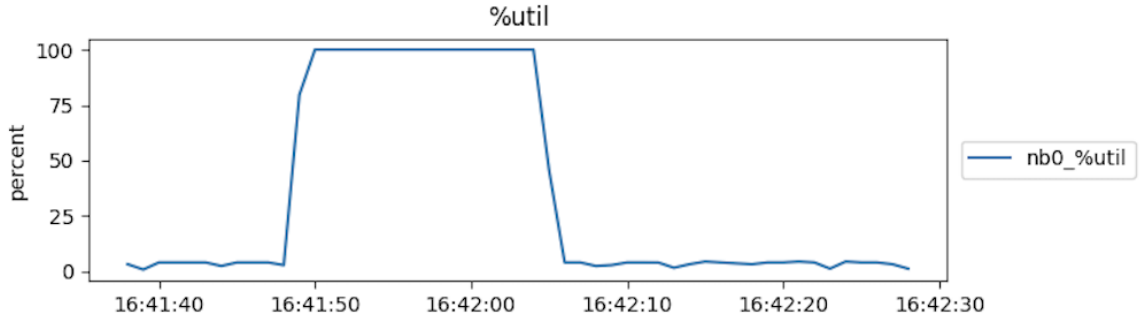
网易严选 FUSION SmartMore 思谋 网易云音乐 有道 youdao LOFTER 万方电子 网易游戏 创云融达 网易云信

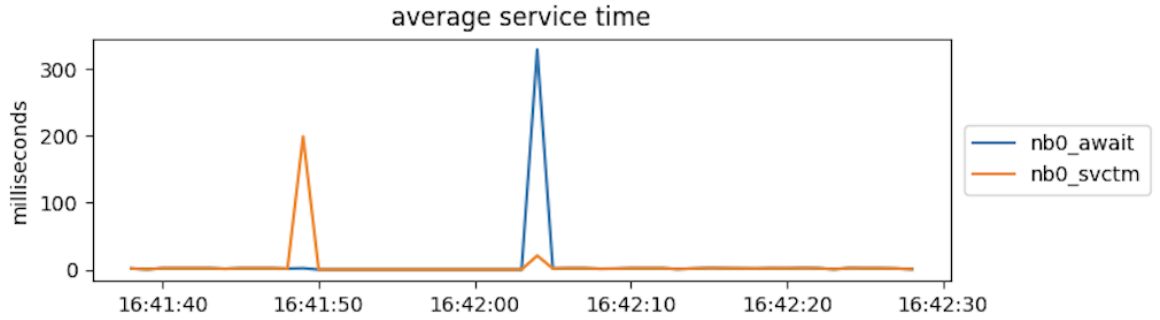
Curve项目介绍

更稳定 慢盘
更稳定 网络丢包 5%
更稳定 网络丢包 10%
更稳定 机器宕机

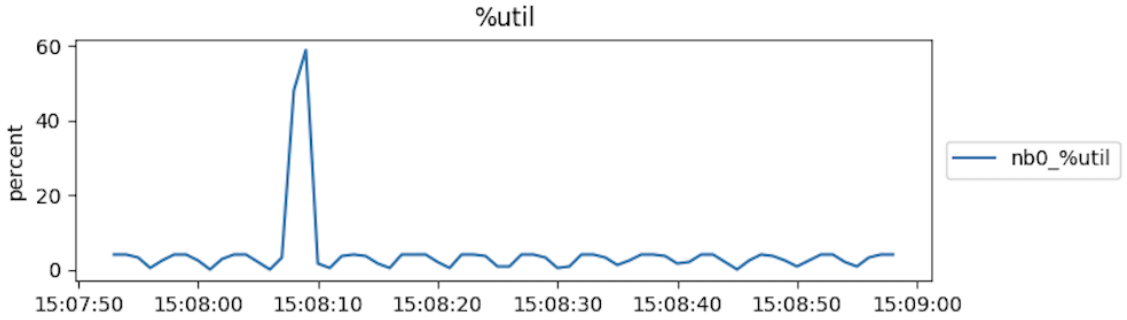
Ceph
Ceph
Ceph

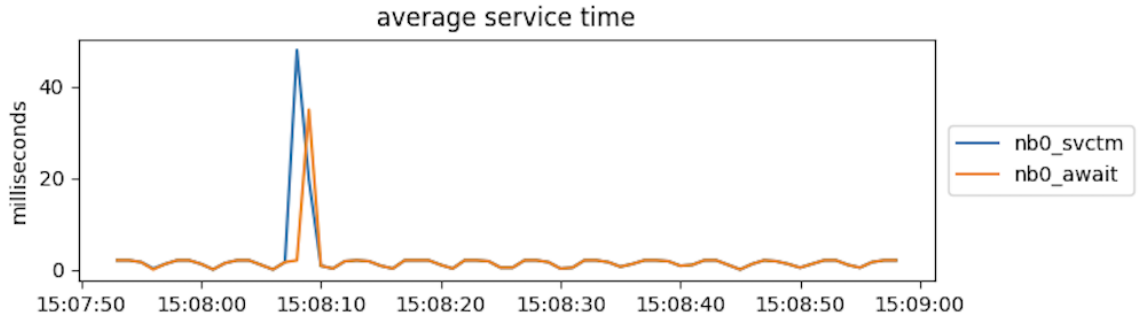
Curve块存储IO时延更稳定，更适合数据库场景





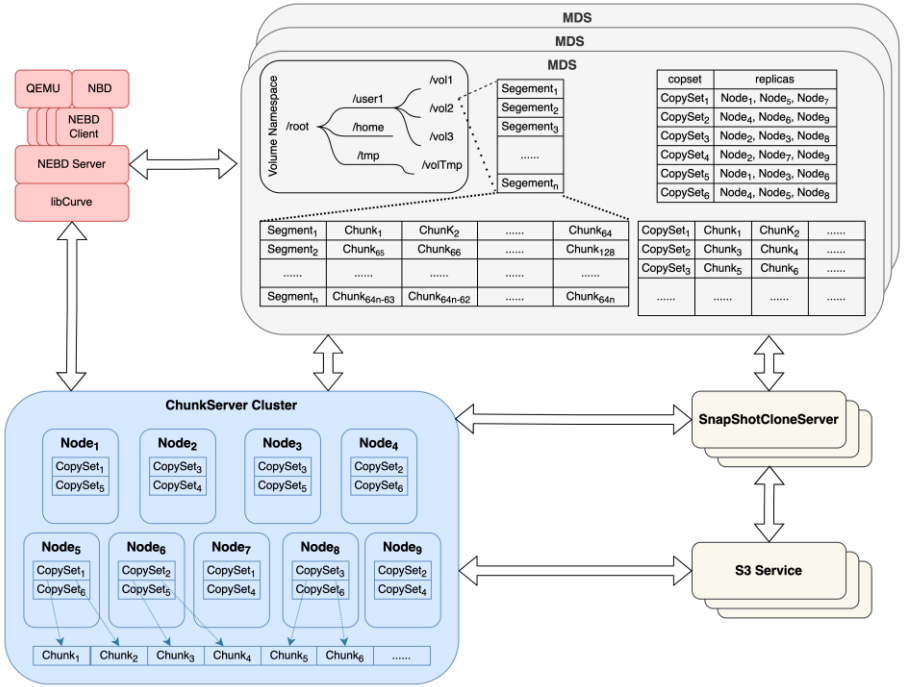
Curve
Curve
Curve





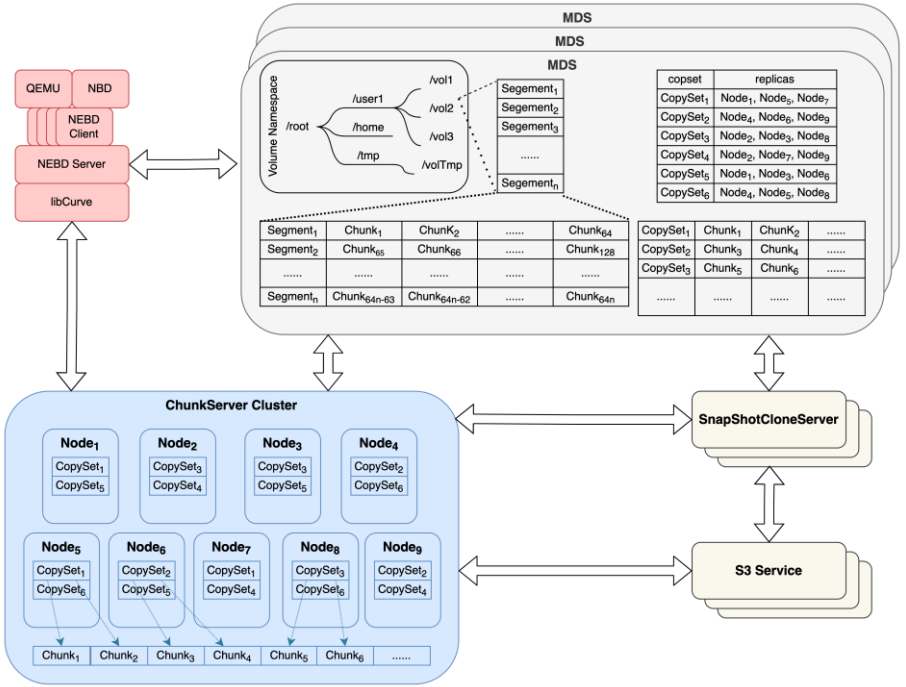
Curve块存储架构介绍

- 元数据管理 MDS+etcd
  - ✓ 管理维护集群topo结构、节点上下线、卷namespace、卷到chunk的映射关系
  - ✓ 卷由chunk组成，卷到segment再到chunk的映射
  - ✓ copyset组为单位分配/容量均衡/负载均衡
  - ✓ chunk到copyset的映射
  - ✓ MDS主备模式，使用etcd选主，元数据存储在etcd
  - ✓ 通过一次性分配一个segment及缓存映射关系来减少访问MDS次数，提升IO性能



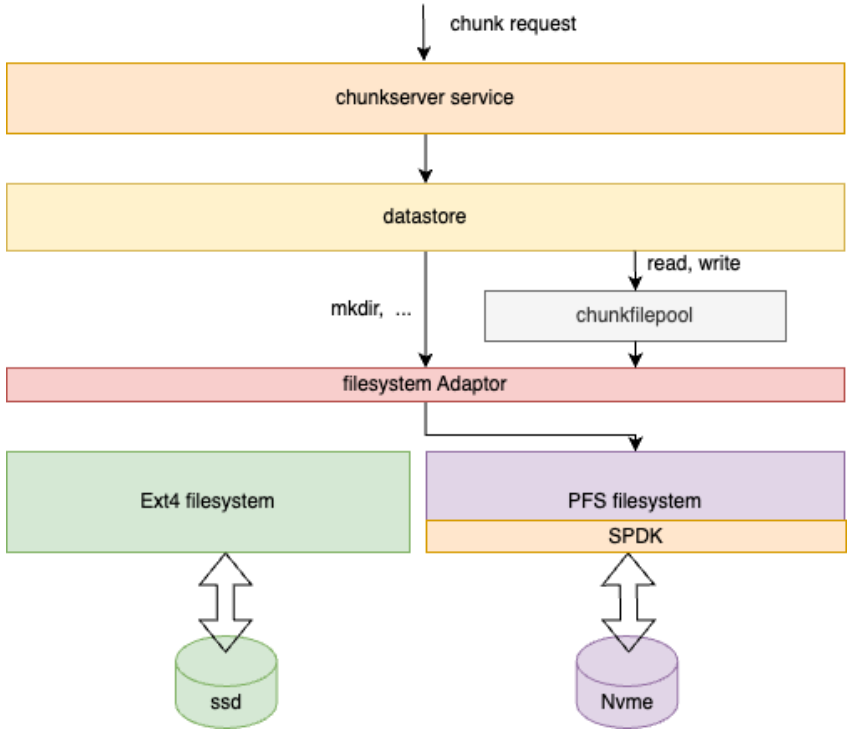
Curve块存储架构介绍

- 数据存储引擎 （ChunkServer+EXT4）
  - ✓ 使用Raft协议管理IO的读写与同步（brpc+braft）
  - ✓ 每个chunkserver通常负责一块盘，通常分配100个copyset
  - ✓ 使用ext4文件系统管理chunkfile
  - ✓ 使用chunkfilepool 降低写放大
  - ✓ 使用brpc的zerocopy能力减少CPU开销（rdma到spdk的zerocopy正在开发中）
  - ✓ 读支持绕过raft层进行加速
  - ✓ 基于麦洛斯ucx库构建brpc的rdma协议支持



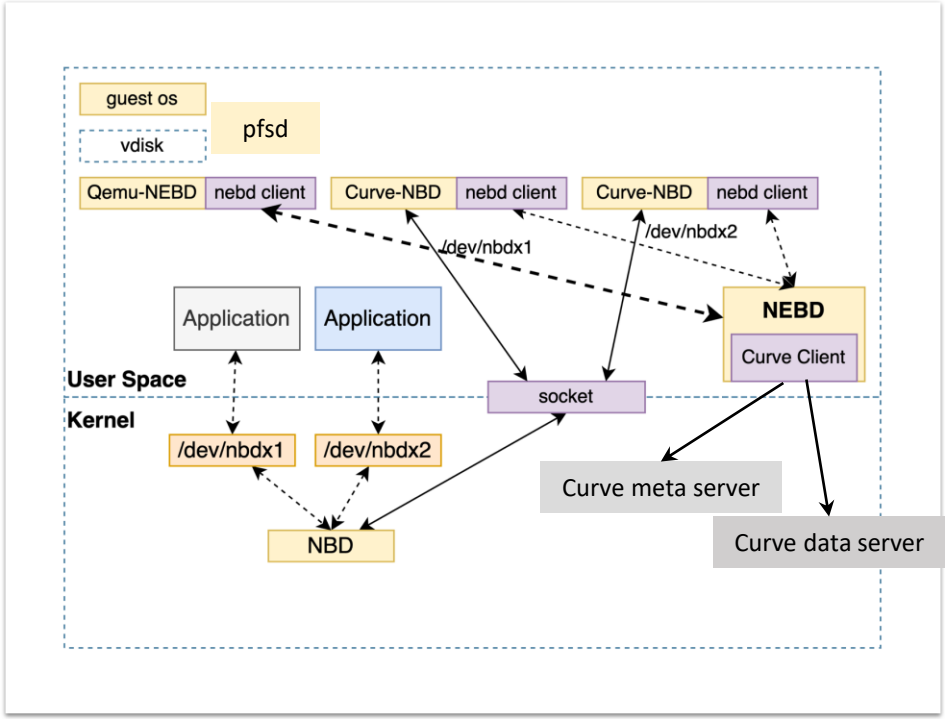
Curve块存储架构介绍

- 数据存储引擎（ChunkServer+SPDK）
  - ✓ SPDK存储引擎基于PolarFileSystem实现
  - ✓ 在chunkserver datastore和chunkfilepool之下实现了一层filesystem Adaptor适配层，向下分别适配PolarFileSystem和Ext4 filesystem
  - ✓ 在PolarFileSystem中基于Intel SPDK实现用户态的nvme驱动程序，通过该驱动直接读写nvme裸盘，从而实现高性能的nvme读写
  - ✓ PolarFileSystem文件系统接口实现了DMA方式和非DMA的两套接口：
    - DMA方式用于核心IO数据流，将从网络上读下来的数据从用户态直接传输到NVME设备，从而实现零拷贝的高效数据传输。这部分需要IO对齐
    - 非DMA方式用于非核心的元数据读写，不需要IO对齐，以提供对元数据方便的读写接口



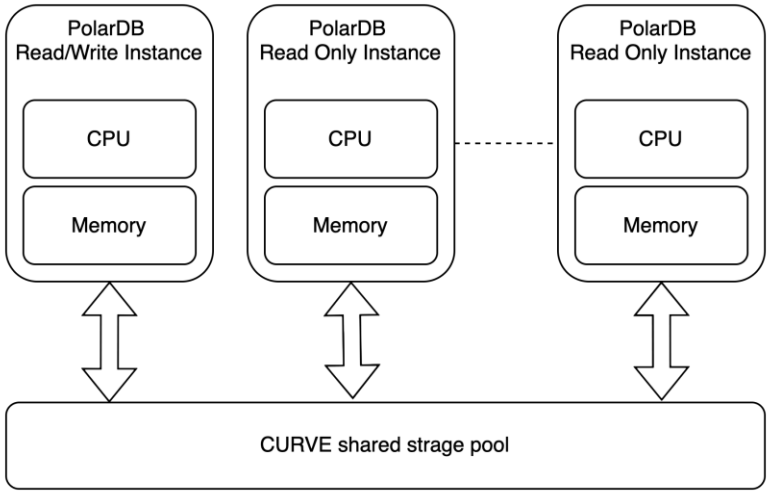
Curve块存储架构介绍

- Curve客户端
  - ✓ 使用brpc与meta cluster和data cluster通信（支持基于ucx的rdma）
  - ✓ 支持同步、异步IO接口，负责IO拆分和跟踪
  - ✓ 支持元数据缓存提升性能
  - ✓ 支持客户端SDK热升级（NEBD服务，性能优先场景下可移除）
  - ✓ 支持QoS、多挂载、IO Fence
  - ✓ 支持 data stripe 增大并发
  - ✓ 支持NBD、QEMU、iSCSI、PFS，支持k8s CSI、openstack cinder驱动



Curve块存储架构介绍

- 客户端架构（PFS + libcurve）
  - ✓ 一写多读，读节点实时可见数据
  - ✓ 类posix文件系统API
  - ✓ 日志文件系统，不需要fsck
  - ✓ PFS使用支持使用libnebd（支持sdk热升级）或libcurve（性能更好）的API与CurveBS交互
  - ✓ Curve块存储集群提供多副本冗余



Curve块存储云原生数据库实践

- 为什么要做存算分离数据库
  - ✓ “分库分表”方案已经无法满足业务复杂度和大规模需求
  - ✓ 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）
  - ✓ 分布式中间件方案已经落后时代：扩容操作效率低、资源投入大、生效时间长；表结构变更效率低、耗时长、影响稳定性；复杂SQL支持能力弱；数据一致性、高可用保障能力差等等问题无法彻底解决
  - ✓ 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

Curve块存储云原生数据库实践

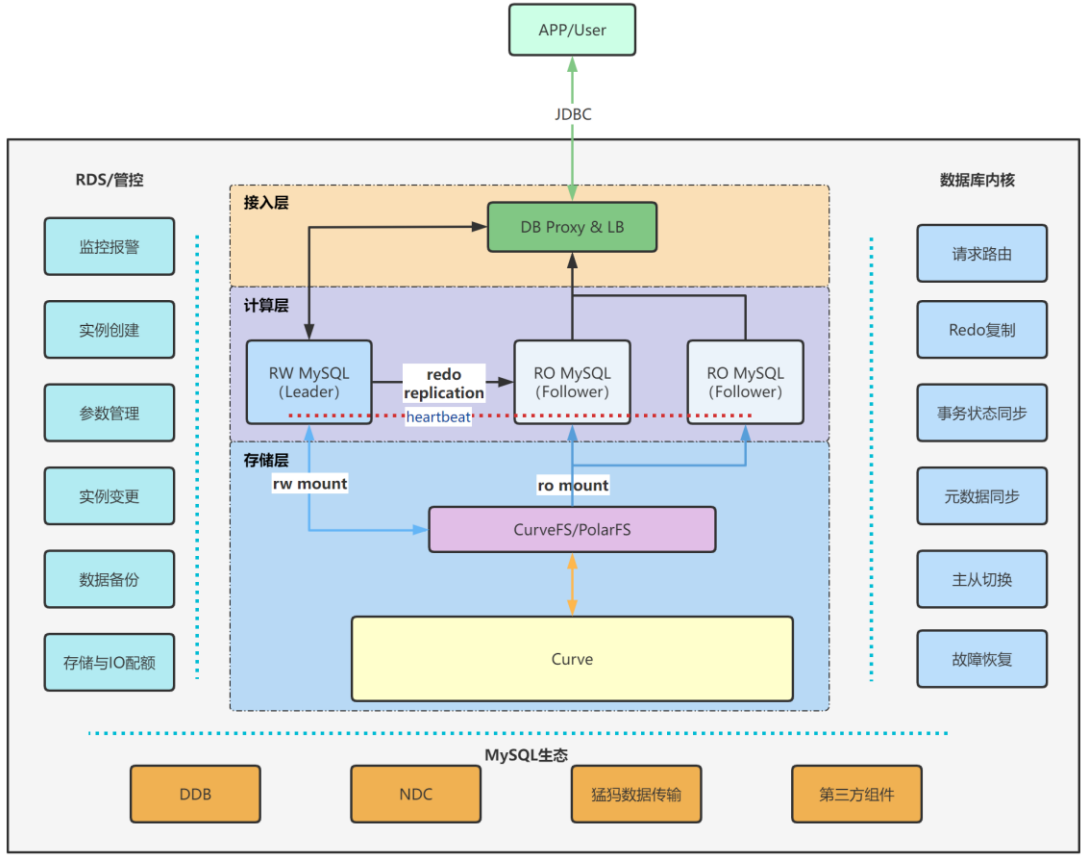
- 存算分离对底层存储系统的要求
  - ✓ 各种故障场景下表现稳定，IO时延波动小
  - ✓ 数据一致性、可靠性高
  - ✓ 高性能、低时延，可最大限度发挥硬件能力
  - ✓ 高可扩展，性能和容量可线性扩展
  - ✓ 部署简单，运维操作门槛低，常见故障自愈能力强，监控指标丰富易用
  - ✓ 支持软硬件异构能力强，占用资源少
  - ✓ 开源存储项目可满足上述要求的极少，Curve块存储是其中之一

Curve块存储云原生数据库实践

- 为什么选择PFS + Curve
  - ✓ Auroa/PolarDB等云原生RDS为代表的shared disk
  - ✓ 以Spanner为代表的shared nothing
  - ✓ Curve作为独立的底层存储项目，自然要选择shared disk架构方案
  - ✓ PFS、Curve在性能和可靠性有保障，历经数年生产环境考验
  - ✓ PFS基于块设备进行读写，Curve块存储适配PFS非常简单易行，只需要很薄的适配层即可

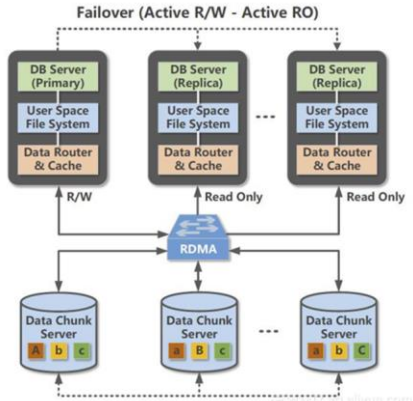
Curve块存储云原生数据库实践

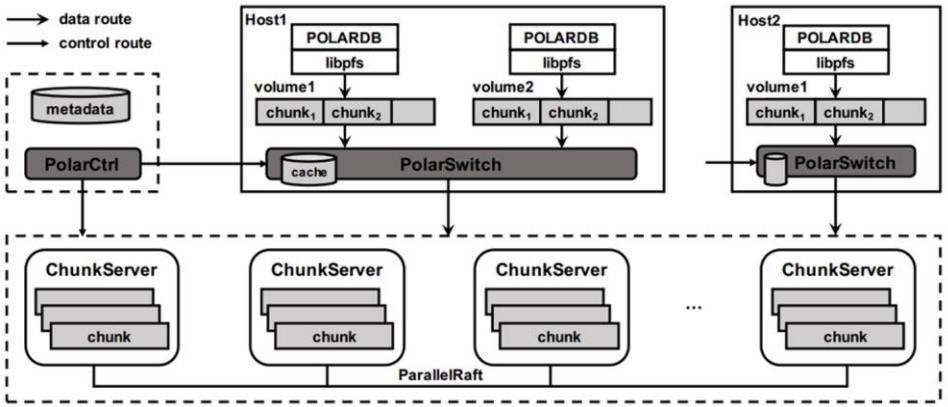
- MySQL + PFS + Curve
  - ✓ 基于共享存储且完全兼容MySQL生态的云原生数据库产品
  - ✓ 当前研发进展：
    - 实现基于共享文件系统（polarfs）的redo主从复制，从库mvcc读等核心能力
    - 基于braft实现了集群管理，支持计算节点动态添加、删除和主从自动切换等
    - 适配了外部xa事务，xtrabackup物理备份工具
    - 通过redo拆分和redo io异步化改造数倍提升了在云盘等高io延迟情况下的事务提交性能
    - 业务灰度测试和上线中



Curve云原生数据库实践

- PolarDB for PG + PFS + Curve
  - ✓ Curve是目前polardb开源社区唯一原生适配的share-storage方案，也是ploardb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
    - benchmarkSQL每分钟事务数提升39%
    - pgbench延时降低21%，TPS提升26%
  - ✓ 当前研发进展：完成功能适配及部署适配，持续性能调优中





Curve云原生数据库实践

- 云原生数据库场景下的存储系统优化实践
  - ✓ 使用ucx对brpc改造，支持rdma，ucx具有产品级稳定性
  - ✓ rdma实现零copy，网路传输绕过os内核，节省时延
  - ✓ spdk支持直接读写nvme，绕过os内核，减少系统调用和跨内核边界数据copy
  - ✓ spdk与rdma结合，实现从网络到nvme读写调用栈 cpu 数据零copy, 提供低时延和大吞吐
  - ✓ pfs使用无锁工作队列，降低shared memory轮询线程，同时使用unix socket作任务通知，数据传输依然是使用shared memory，极大降低cpu使用率，避免100%空转造成的浪费
  - ✓ 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云原生数据库实践

- ✓ 支持docker化部署pfs+libcurve，简化部署使用
- ✓ Curve本身的优化增强：
  - 支持卷级别IO Fence，leader切换时不会使得inflight io损坏数据
  - Raft Apply data不执行sync，提升数据写入性能

测试环境为1台客户机，3台存储服务器，每台配置如下：
CPU：Intel(R) Xeon(R) Gold 5220R CPU @ 2.20GHz 96 核心
RAM：256GiB
网卡：Mellanox Technologies MT27710 Family [ConnectX-4 Lx] x2
NVMe: SAMSUNG MZQL21T9HCJR-00B7C__1 1.8T x4

4K 16 随机随机写

QD	RDMA-IOPS	fio clat(us) avg	fio clat(us) 90	TCP-IOPS	fio clat(us) avg	fio clat(us) 90
1	3124	316	326	2038	486	545
8	21.7k	364	343	14.5k	548	586
16	36.0k	440	392	26.1k	609	676
32	55.9k	568	652	44.6k	713	832
64	69.9k	911	1942	65.3k	974	1467
96	76.5k	1249	3097	76.2k	1254	2278
128	81.1k	1575	4113	82.0k	1556	3064

4K 16 随机随机读

QD	RDMA-IOPS	fio clat(us) avg	fio clat(us) 90	TCP-IOPS	fio clat(us) avg	fio clat(us) 90
1	12.0k	80	88	4954	198	251
8	105k	79	85	49.8k	158	237
16	79.7k	98	116	34.1k	231	318
32	127k	122	143	48.3k	327	400
64	124k	236	253	78.4k	404	611
96	125k	506	510	75.6k	843	1516
128	133k	958	955	84.8k	1505	3195

Curve云原生数据库实践

- **rdma和tcp的对比**
  - ✓ 默认参数情况下，基于rdma的curve，rc3基线性能比tcp好38.7%（7755.53 vs 5591.73）
  - ✓ 调参情况下（增加innodb_redo_write_max_size和innodb_redo_write_ahead_size到16KB），rdma比tcp性能好72.3%（21173.73 vs 12285.96）
  - ✓ redo 4分片+4k对齐场景，基于rdma的curve性能比tcp性能好37.6%（13313.33 vs 9676.92）

Curve云原生数据库实践

- 后续规划
  - ✓ msyql方面：支撑内部落地业务、配合外部开源（待定）
  - ✓ polardb方面：完成arm一体机项目，社区用户测试及落地
  - ✓ Curve自身：
    - 持续的性能优化及功能增强，如rdma自动fallback tcp
    - spdk的性能瓶颈分析及优化
    - raft性能优化
    - 更高性能硬件选型及适配等

THANKS

扫码关注Curve公众号