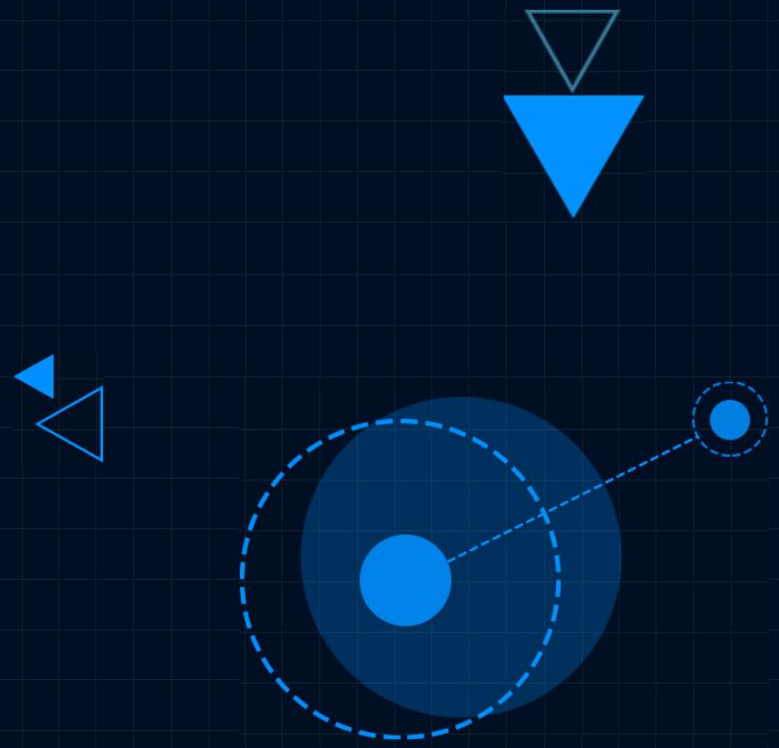
CURVE分布式开源存储系统架构

向东

架构师 网易杭州研究院



提纲

CURVE简介

CURVE Block Strorage (CurveBS)

CURVE File System (CurveFS)

Roadmap

CURVE简介

- CURVE是分布式存储系统

  高性能

  容易运维

  云原生
- CURVE由两部分组成

  CURVE Block Storage (CurveBS)

  　　CURVE高性能分布式块存储

  CURVE File System (CurveFS)

  　　CURVE分布式文件存储，(底层支持S3、块存储、分布式块存储)

CEPH存储系统的问题

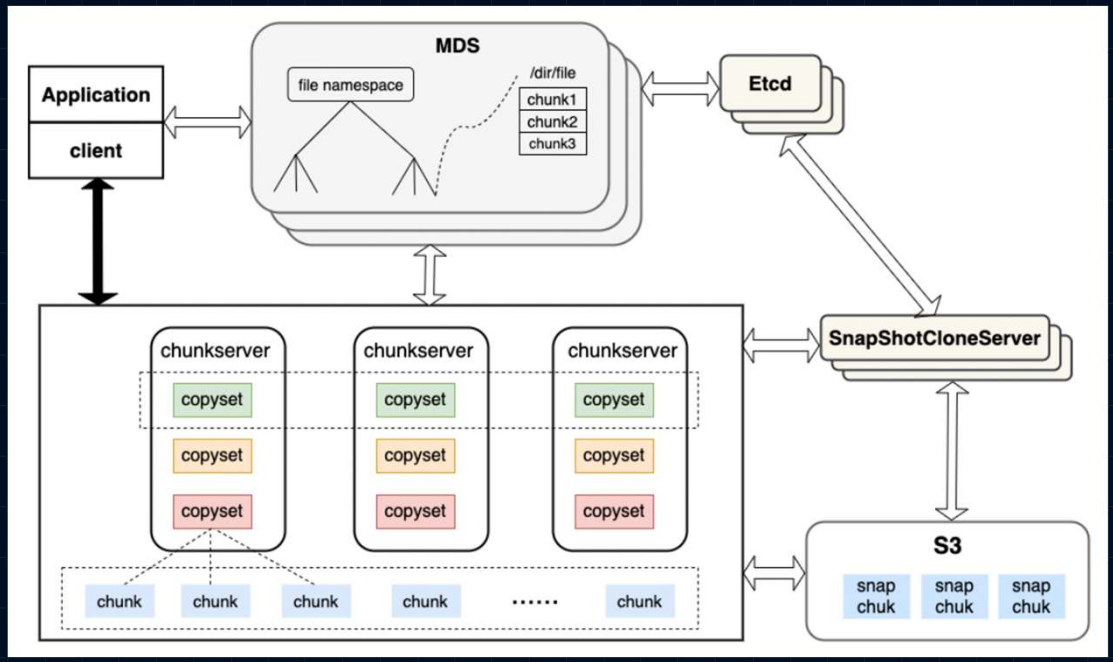
- CEPH在开发方面的劣势
  代码量庞大，比较难做到自主可控，200W+行代码
  比较难做到IO链路数据监控/分析
- CEPH在性能方面的劣势
  在底层硬件出现故障的时，IO出现抖动
- CEPH在运维方面的劣势
  扩容/出现慢速磁盘/更换磁盘的都会引起IO抖动

CURVE的现状

- CURVE 2020年7月16日开源
  https://github.com/opencurve/curve
  https://www.opencurve.io/
- 版本更新 (半年1个大版本、1～2个月1个小版本)
  v1.2.0 支持Qos、Discard、数据静默检查
  v1.3.0 增加了部分性能优化
  详细版本更新内容：https://github.com/opencurve/curve/releases
- CurveBS在公司内部广泛应用
  nbd方式来支持块存储、虚拟机
  CSI方式支持容器
- CurveFS目前开发中

CurveBS框架

- 元数据管理 MDS +Etcd
  卷由chunk组成，卷到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
- chunkServer负责IO的读写与同步
  每个chunkServer负责一块盘
- SnapShotCloneServer负责snapshot，支持保存到S3对象存储上



CurveBS设计与CEPH对比

- 开发框架

  使用M:N的线程调度框架在多核服务器上提供更好的扩展性和更高的性能

  Chunk File Pool降低元数据开销，接近直接读写裸盘性能

  无锁队列

  零拷贝设计节省cpu开销

CurveBS设计与CEPH对比

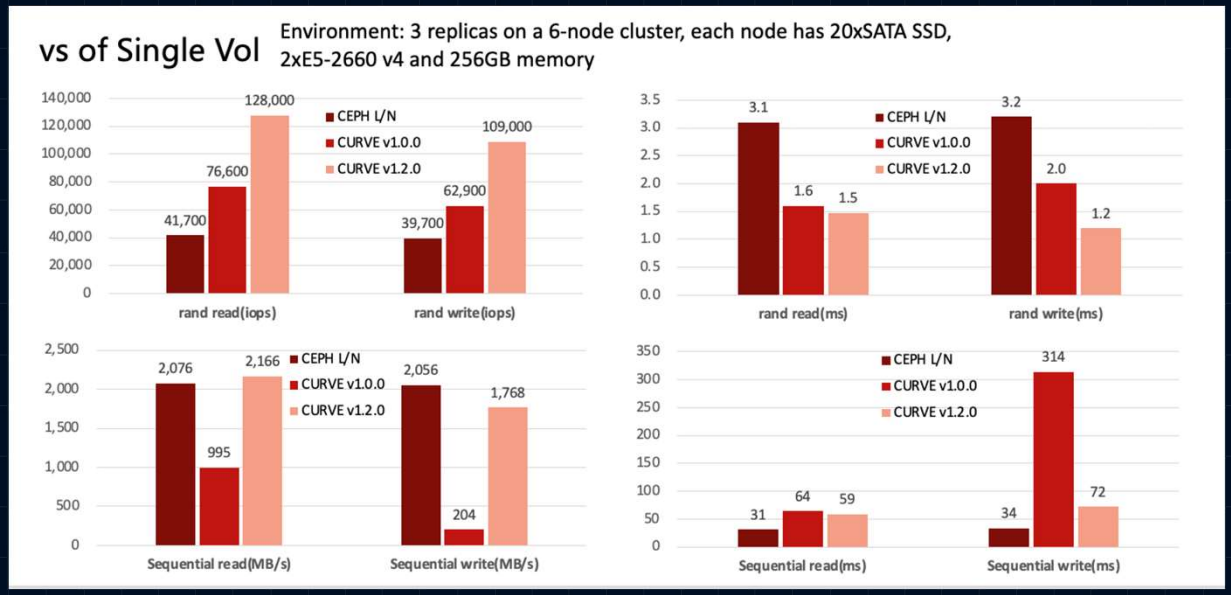
数据一致性协议	CURVE(RAFT)	BLUESTORE
	●	●
写成功确认	多数盘写成功	所有盘写成功
数据读取	copyset组中的Leader	PG中的节点
慢盘/盘故障的影响	不会打断IO	不时出现IO延迟
是否网络延迟敏感	Y	N
改进措施	对于写请求的RAFT并发	

CurveBS设计与CEPH对比

元数据管理	CURVE	BLUESTORE
元数据	预先创建的Chunk File池	RocksDB
元数据开销	几乎没有Ext4的元素开销	增加了读写放大
性能	高	需要针对RocksDB优化
改善措施	降低非覆盖写的写日志操作	
copyset映射/chunk映射管理	开销低	CRUSH开销极低

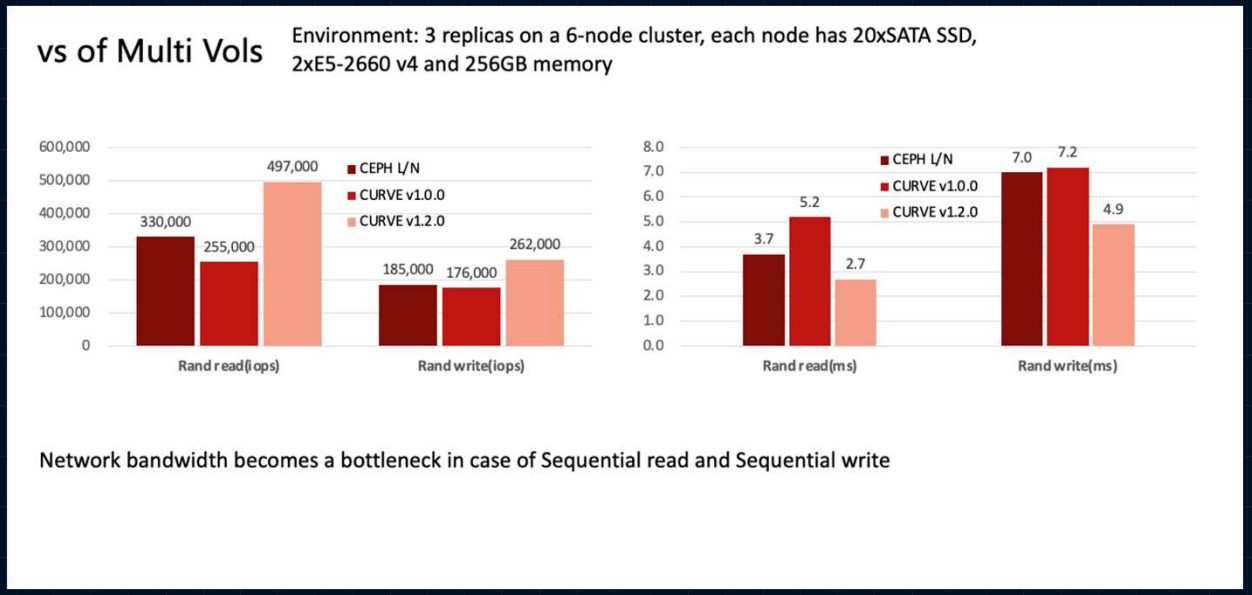
CurveBS与CEPH性能对比

- 单卷测试对比



CurveBS与CEPH性能对比

- 多卷测试对比



CurveFS设计目标

- 高性能设计
- 提供兼容的POSIX文件接口
- 底层支持云端块存储、对象存储、块存储
- 支持数据生命周期管理
- 支持云原生的文件系统

Roadmap

- 支持CurveBS和对象存储作为CurveFS的底层
- 兼容POSIX文件接口基于FUSE
- CurveFS的Cache模块
- CurveFS的云原生支持
- RAFT的优化
  Multi RAFT的并发写
  降低对于非覆盖写的写放大
- 支持数据多层管理

Thanks_

CURVE