Curve

新一代云原生分布式存储—Curve 上

D I G I T A L S A I L

李小翠

网易数帆存储团队

Curve

DIGITAL SAIL

目录

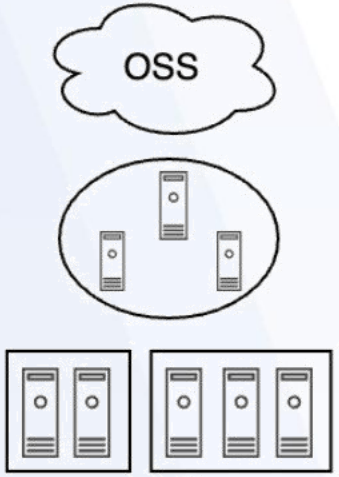
存储的发展

互联网时代，数据大爆炸

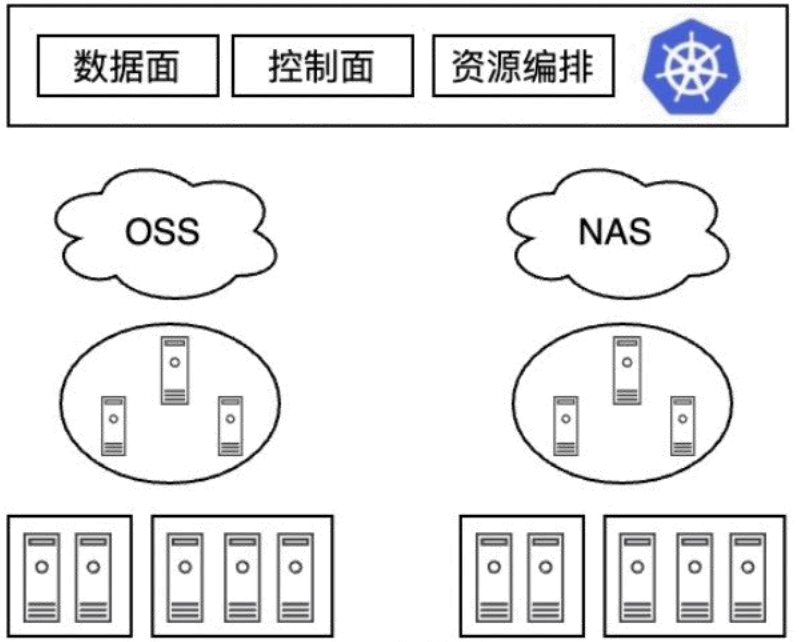
单机存储

集中存储

分布式存储



云存储



云原生存储

小型主机

容量有限

大型主机

成本高

单点问题

扩容困难

各存储设备通过网络互联

大规模

弹性扩容

底层构建在分布式存储之上

云的概念

成本：共用基础设施

弹性：随意扩缩容

速度：更快的构建发布业务

底层构建在分布式存储之上

云原生的概念：

易用性：跨平台，超融合，弹性

分布式存储的分类

Curve

按照各种应用场景所需的存储接口分类

对象存储

接口为简单的 Get、PUT、DEL 和其他扩展

文件存储

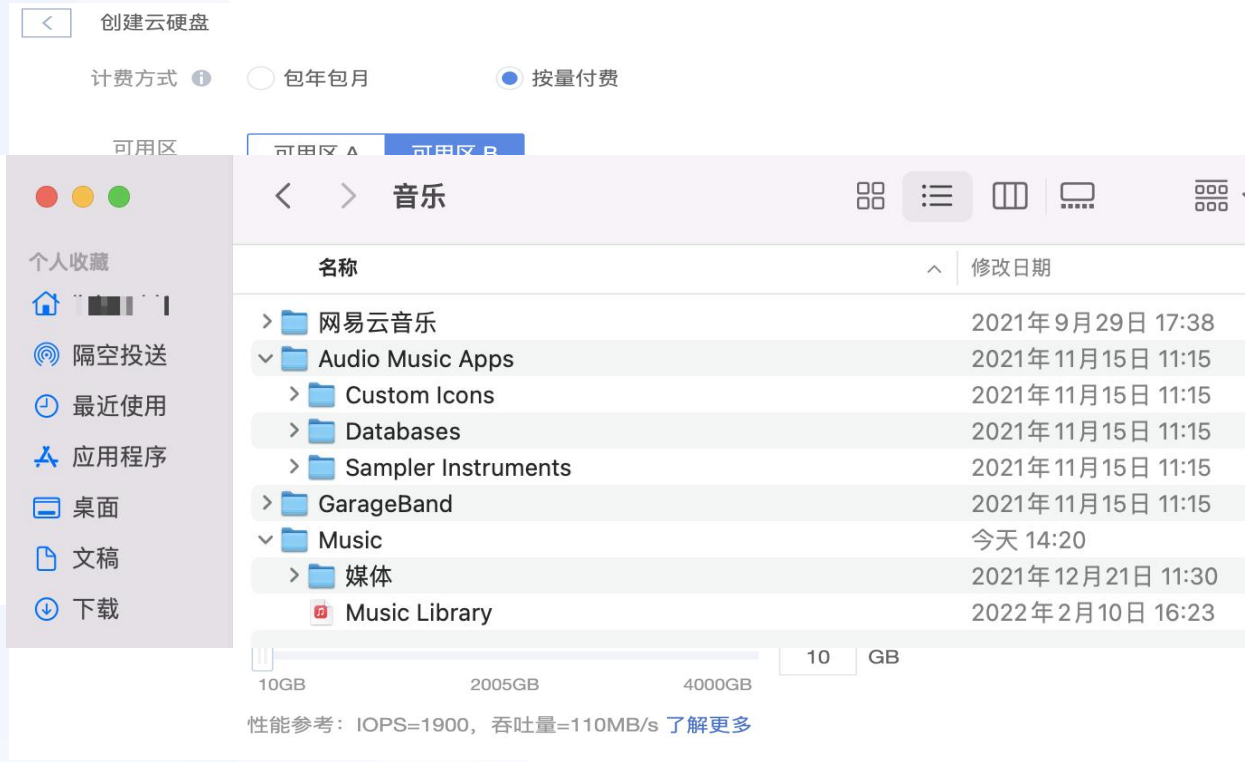
通常意义是支持 POSIX 接口

传统意义的文件系统： Ext4

块存储

对指定地址空间进行随机读写

传统意义的块存储：磁盘



分布式存储的要素

Curve

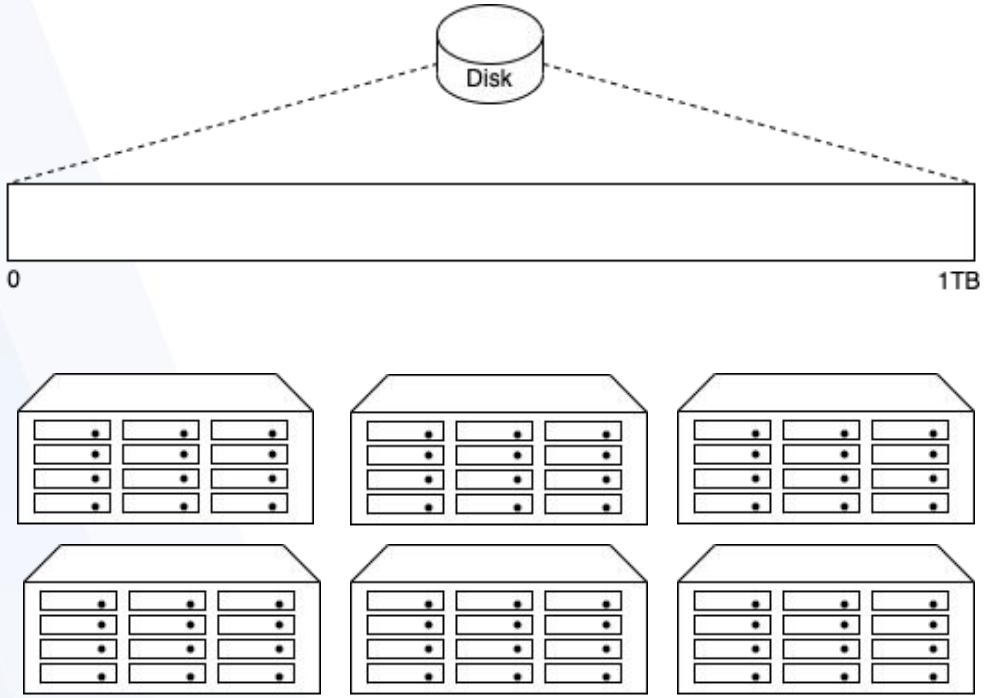
如何构建分布式文件系统？ 以分布式块存储为例。

要什么

- 提供大容量的块设备
- 可以在指定地址空间内随机读写 write(offset, len)
- 服务质量要求：数据不能丢、服务随时可用、弹性扩缩容

有什么

- 成百上千台存储节点
- 磁盘故障、机器故障、网络故障概率性发生



分布式存储系统需要满足接口需求，并且有持续监控、错误检测、容错与自动恢复的能力

以达到高可靠、高可用、高可扩

分布式存储的要素

Curve

要素拆解

数据分布 —— 无中心节点/中心节点 **均衡**

地址空间的每段数据会分布在不同机器的磁盘上，如何找到这些数据？

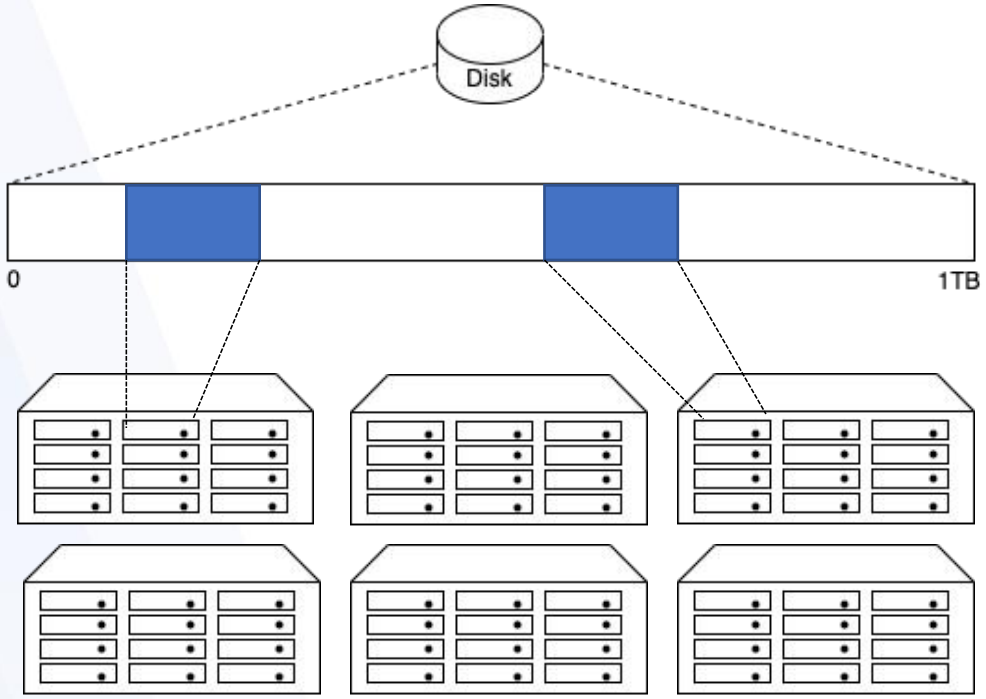
可靠性 & 可用性 —— 多副本/EC **服务不可用时间**

数据一致性 —— 一致性协议

如何保证数据不丢？如何保证各种硬件故障的时候读写都正常？

可扩展性 —— 和数据分布的方式相关

所用容量都用完后，可以新增机器扩展容量



分布式存储的要素 — 数据分布

Curve

无中心节点：哈希算法

- 映射信息无需记录，直接通过计算获得
- 伪随机算法在服务器数量特别大的时候接近均衡
- 节点故障（DiskNums）变更会涉及其他数据的迁移

INPUT (Offset, Len)	HASH	HASH mod 72 (DiskNums)
(0, 4MB)	1633428562	58
(4MB, 8MB)	7594634739	3
(8MB, 16MB)	3421657995	51

有中心节点：持久化对应关系

- 需要将数据分布（元数据）持久化
- 中心节点感知集群的信息，进行资源实时调度
- 节点故障不会涉及其他的数据迁移

KEY (Offset, Len)	VALUE (DiskID)
(0, 4MB)	70
(4MB, 8MB)	60
(8MB, 16MB)	50

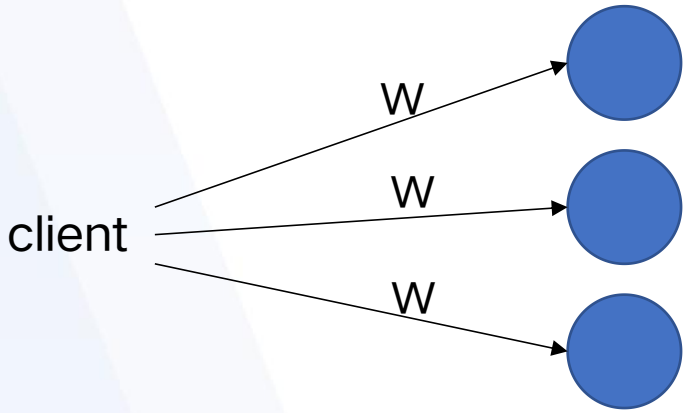
分布式存储的要素 — 一致性协议

Curve

多副本： 写三次？ 一致性协议　一致性：WARO（Write-all-read-one）、Quorum

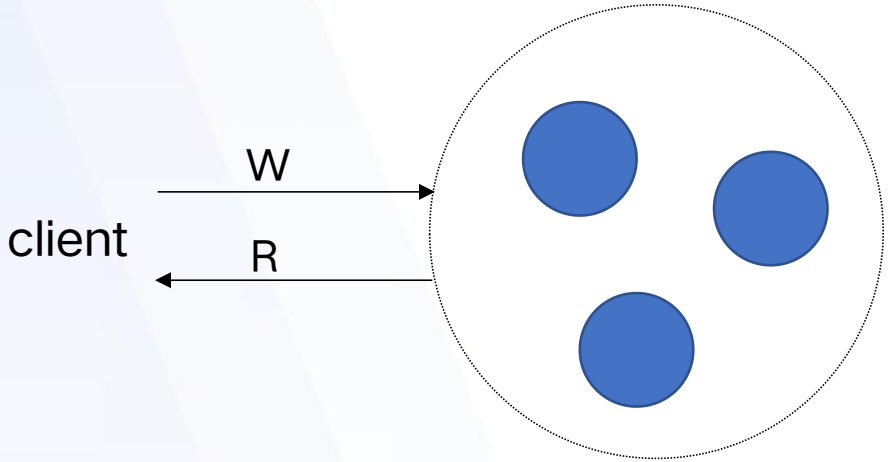
WARO

- 所有副本写成功
- 读可用性高：可以读任一副本
- 写可用性较低，任一副本异常写失败



Quorum

- 大多数副本写成功
- 读写服务可用性做一个折中
- 写性能提升，速度取决于写的较快的大多数



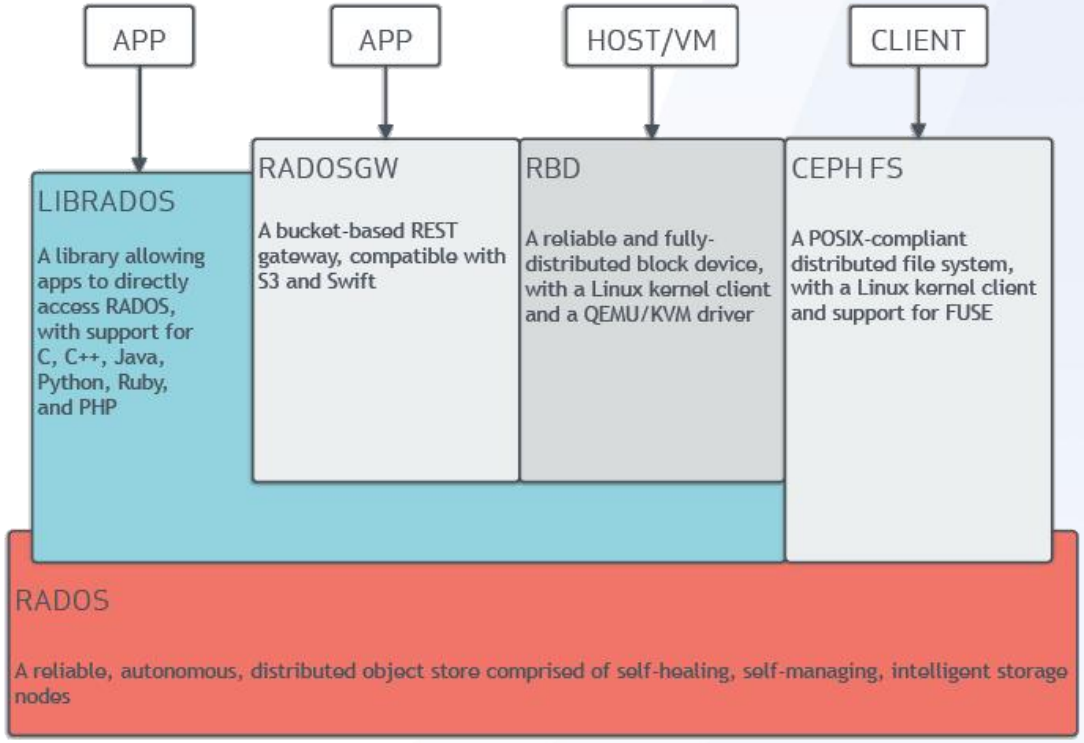
Curve

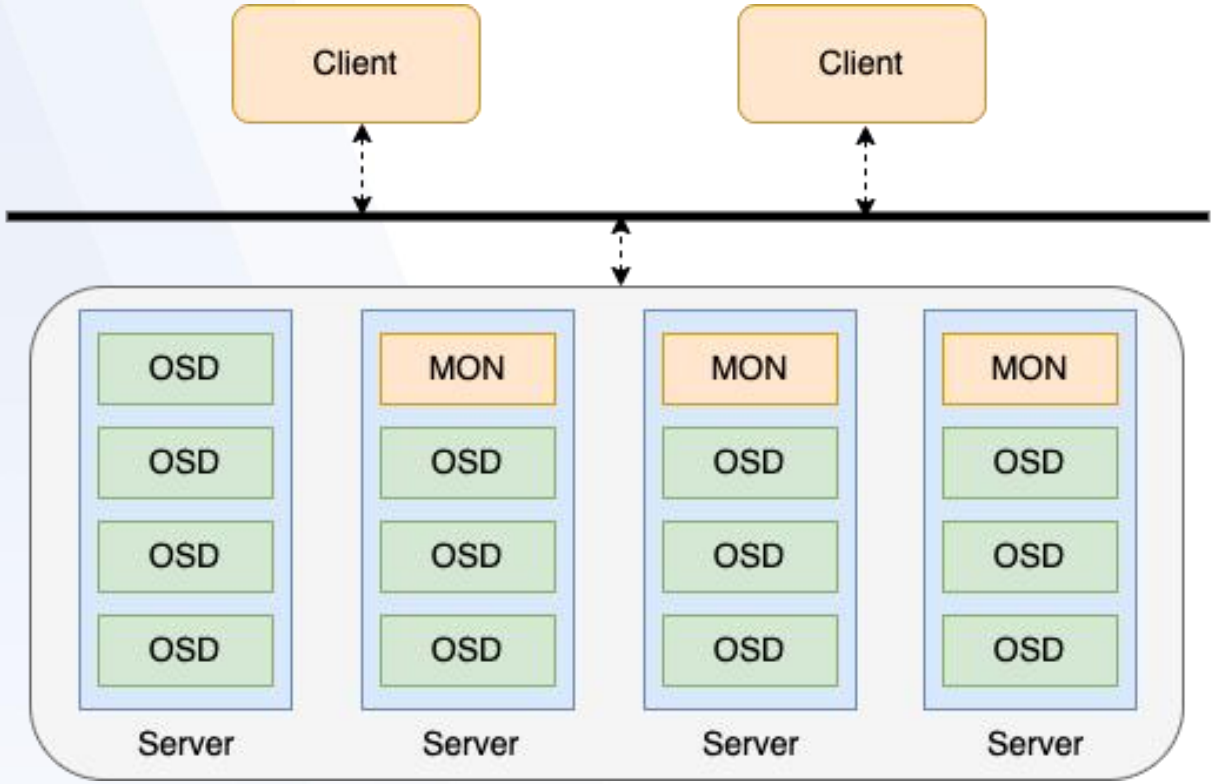
DIGITAL SAIL

目录

架构简介 — 总体架构

开源分布式存储界的扛把子　支持块存储、文件存储、对象存储





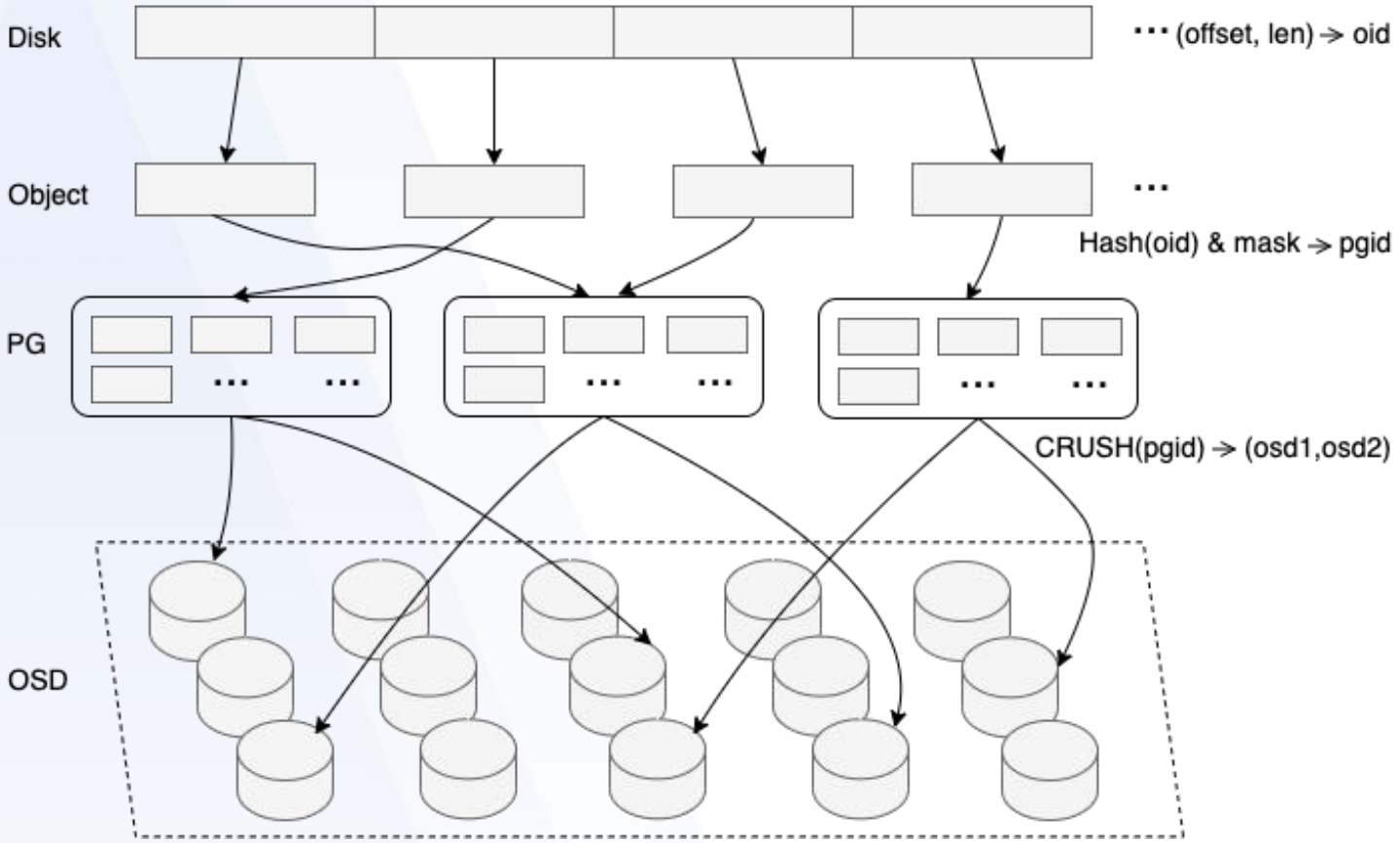
架构简介 — 概念介绍

Curve

object：存储单元

PG：Placement Groups
归置组
归置组中的成员为副本

OSD：Object Storage Device,
管理一个磁盘的进程



架构简介 — 数据放置

Curve

使用多级哈希的方式

根据 offset, len, name.. 生成ObjectID

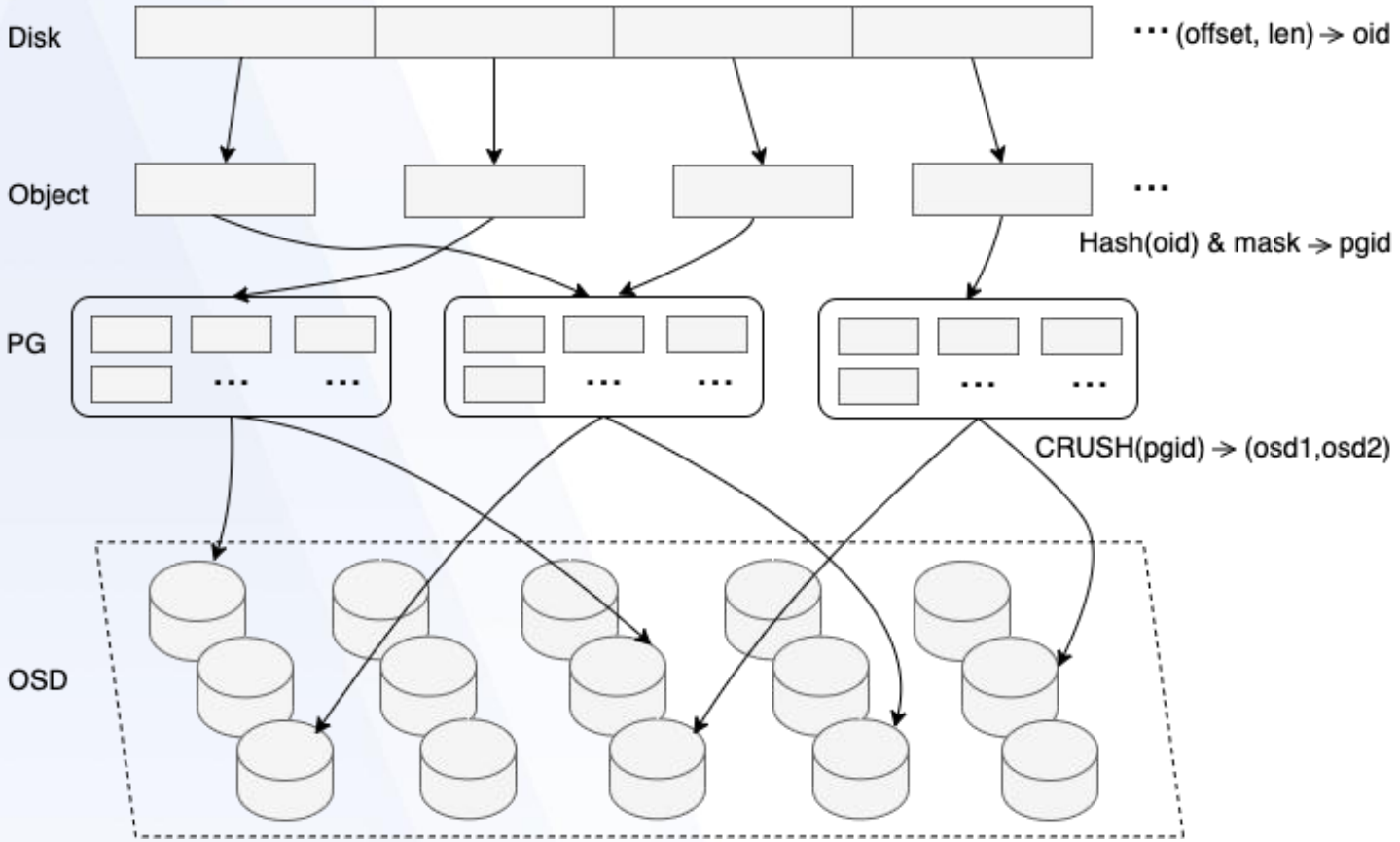
rbd\udata.6855c174a277a30.000000000005c2

对ObjectID进行哈希并取模（复制组数量）得到pgid

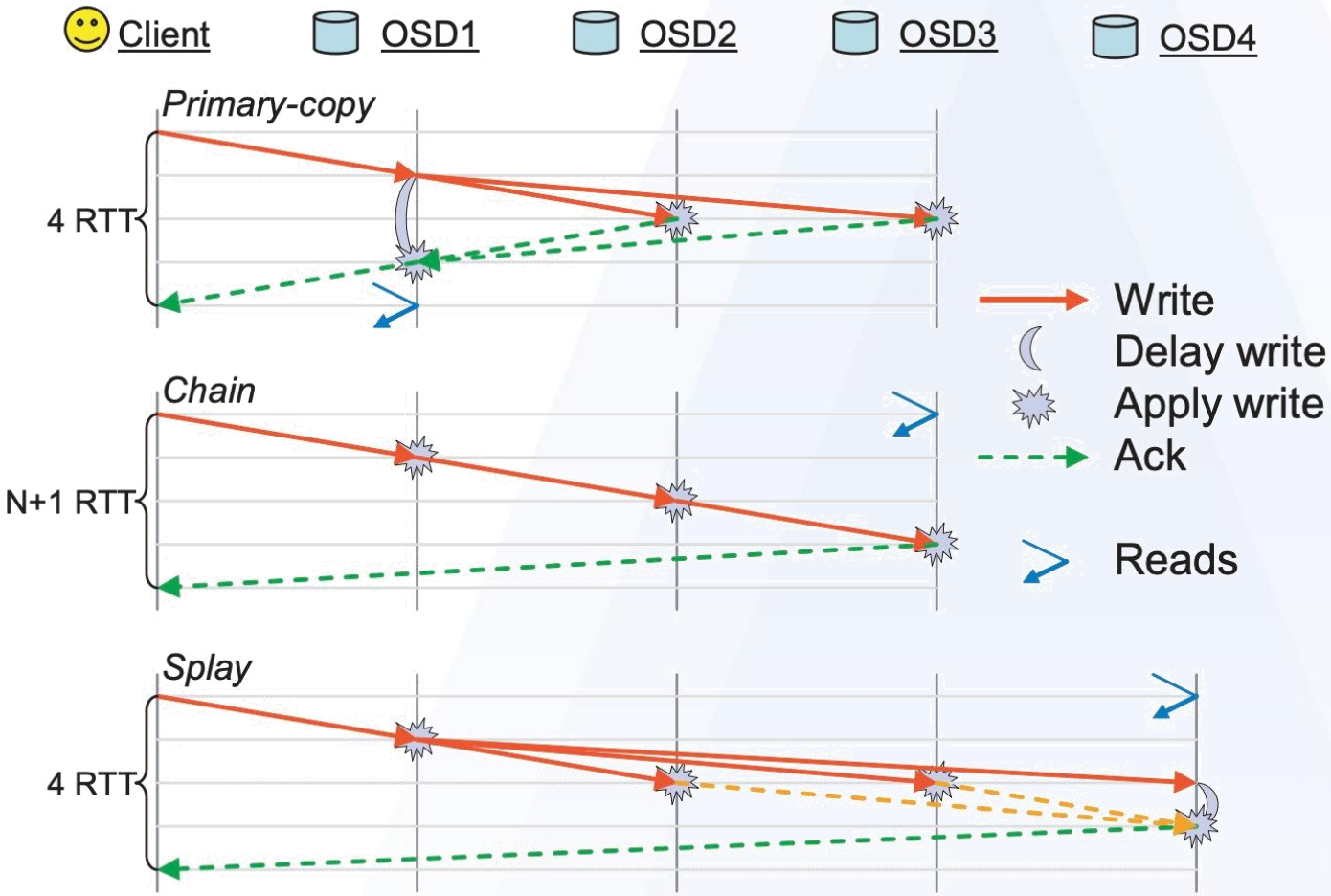
head_D35c9011

使用CRUSH算法根据pgid获得指定的副本个数的id

osd.1， osd.2， osd.3



架构简介 — 多副本一致性协议



使用WARO一致性协议

- 所有副本写完成返回客户端
- 延迟取决于所有副本中最慢的那一个

复制策略

- 主动拷贝、链式复制、splay复制

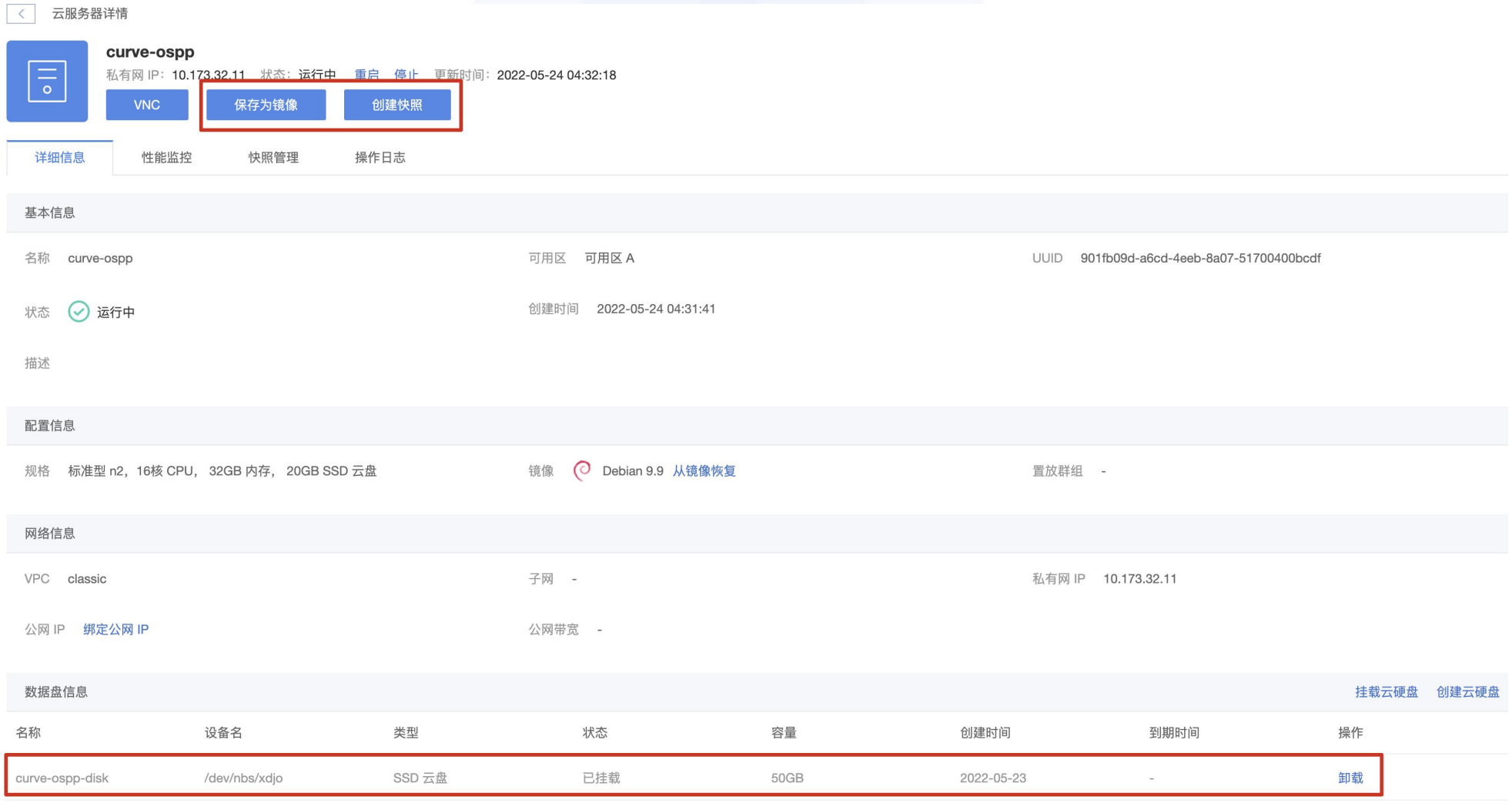
异常处理

- PG有23种状态：Peering，Degraded等
- 强一致性协议对异常的容忍较差

块存储场景

Curve

为云主机提供云盘，云盘提供随机读写、快照（数据备份，灾备使用）、镜像（模板，自定义）功能。

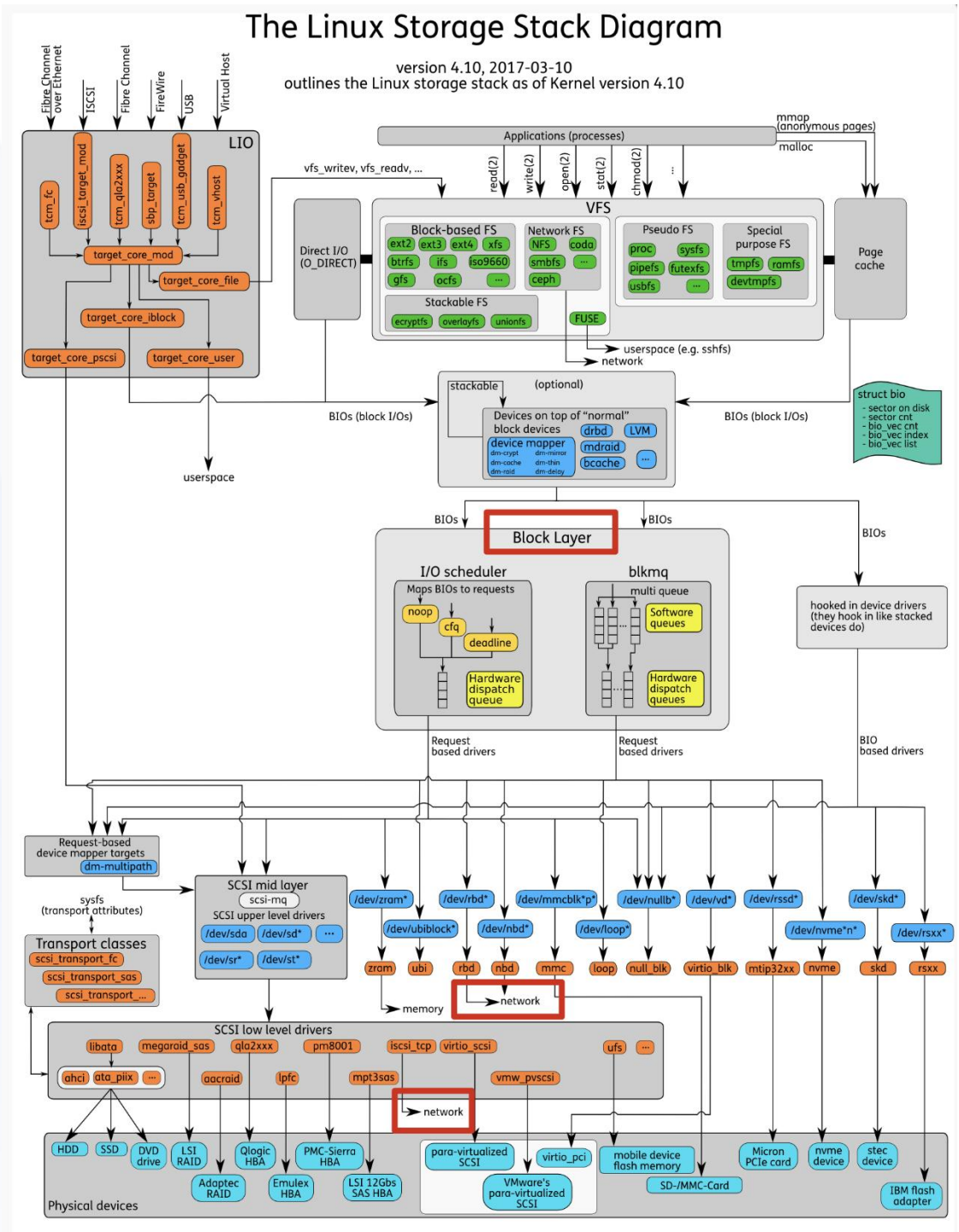


块存储场景

为物理机提供块设备

Linux IO栈

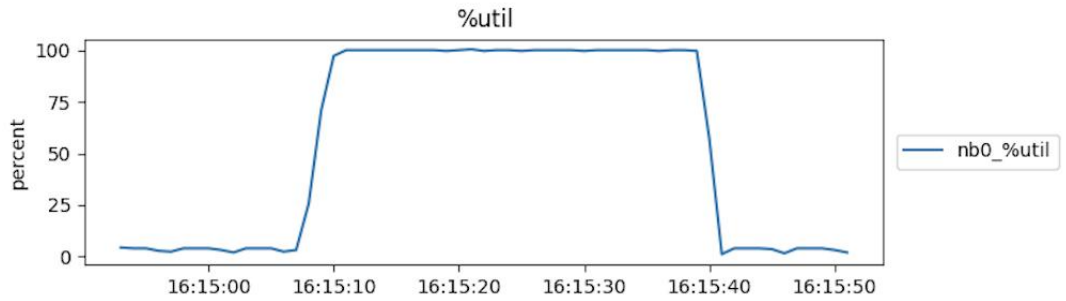
应用程序 -> 文件系统 -> 块设备层 -> **不同协议/驱动**

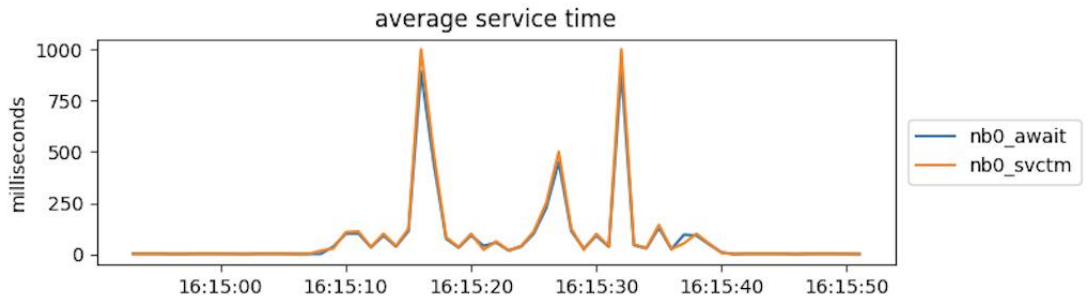


使用中的问题

Curve

- **io抖动（一致性协议）：** 异常场景（比如阵列卡一致性巡检，坏盘，**慢盘**，网络异常），服务升级





- **性能差（一致性协议）：** 在通用硬件下，无法支撑数据库、kafka等中间件对存储性能和稳定性要求
- **容量不均衡（数据放置）：** 集群各节点容量不均衡需要人为干预
- 上述问题和架构涉及、核心功能的选型有关，在已有开源版本上改进代价很大

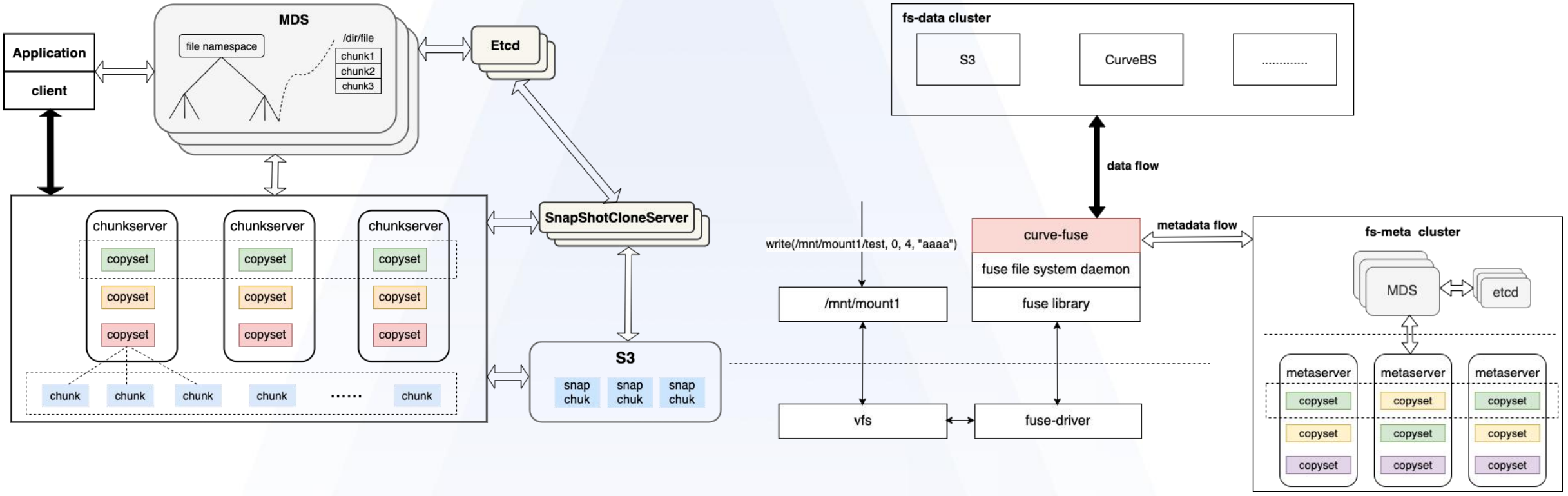
Curve

DIGITAL SAIL

目录

架构简介 — 总体架构

支持块存储、文件存储（多种存储后端）



架构简介 — 概念介绍

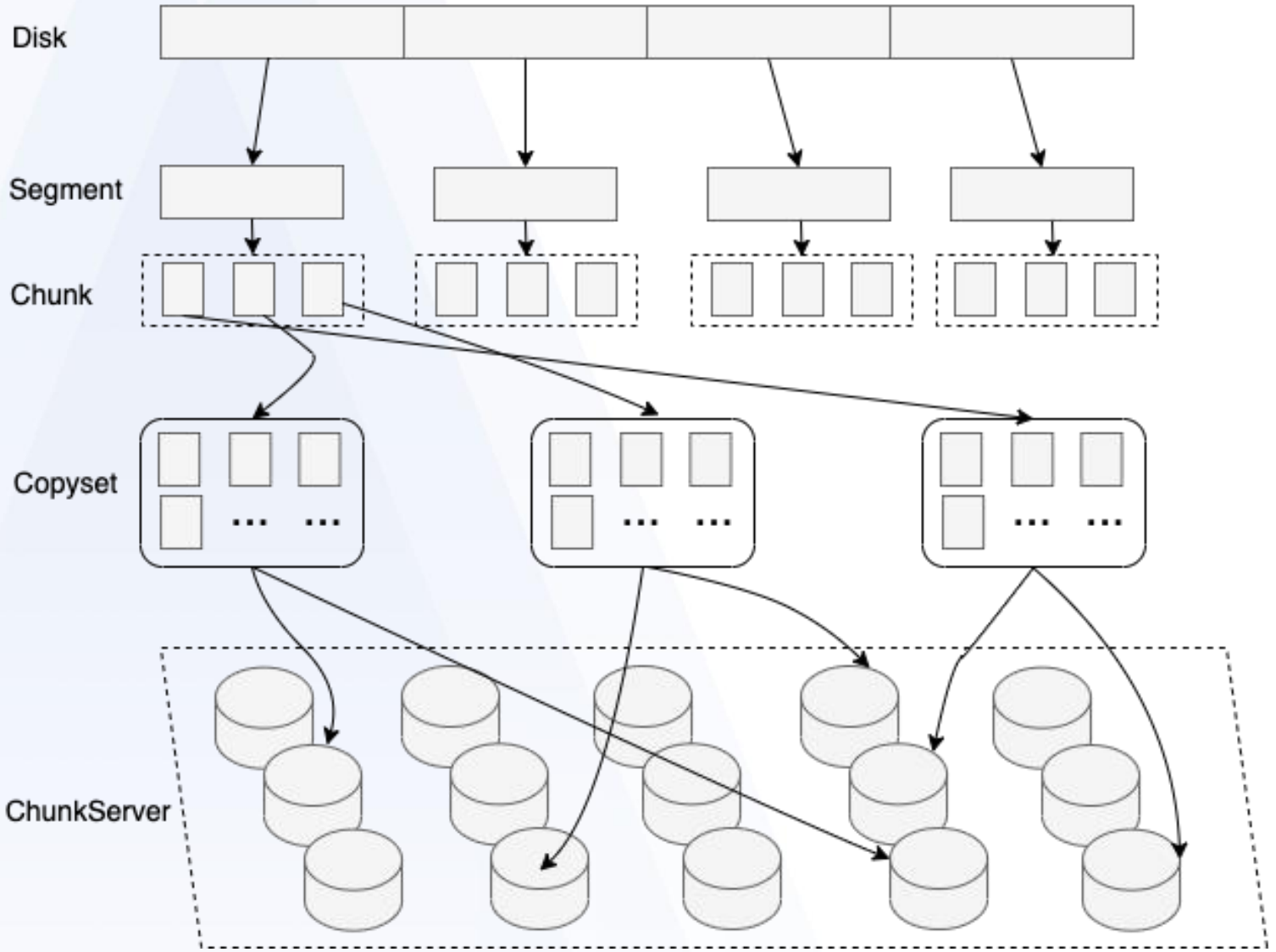
Curve

Segment: 空间分配的基本单元

Chunk: 数据分片

Copyset: 复制组

ChunkServer: 管理一个磁盘进程



架构简介 — 数据放置

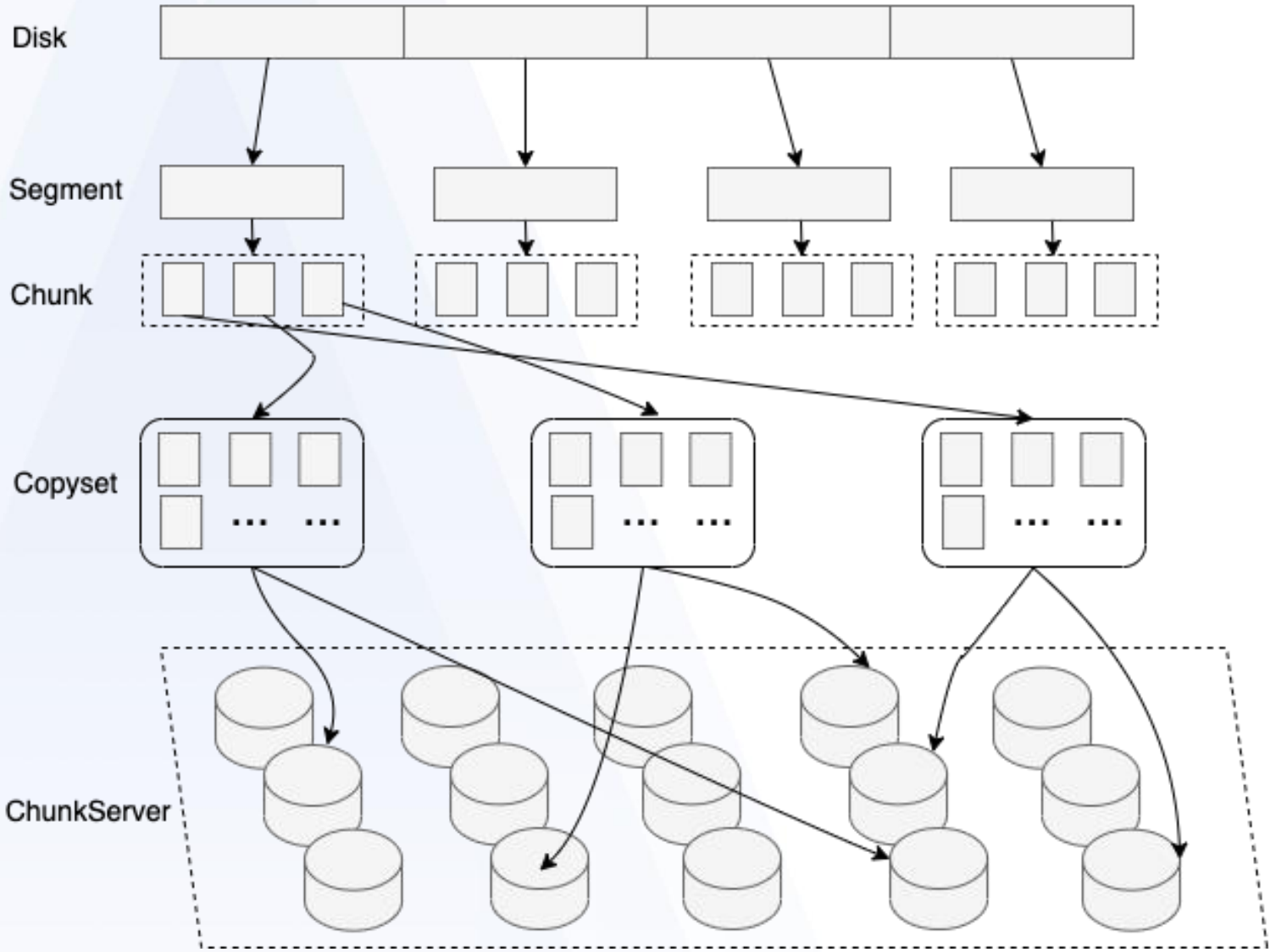
Curve

Copyset的放置

- 由中心节点MDS以Scatter-width均衡为目标进行创建

Chunk的分配

- 由中心节点MDS在各Copyset中根据权重进行选择

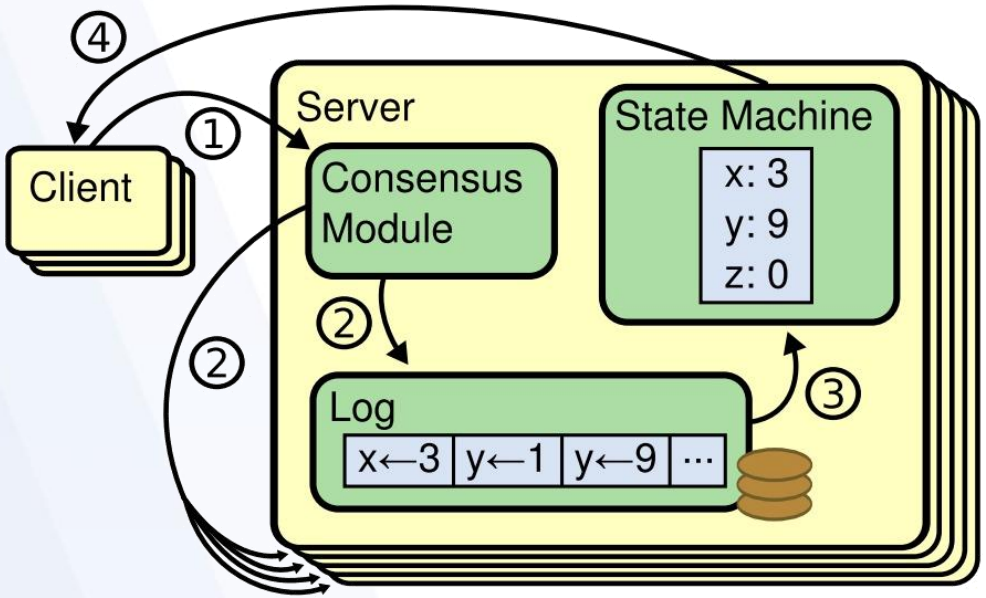


架构简介 — 一致性协议

Curve

Quorum一致性协议

- 大多数副本写完成返回客户端
- 延迟取决于所有副本中最快的大多数

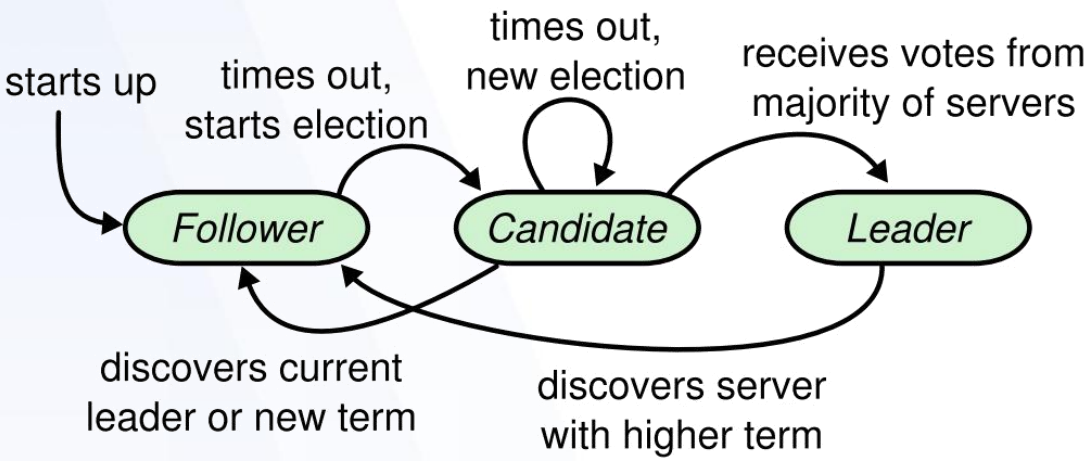


复制策略

- 主动拷贝，由 leader 向 follower 并发拷贝

异常处理

- 自动leader选举
- 2N+1 副本数可以容忍 N 副本数异常



主要亮点

Curve

基于在架构上的选择和优秀的工程实践，Curve 在性能、运维、稳定性、工程实践质量上都优于Ceph

高性能

易运维

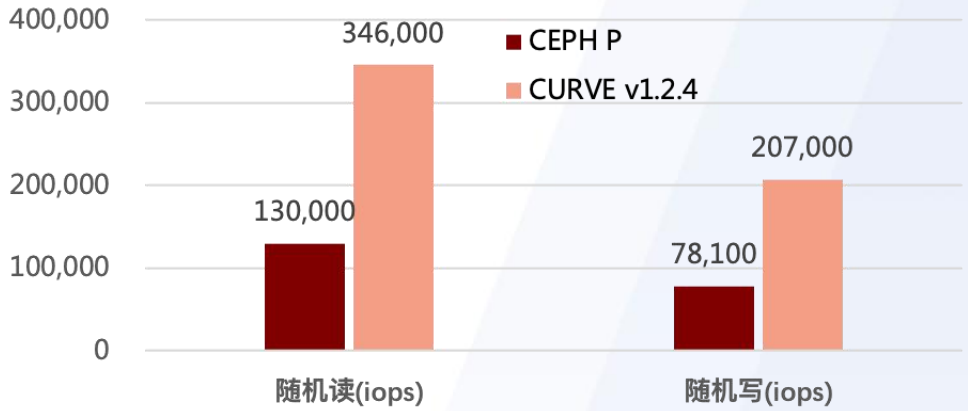
更稳定

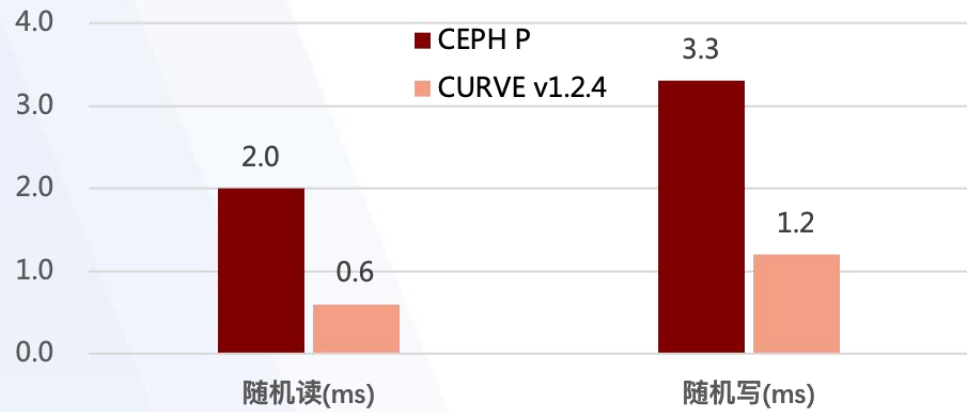
高质量

主要亮点 — 高性能

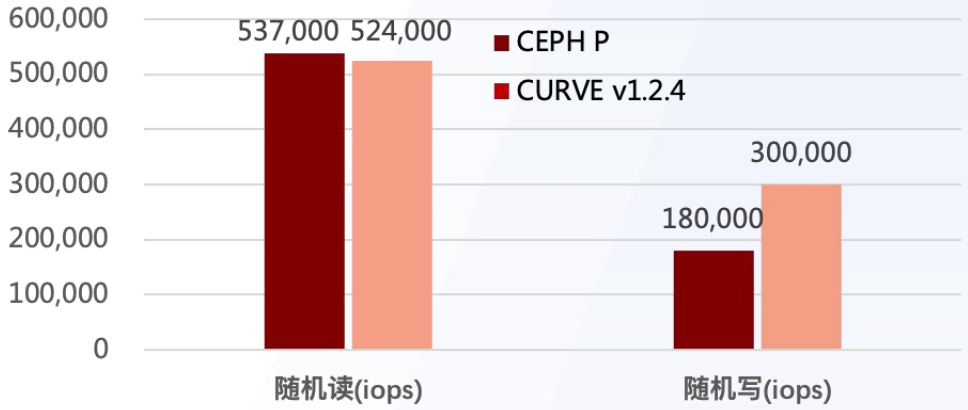
高性能　　NVME　块存储场景，Curve随机读写性能远优于Ceph

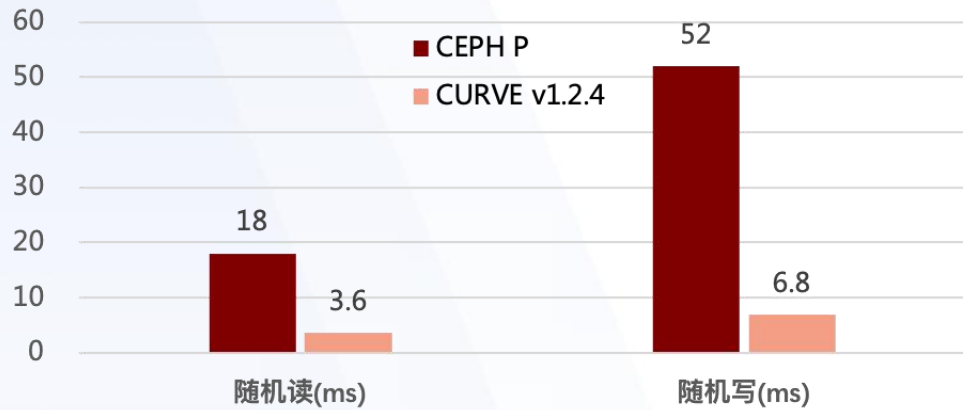
单卷





多卷





测试环境：3台服务器*8块NVME， Intel(R) Xeon(R) Gold 5318Y CPU @ 2.10GHz ，256G，3副本，使用自带fio

主要亮点 — 易运维

Curve

运维场景	Curve	Ceph
加盘	对IO无影响	秒级io影响
服务端升级	对IO无影响	重启管控面IO无影响，重启osd io秒级影响
客户端升级	热升级，秒级抖动	不支持热升级，需要业务停服
集群监控	丰富的metric	metric类型较少

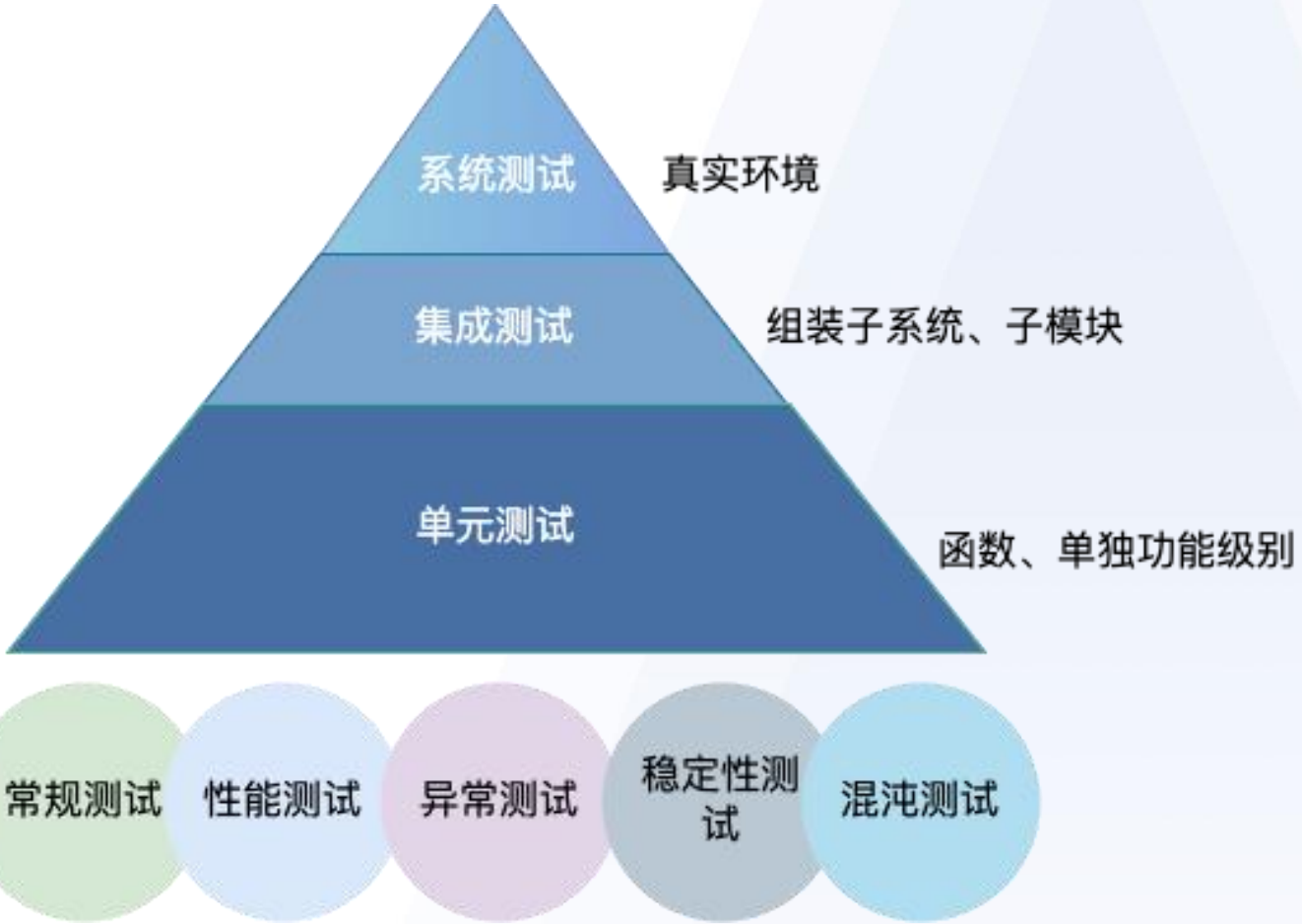
主要亮点 — 更稳定

Curve

异常场景	Curve	Ceph
坏盘	基本无抖动	无明显抖动
慢盘	io持续抖动，但util未100%	io持续抖动，util持续100%
网络丢包	随着loss增大，还有部分io	随着loss增大，无法进行io
机器宕机	io略微波动	io卡住10s以上
机器卡住	io抖动4s	不可恢复

主要亮点 — 高质量

良好的模块化和抽象设计；完善的测试体系



单元测试覆盖率	lines	functions	link
Curve	85.4%	89%	curve
Ceph	37.1%	43.3%	ceph

应用情况

Curve

Curve 在网易集团内有大规模的生产应用

为核心业务提供稳定的存储服务，单集群存数万个卷，储容量PB级别

✓ 网易集团内部业务：

- **网易严选，网易云音乐**
  网易有道，网易游戏
  网易Lofter，云信

在集团外有联合开发用户和测试用户

✓ 网易外部用户：

- **超聚变，创云融达信息技术**
- **扬州万方电子技术，思谋科技**

网易严选　有道 youdao　网易游戏 游戏热爱者　FUSION　创云融达 ORIGINALITY CLOUD

LOFTER　网易云信　网易云音乐　万方电子 WANFANG ELECTRONIC　SmartMore 思谋

Curve

DIGITAL SAIL

目录

OpenCurve_bot

扫一扫上面的二维码图案，加我为朋友

FAQ