Curve

Curve云原生分布式存储—高性能、更稳定、易运维

D I G I T A L S A I L

吴汉卿

网易数帆存储团队

大纲

Curve

- 项目介绍
- 项目应用场景
- 主要亮点
- 技术架构先进性
- 应用情况

项目介绍

Curve

Curve 是一款**高性能、易运维、云原生**的开源分布式存储系统。

应用于以下场景:

- 为 OpenStack 云主机、物理服务器等提供**高性能块存储**
- 为 Kubernetes 提供**持久化存储卷**
- 赋能主流云原生数据库，完美支持**存算分离**
- 为公有云、混合云的大数据/AI训练场景提供高性价比**共享文件存储**

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 易景捷

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。

CNCF沙箱项目：意味着**全球顶级开源基金会**对网易存储技术演进的认可。

生产用户

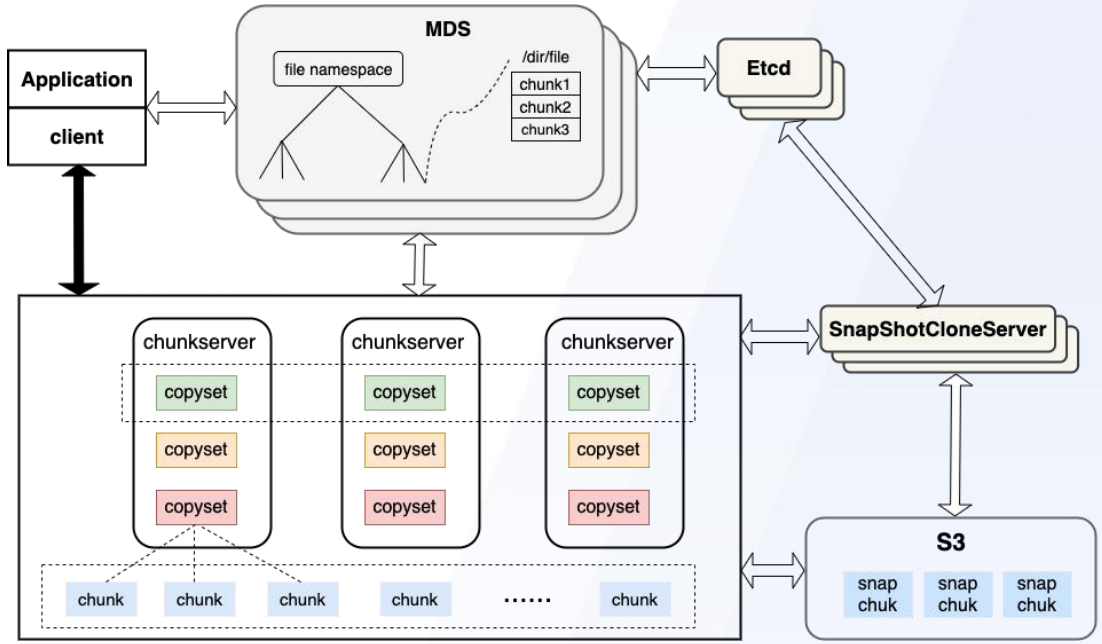
网易严选 | HUAWEI | 品高软件 BINGO www.bingosoft.net | 网易云音乐 | 有道 youdao

LOFTER | 万方电子 WANFANG ELECTRONIC | 网易游戏 | 创云融达 ORIGINALITY CLOUD | 网易云信

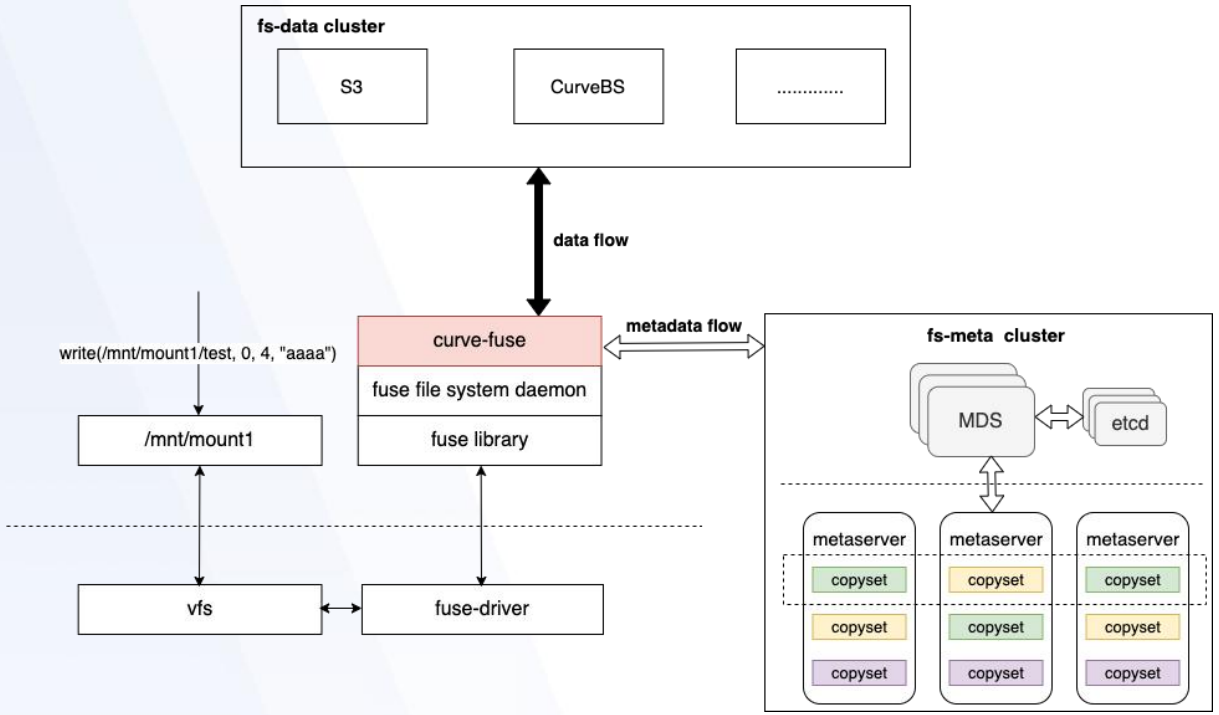
项目介绍

Curve

块存储支持快照、克隆和恢复，支持QEMU、NBD、CSI



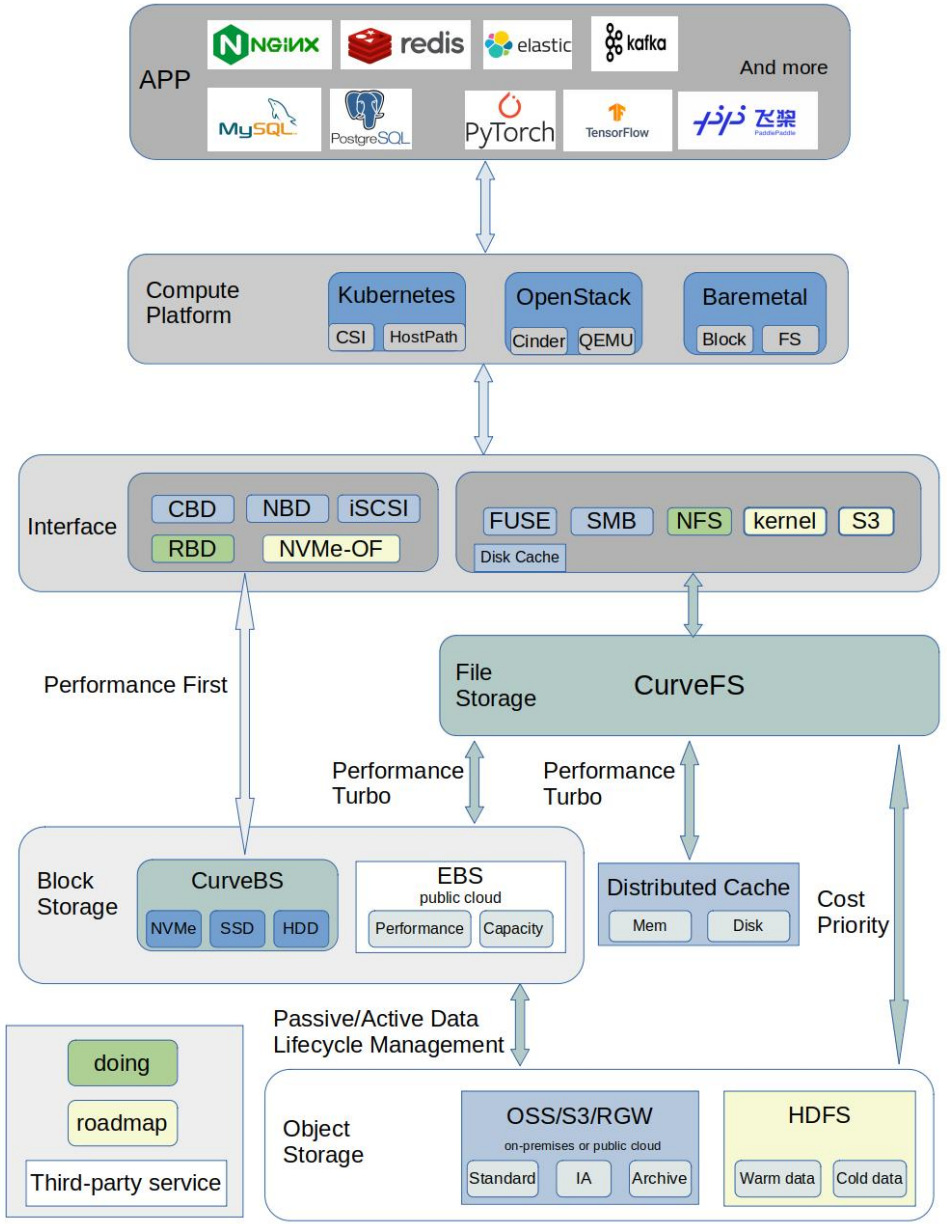
文件存储基于 Fuse 支持 Posix 文件系统接口，支持 CSI



应用场景

Curve 目前成熟的应用场景：

- 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



主要亮点

Curve

高性能

易运维

更稳定

高质量

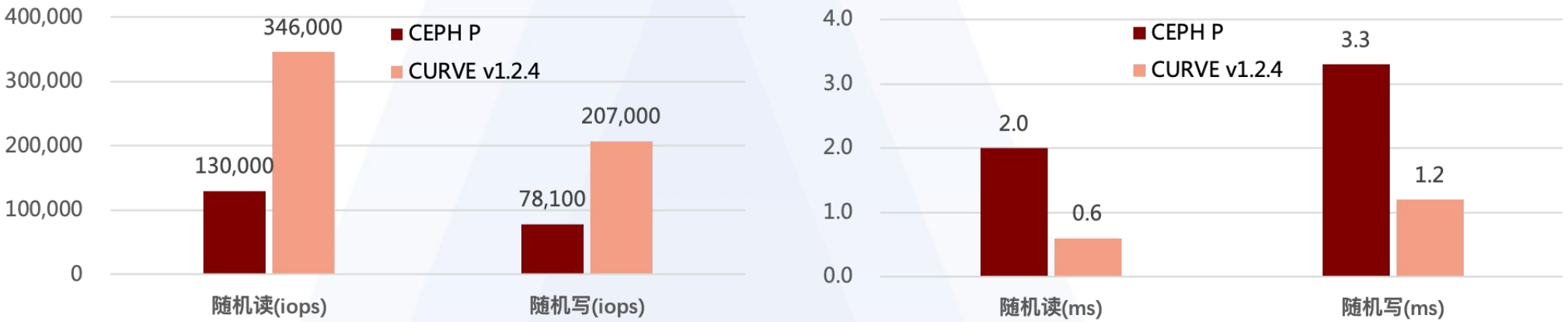
高可扩

高可用

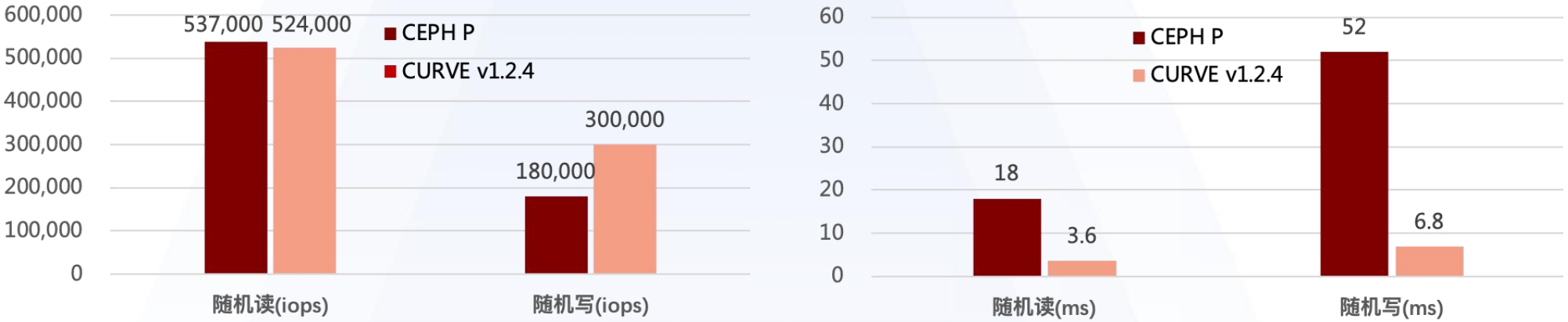
主要亮点

Curve

高性能　小IO随机读写



单卷 - 4k随机读写性能对比



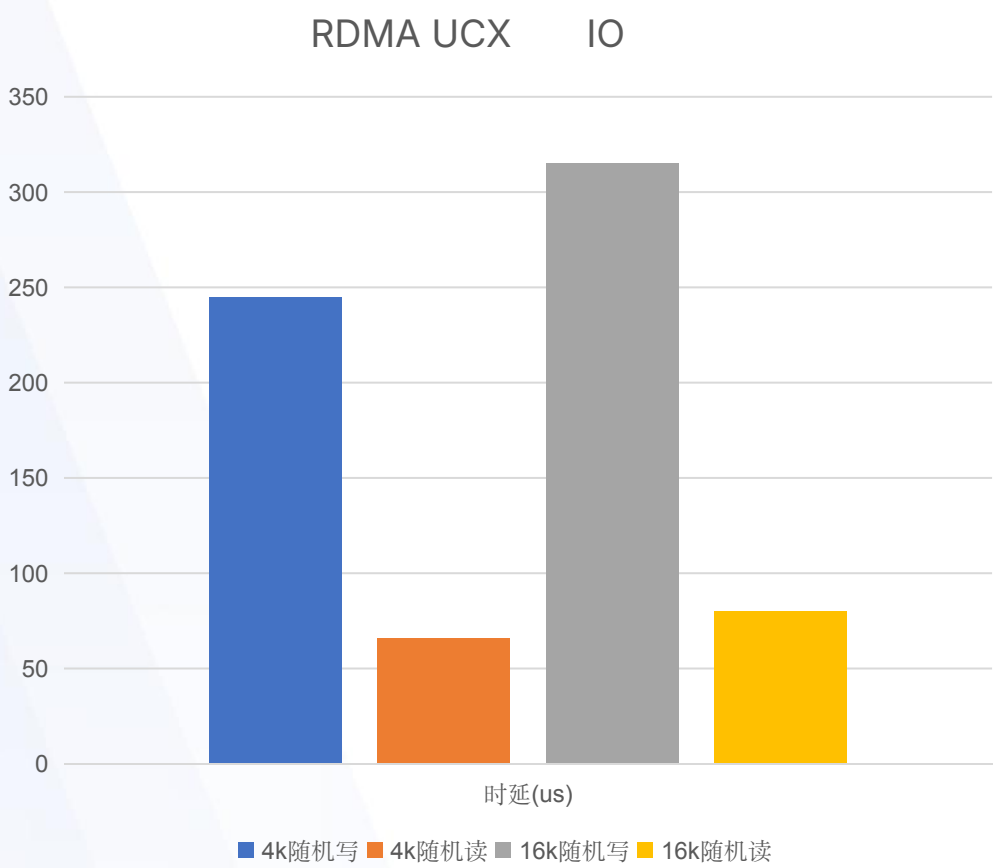
多卷 - 4k随机读写性能对比

主要亮点

Curve

高性能　云原生数据库场景下最新进展

- **SPDK&RDMA：**适配完成，内部业务上线
- **PolarFS性能优化**
  - ✔ 使用无锁工作队列，降低shared memory轮询线程，使用unix socket作任务通知，避免CPU 100%空转
  - ✔ 优化pfs_lseek锁，对多线程更加友好
  - ✔ 一写多读保护（IO fence）支持
  - ✔ 支持大于4M的读写
  - ✔ 减少pfs journal的补偿读



主要亮点

Curve

高性能　后续优化方向

- 混合存储引擎（目前完成方案选型测试，在业务验证阶段）
- 更高性能硬件适配
- 专用硬件offload网络和存储IO
- 高性能、低开销的存储引擎
- 其他

主要亮点

Curve

易运维 CurveAdm部署运维工具

- 适配常见操作系统发行版
- 封装简化高级运维操作
- 高效的多集群管理能力
- 集成周边组件部署能力（PolarFS、tgt等）
- 支持一键安装、一键升级、一键收集问题日志等

```
→  curveadm git:(master) ✗ curveadm -h
Usage:  curveadm COMMAND [OPTIONS]

Deploy and manage CurveBS/CurveFS cluster

Management Commands:
  client      Manage client
  cluster     Manage clusters
  config      Manage cluster topology
  hosts       Manage hosts
  pfs         Manage pfs
  playground  Manage playground
  target      Manage SCSI target of CurveBS

Commands:
  audit       Show audit log of operation
  clean       Clean service's environment
  completion  Generate completion script
  deploy      Deploy cluster
  enter       Enter service container
  format      Format chunkfile pool
  map         Map a volume to nbd device
  migrate     Migrate services
  mount       Mount filesystem
  precheck    Precheck environment
  reload      Reload service
  restart     Restart service
  scale-out   Scale out cluster
  ssh         Connect remote host
  start       Start service
  status      Display service status
  stop        Stop service
  support     Get support from Curve team
  umount      Umount filesystem
  unmap       Unmap nbd device
  upgrade     Upgrade service

Options:
  -d, --debug     Print debug information
  -h, --help      Print usage
  -u, --upgrade   Upgrade curveadm itself to the latest version
  -v, --version   Print version information and quit

Examples:
  $ curveadm playground --kind curvebs  # Run a CurveBS playground quickly
  $ curveadm cluster add c1             # Add a cluster named 'c1'
  $ curveadm deploy                     # Deploy current cluster
  $ curveadm stop                       # Stop current cluster service
  $ curveadm clean                      # Clean current cluster
  $ curveadm enter 6ff561598c6f         # Enter specified service container
  $ curveadm -u                         # Upgrade curveadm itself to the latest version

Run 'curveadm COMMAND --help' for more information on a command.
```

主要亮点

Curve

易运维 块存储场景，Curve运维更友好

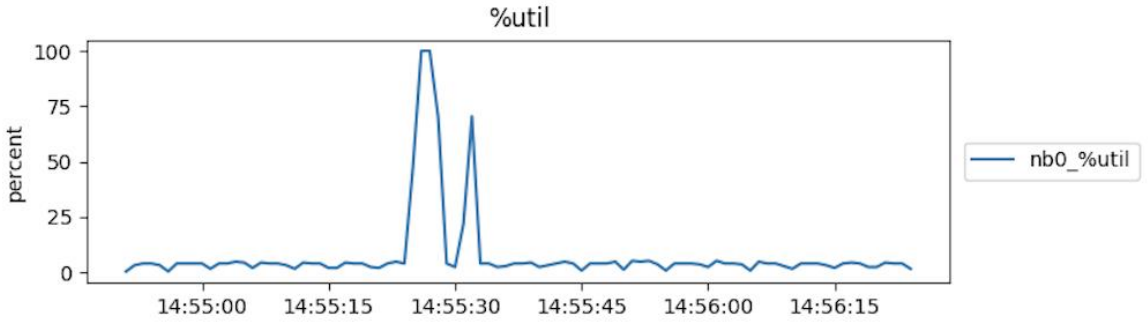
运维场景	Curve	Ceph
集群环境	支持x86、Arm	支持x86、Arm
集群部署	支持一键部署	支持一键部署
集群扩容	物理池方式扩容，卷可以跨物理池	物理池方式扩容，卷不可以跨物理池
加盘	对IO无影响	秒级IO影响
服务端升级	对IO无影响	重启管控面IO无影响，重启OSD IO秒级影响
客户端升级	热升级，秒级抖动	不支持热升级，需要业务停服
集群监控	丰富的metric	metric类型较少

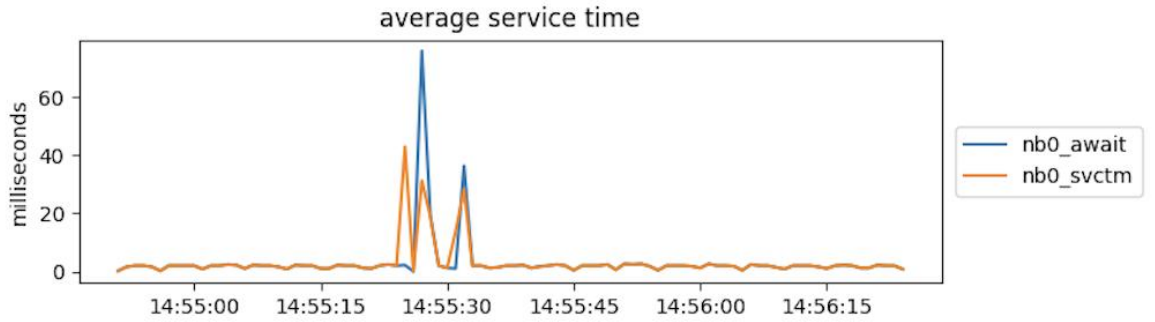
主要亮点

Curve

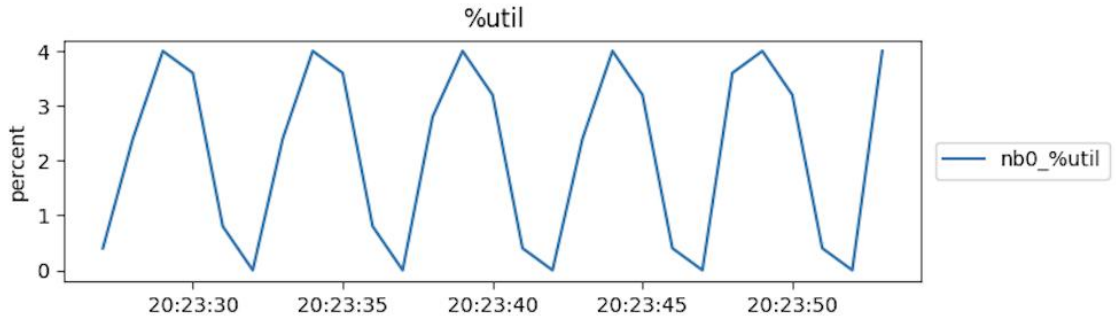
易运维　加盘

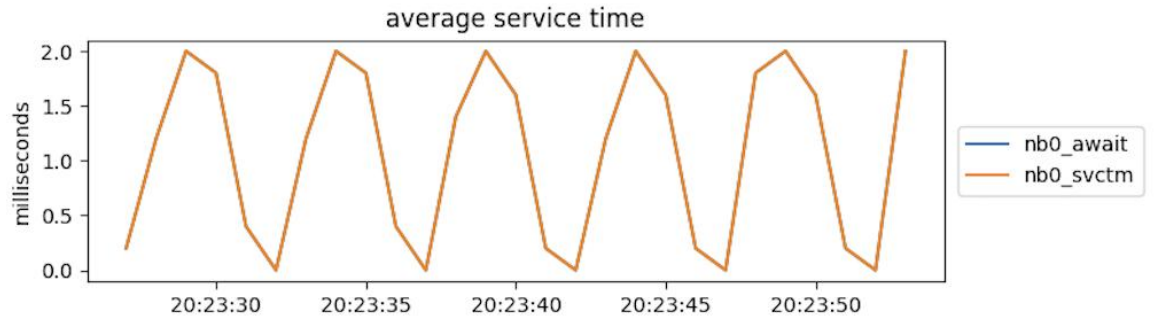
Ceph





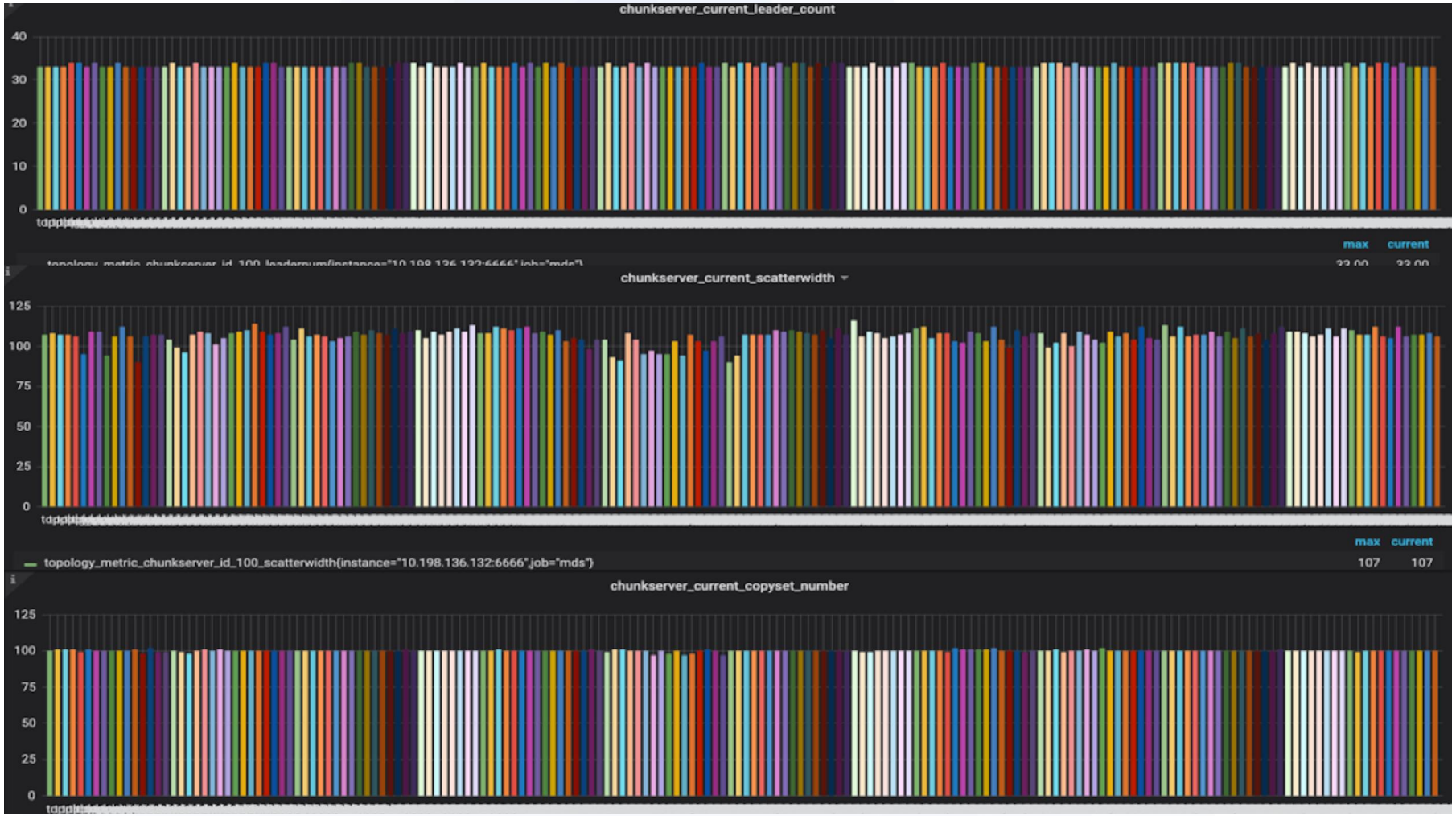
Curve





主要亮点

易运维　集群自动均衡



主要亮点

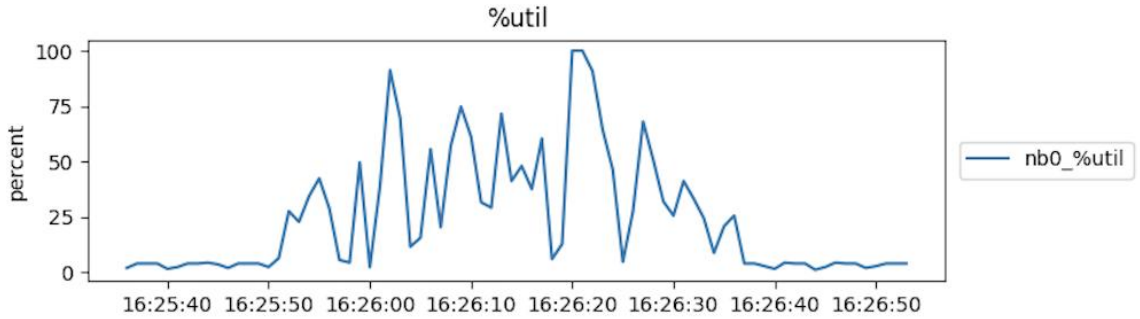
Curve

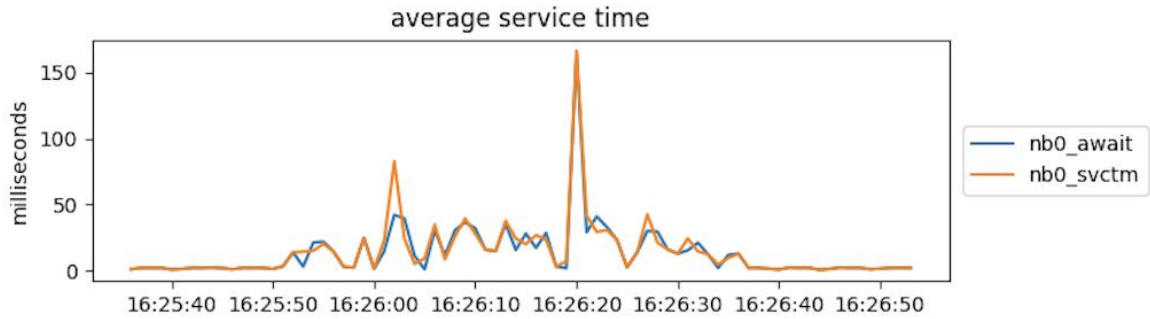
更稳定 块存储场景，常见异常Curve稳定性优于Ceph

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	随着丢包比例增大，还有部分IO	随着丢包比例增大，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

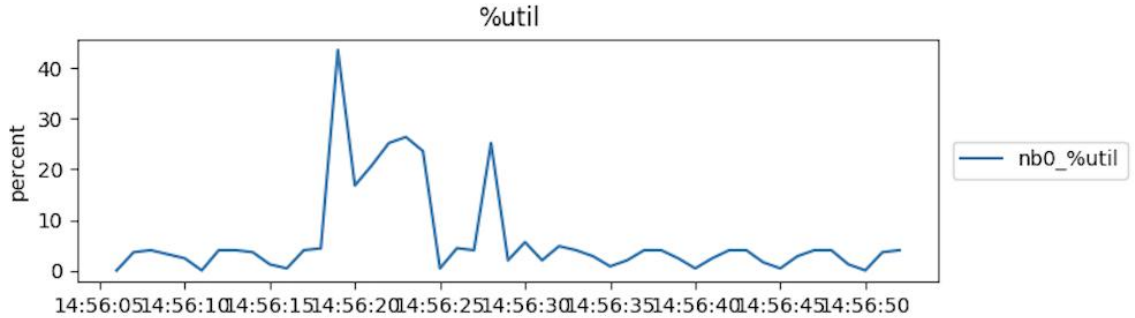
主要亮点

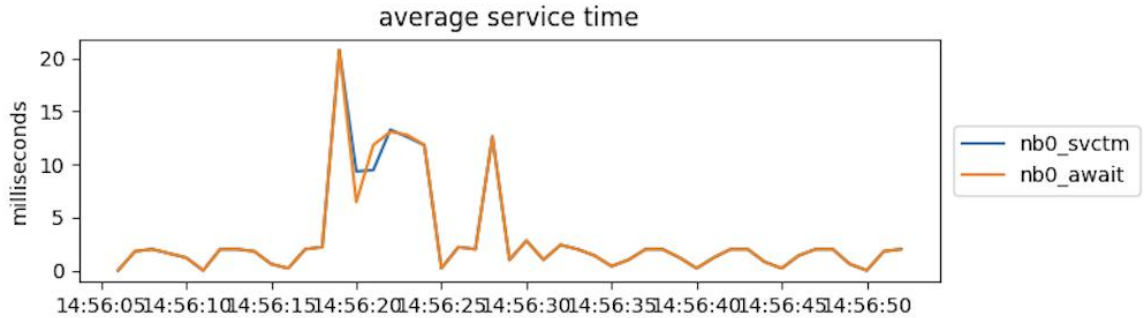
更稳定 网络丢包 5%





Ceph



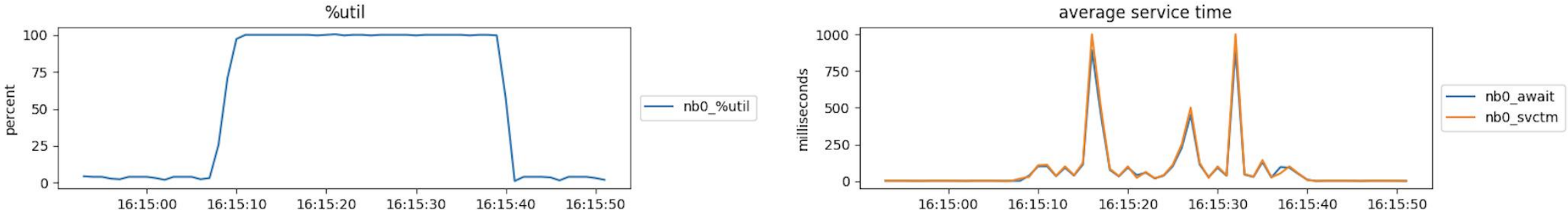


Curve

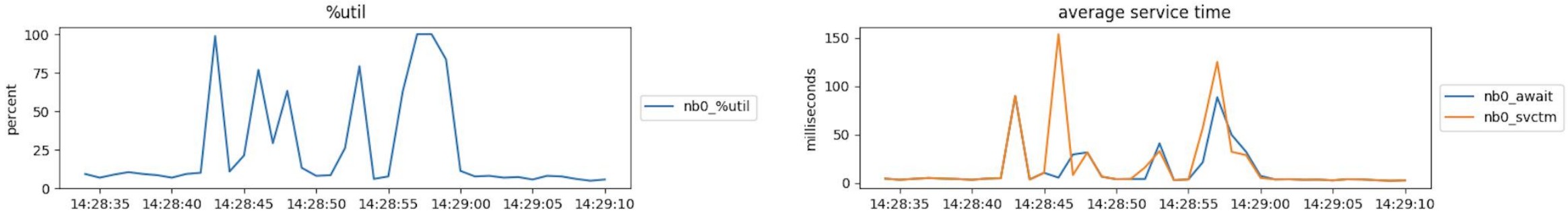
主要亮点

Curve

更稳定 慢盘



Ceph

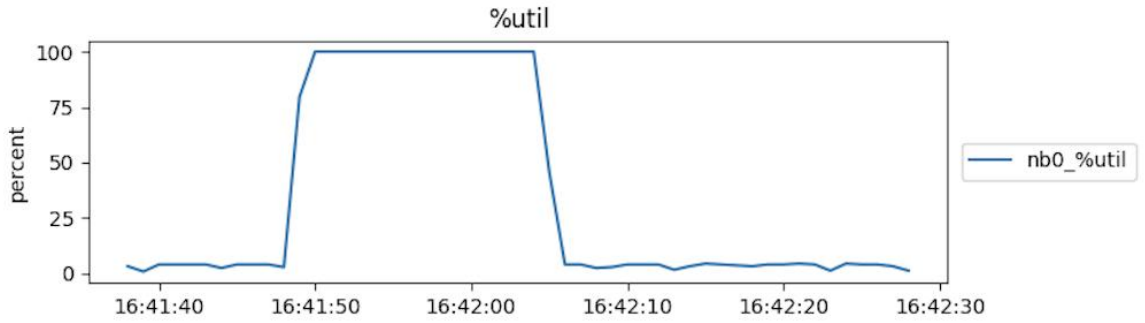


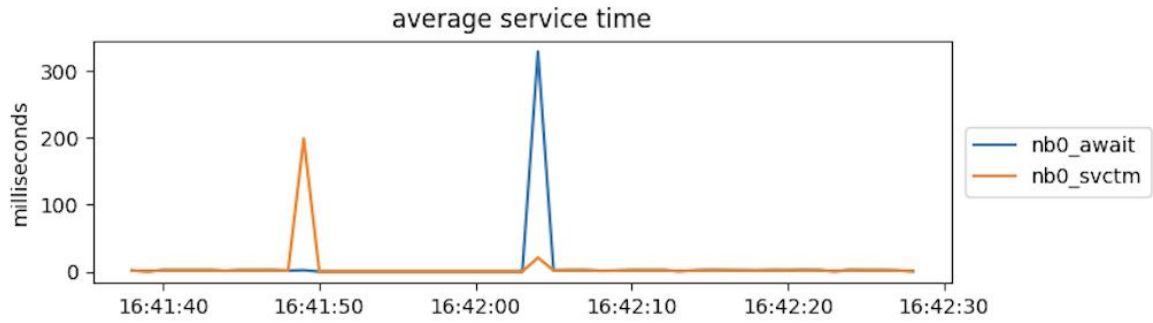
Curve

主要亮点

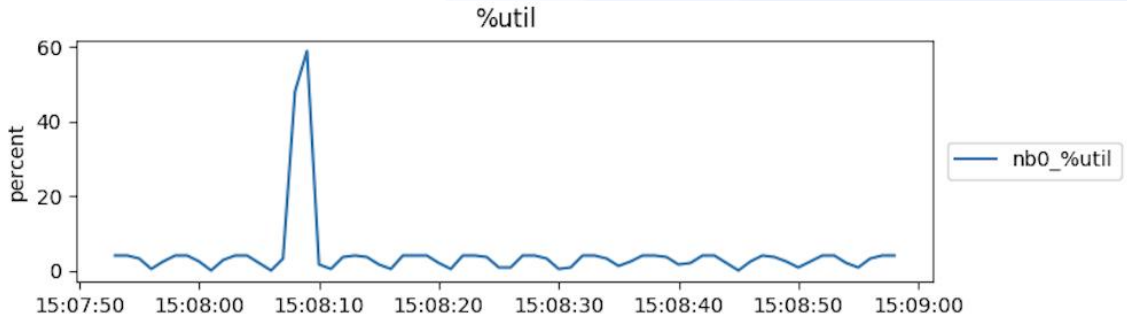
Curve

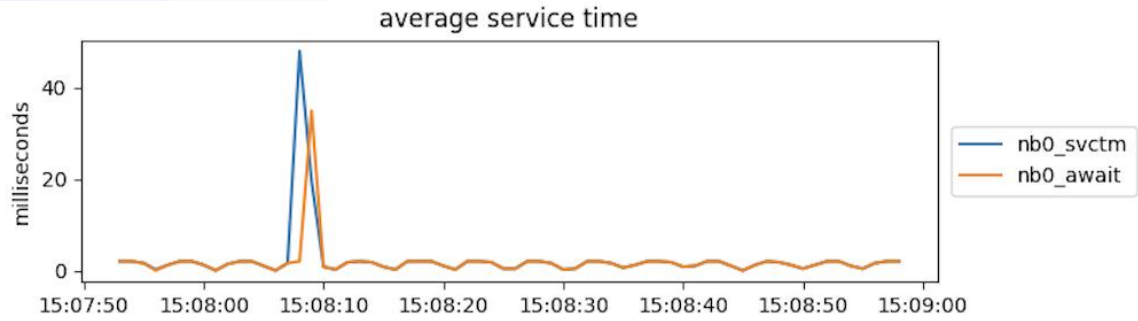
更稳定 机器宕机





Ceph

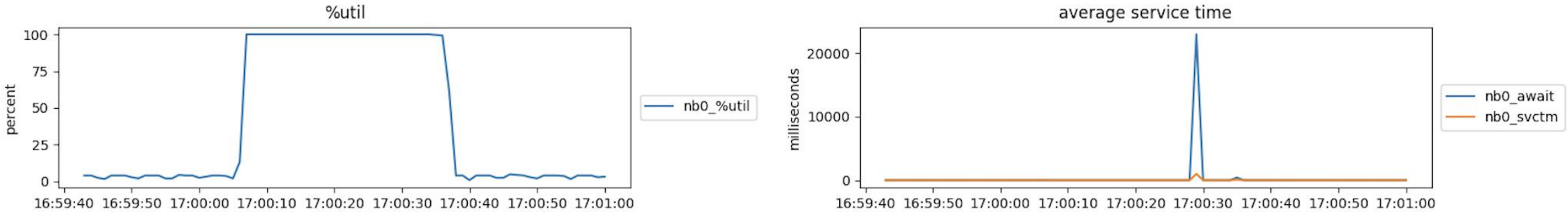




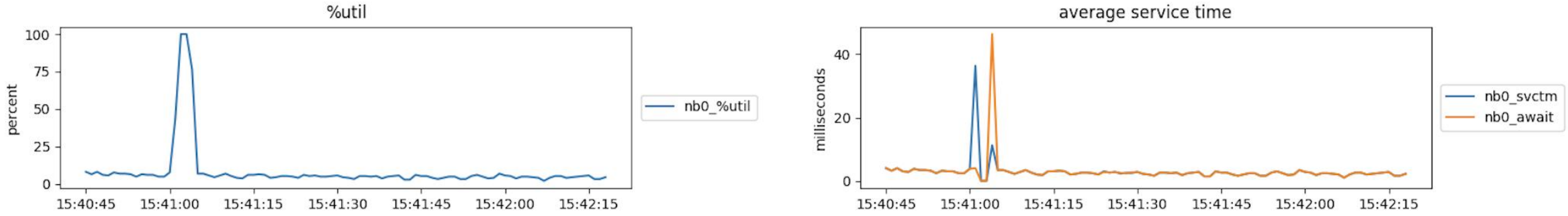
Curve

主要亮点

更稳定 机器卡住



Ceph

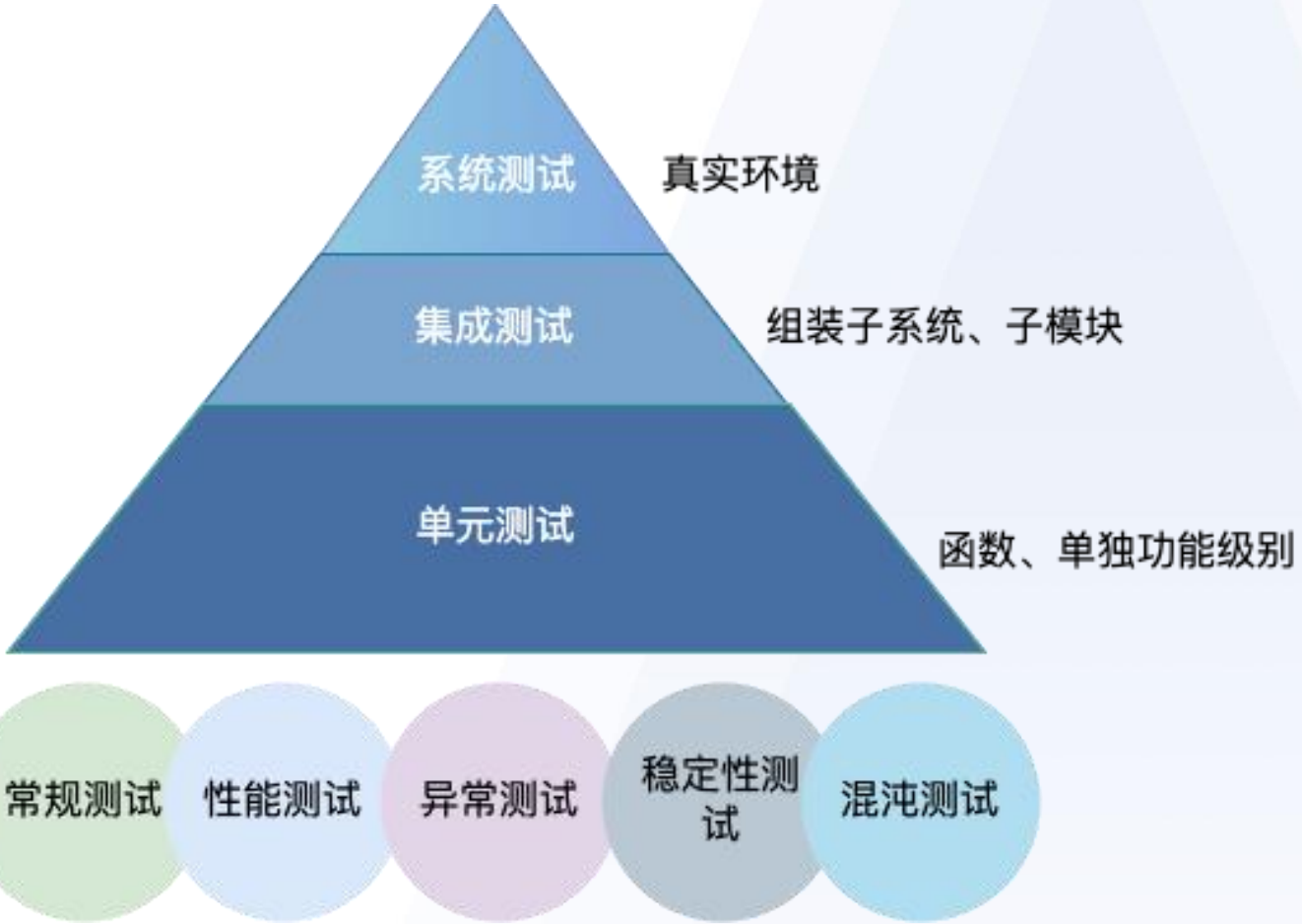


Curve

主要亮点

Curve

高质量 良好的模块化和抽象设计；完善的测试体系



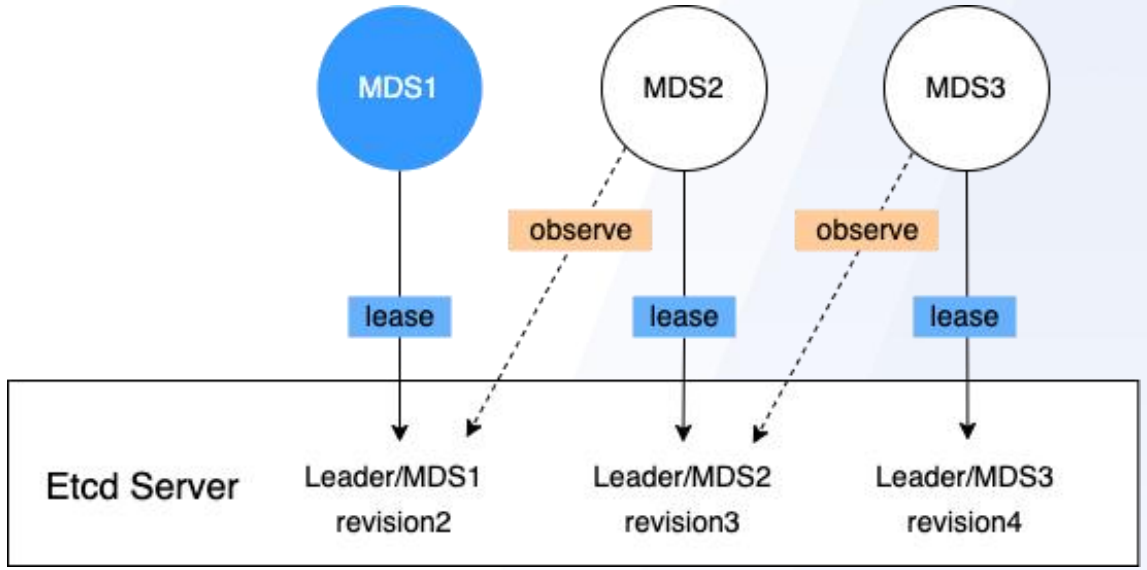
单元测试覆盖率	lines	functions	link
Curve	85.4%	89%	curve
Ceph	37.1%	43.3%	ceph

主要亮点

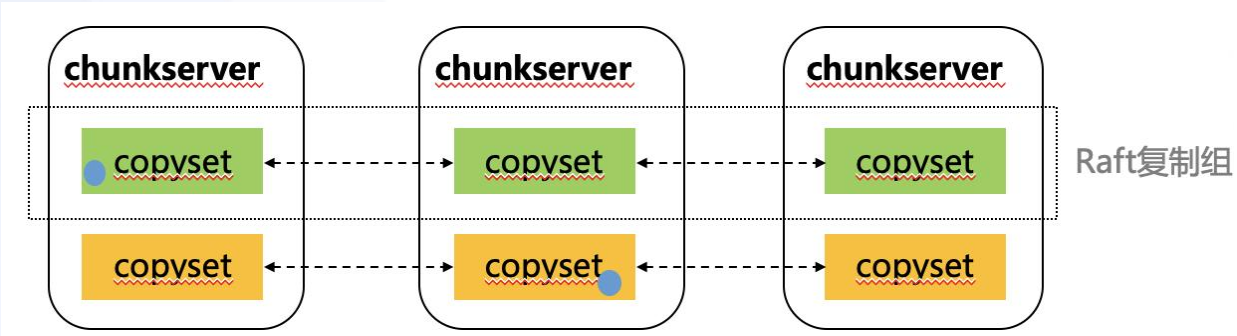
Curve

高可用　核心组件支持多实例部署，系统无单点故障

MDS元数据：通过etcd选主，实现高可用



Chunkserver 数据节点：通过raft实现高可用

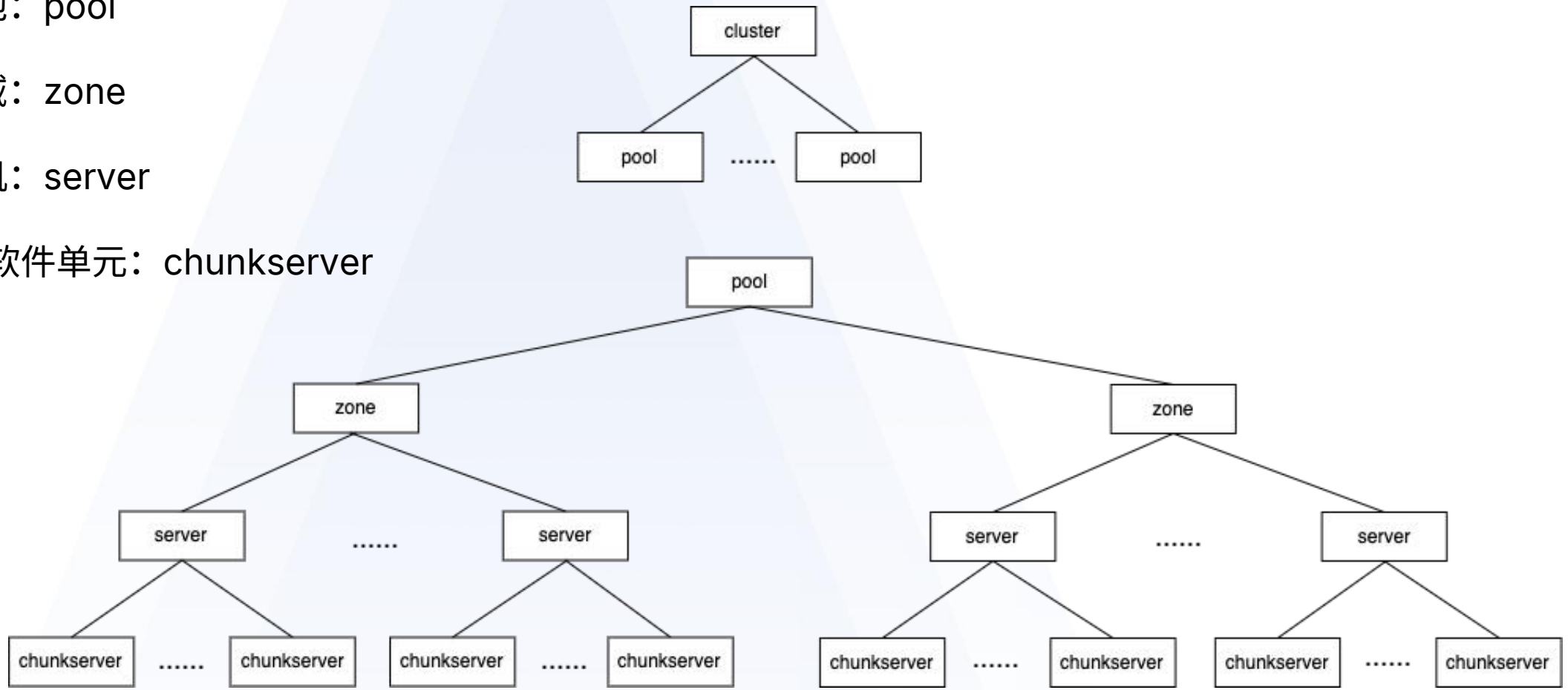


主要亮点

Curve

高可扩展　支持以pool为单位进行扩容，无需进行数据重平衡

- 物理池：pool
- 故障域：zone
- 物理机：server
- 磁盘/软件单元：chunkserver



技术架构先进性-服务端

Curve

中心化节点架构

利用中心节点感知集群节点和负载、容量、异常，进行资源的实时调度与恢复

Quorum机制

异常情况容错更优异，在异常恢复较快的情况下，可以在保证可靠性的同时，提高服务的可用性

文件池chunkfilepool

集群初始化时，创建指定大小的文件，降低IO过程中文件元数据更新的开销，降低IO延迟

IO路径一次持久化

对于用户io，raft大多数写完wal，apply写入缓存即可返回

raft轻量级快照

读写操作是幂等的，raft快照只需要保存文件名，对io无影响

快照上传S3

快照存储到支持S3接口的对象存储，不限制数量

技术架构先进性-客户端

Curve

条带化设计

大io使用条带化技术，在客户端进行分片，提高IO的并发度，让更多的节点参与IO处理

ApplyIndex读

客户端带着appyIndex读，无需走一致性协议，使得读不会被写阻塞，极大的提高了读性能

支持多挂载

Curve块存储一个文件可以挂载到多个客户端，提供块级别的一写多读级别的共享，并提供异常场景下的IO Fence能力保证数据一致性

支持热升级

client端使用client-server架构，版本升级只需要更新server，无需业务停服，对io秒级影响

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，云信等**

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**
  - **ZStack、易思捷、智星云**
  - **新浪、合合信息等**

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

THANK YOU

D I G I T A L S A I L

扫码即可关注