Ceph分布式对象存储系统研究与优化

汪黎

liwang@ubuntukylin.com.cn

提纲

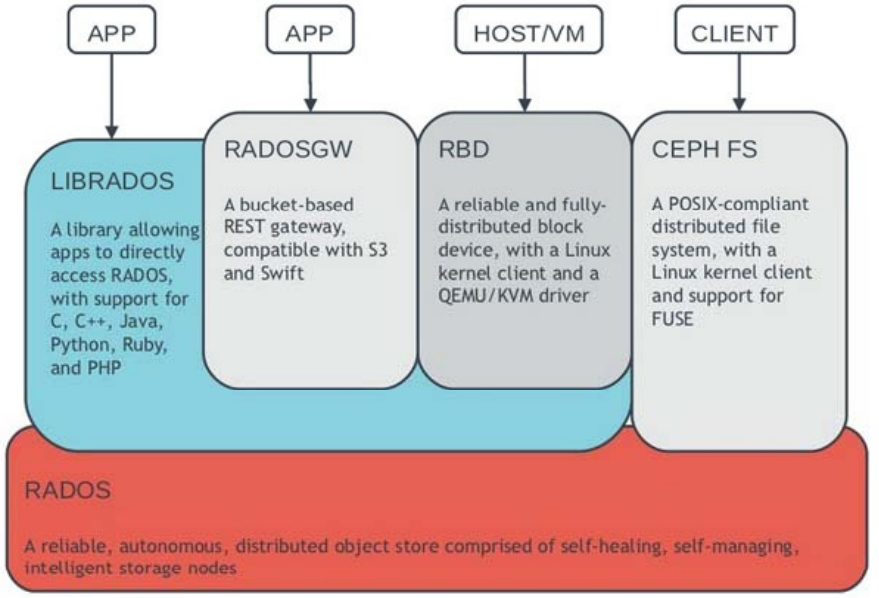
- Ceph介绍
- 我们的工作

概述

- Ceph是Sage Weil 在 UCSC开发的大规模的、可扩展的、开源的分布式存储系统
- Ceph client included in Linux kernel since 2.6.34
- Supported by Openstack since Folsom
- Maintained by Inktank inc.

Ceph架构

- RADOS
  - Ceph的核心组件
  - 提供高可靠、高可扩展的分布式对象存储架构
  - 利用本地文件系统存储对象
- Client
  - Cephfs
  - RBD
  - Radosgw
  - Librados



Ceph Clients

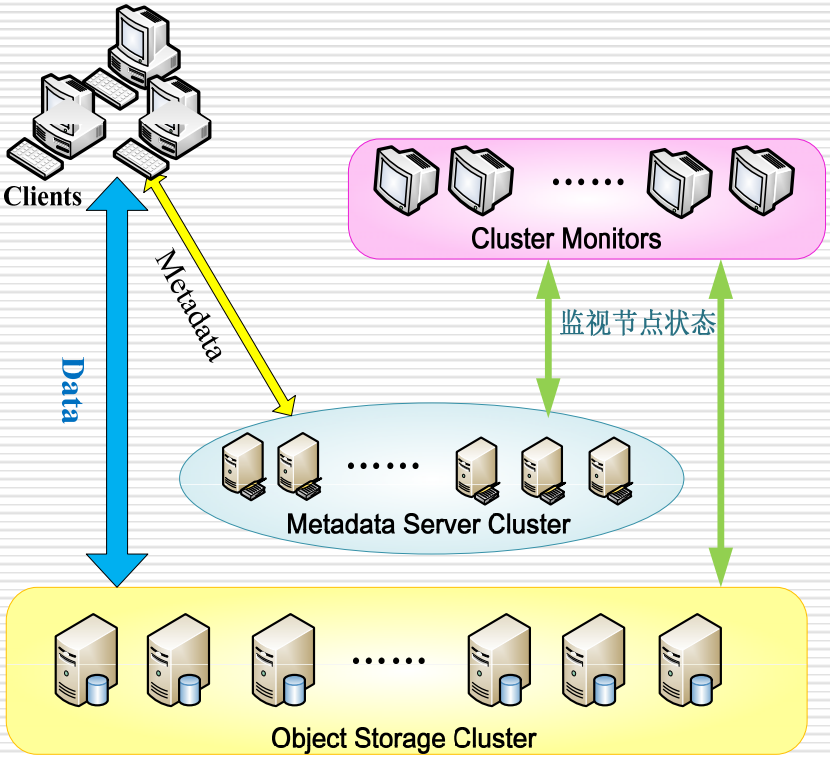
- Cephfs
  - 提供POSIX兼容的分布式文件系统接口，支持kernel、fuse
- RBD
  - 提供高可靠的，分布式的块接口
- Radosgw
  - 提供S3、Swift兼容的对象接口
- Librados
  - 提供rados库调用接口

Rados特色

- 高可靠性
  - 多副本
  - 自动隔离失效节点
  - 数据自动恢复
- 高可扩展性
  - 数据分布式存储
  - 数据透明扩容
- CRUSH副本分布算法

Cephfs组成

- OSD
  - 存储文件数据和元数据
- Monitor
  - 监视整个集群状态
  - 维护并通告集群拓扑结构
- MDS
  - 缓存和同步元数据
  - 管理名字空间
- Client
  - 提供POSIX接口



Cephfs特色

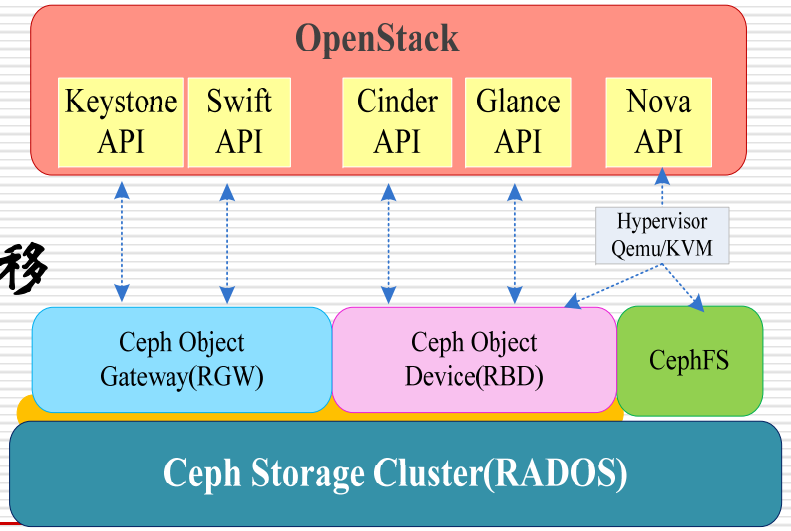
- 数据与元数据分离
  - 数据由OSD服务
  - 元数据由MDS服务
  - 但都存储在OSD上，作为不同的对象，MDS作为元数据cache
- 支持文件分片存储
- OSD、Monitor、MDS均支持集群运行，支持动态增删节点，无单点失效
- 动态子树划分

capability

- 为了维护数据一致性, 由mds给client颁发file cap, 决定client允许进行的操作
- Fb: 允许client进行带缓存的写操作
- Fc: 允许client进行带缓存的读操作
- 对于同一个文件，当有大于1个的writer或1个writer和1个或多个reader时，client将进行无缓存的同步读写，除非打开LAZYIO

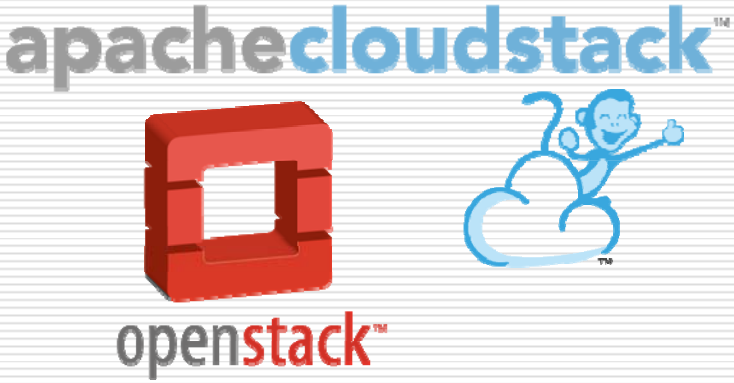
Ceph for OpenStack

- rbd
  - 为Glance、Cinder提供镜像存储
  - 提供Qemu/KVM驱动支持
  - 支持openstack的虚拟机迁移
- rgw
  - 替代swift
- CephFS
  - 提供共享的文件系统存储
  - 支持openstack的虚拟机迁移
- 还可替代hdfs



Ceph应用

- Ceph遵循LGPL协议
- Ceph支持的云计算和大数据平台：OpenStack、CloudStack、OpenNebula、Hadoop等
- Ceph合作伙伴：SUSE、Citrix、Canonical、DELL等

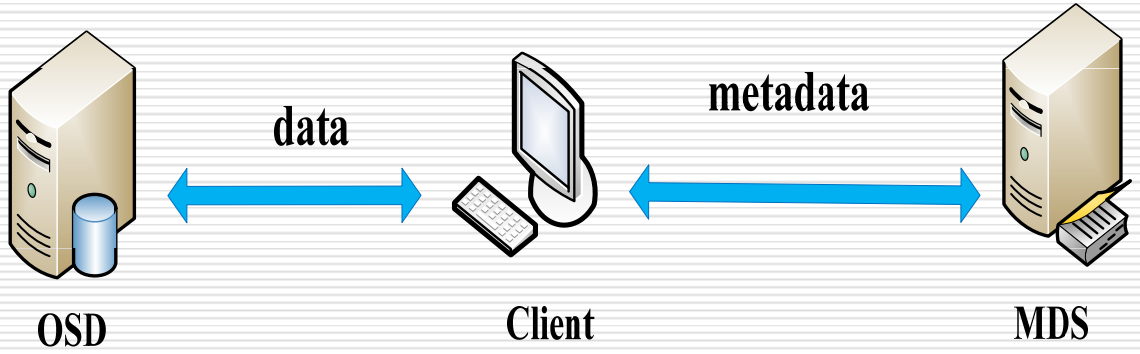


我们的工作

- 基于Ceph的优化
  - 面向海量小文件访问的性能优化
    - Inline data support
  - 面向虚拟化场景的存储优化
    - Fallocate support
  - 面向Cephfs文件系统的优化
    - Cephfs quota support
  - 面向rbd克隆镜像的优化
    - Copy on read

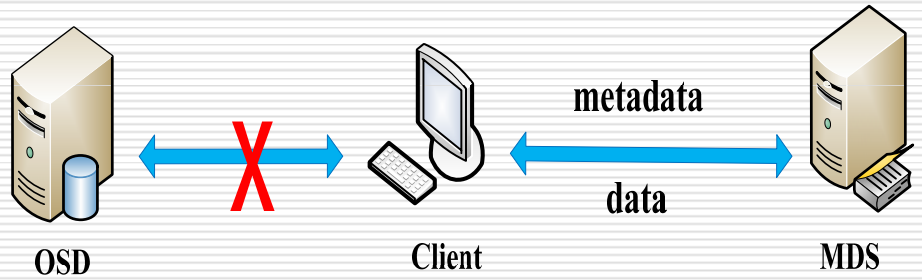
Inline data support

- 动机
  - 元数据与数据分离存储
  - 文件访问需要两次通信：从MDS取元数据、从OSD取数据
  - 小文件访问受限于网络延迟



3.1 Inline data support

- 方法
  - 针对小文件，数据存储在元数据中
  - 打开文件时，元数据服务器将数据与元数据同时推送给客户端
- 优势
  - 节省了数据存储位置计算时间
  - 节省了与OSD的通信



Inline data support

- 文件状态设计
  - Inline 状态：文件所有数据都存放在元数据区域，OSD上不存数据
  - Disable状态：文件所有数据从元数据区域完全迁移至OSD，元数据区域不再存放文件数据
- 状态变迁
  - 记录并维护inine data版本，记录inline状态
  - 状态只能从inline状态变迁到disable状态，不可逆转

Inline data support

- 状态迁移条件
  - 文件大小超过阈值
  - 当client无法执行带缓存的read, write, 而只能同步读写数据时
    - 典型情况: 1 writer + 1 or more readers或 2 or more writers
    - 为保证数据一致性, 此时ceph取消client buffer,强制client进行无page cache的同步读写
    - 继续inline对mds负载较重
    - 迁移处理比较复杂, 因为client不定拥有完整且最新的inline data

Some tricks

- inline data迁移异步进行
- 避免反复推送uptodate inline data
- inline data功能可以开关
- 避免multiple clients同时迁移inline data, 导致修改丢失
- 避免一个client迁移inine data成功,但在未通知mds之前失效, 以后新的inline data不能覆盖osd上旧的inline data

Inline data support

- 数据结构设计
  - inode添加两个域
    - uint64_t inline_version
    - bufferlist inline_data
  - inline_version
    - 标识inline data的版本
    - 维护数据的一致性和正确性
    - 从1开始单调递增，最大值为2 ^ 64 - 1，即CEPH_INLINE_DISABLED，定义为inline data support的disabled状态
  - inline_data
    - 存放小文件的实际文件内容

文件操作流程

- **Create**
  - 初始化inline_data及inline_version
  - Client: inline_version=0
  - MDS：inline_version=1
- **Open**
  - 元数据查询：getattr
  - MDS向Client发送元数据
  - 为客户端分配操作文件的capbility，如FC、FB权限

文件执行流程

- Read
  - 在客户端inline_version<CEPH_INLINE_DISABLED时，表示是inline data状态
  - 判断客户端是否有读缓存的权限
    - 如果有，返回inline data
    - 如果没有，则将inline data迁移到OSD，执行原read流程，再将inline_version设置为CEPH_INLINE_DISABLED

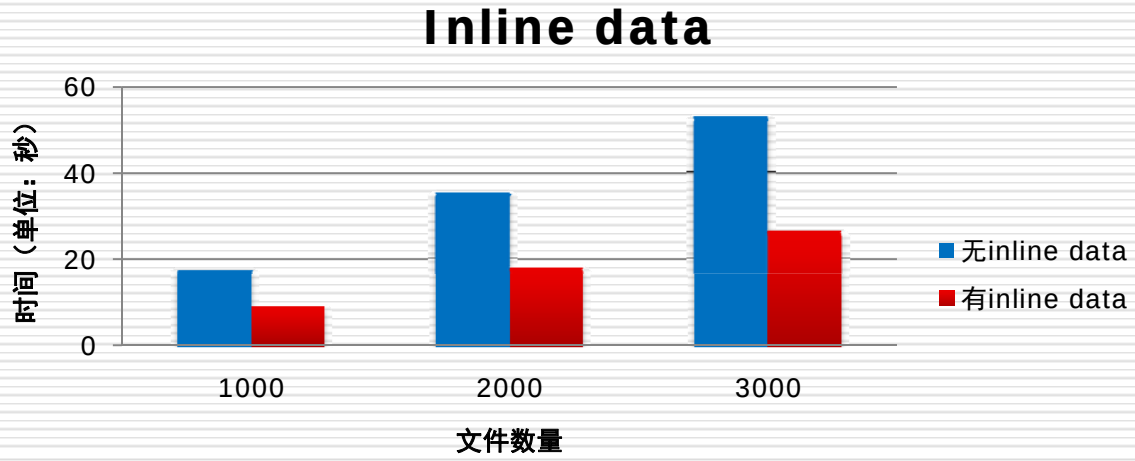
文件执行流程

Write

- 在Client inline_version< CEPH_INLINE_DISABLED时，表示是inline data状态
- 判断Client是否有FB权限
  - 如果有FB权限，判断写文件范围是否超出设置的inline阈值
    - 如果未超出，则将文件数据写入到inline data，inline_version+1, mark caps dirty, 异步将inline data写到mds
    - 如果超出，执行如下步骤：
      - 将 inline data迁移至OSD
      - 执行原有写操作写OSD
      - 迁移完成后，将Inode中的inline data清空，将inline_version置为CEPH_INLINE_DISABLED
  - 如果没有权限，执行超出inline 阈值的相同流程

Inline data support

- 测试环境：
  - 操作系统：Ubuntu 12.04 x86_64
  - ceph版本：0.56.6
  - Ceph部署：15个OSD， 1个MDS， 1个MON
- 测试方法：
  - 顺序读取一定数量的1k小文件，统计总共花费的时间
- 测试结果：

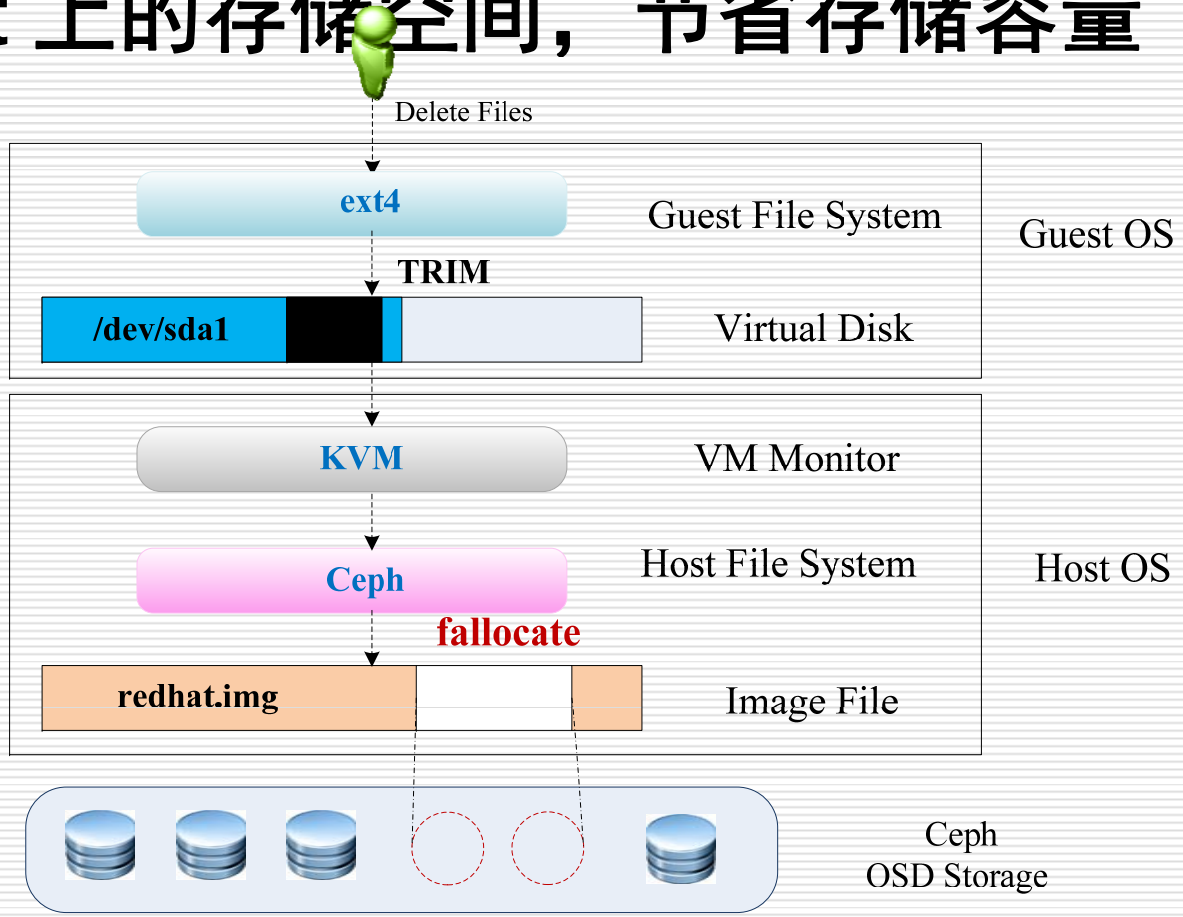


Fallocate support

- 动机
  - Ceph用于虚拟镜像文件存储时，不能在虚拟机文件系统删除文件时回收空间
- 方法
  - 实现fallocate，支持将文件的指定区域清空，释放对应的存储对象
- 优势
  - 可在虚拟机删除文件时回收主机存储空间

fallocate用途

- Ceph用作虚拟机镜像文件存储时，可以释放host 上的存储空间，节省存储容量

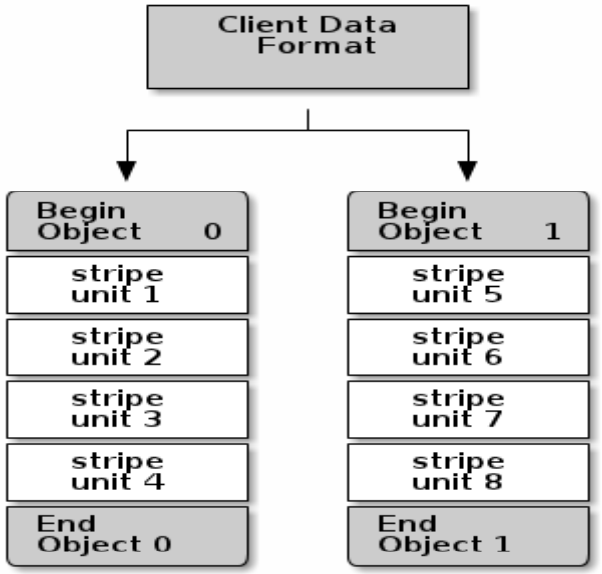


算法

- OSD object处理
  - 当hole覆盖了整个object时，将该object从osd上删除
  - 如果hole只涉及object的一部分，将该部分写0
- Page cache
  - 如果hole覆盖page cache中的整个page，将该page写回，删除
  - 如果hole只涉及page的一部分，将该部分写0

Some tricks

- 如果hole覆盖第一个object,由于其包含有backtrace等元数据信息,不将其删除,而是将其truncate到0
- Data striping支持
  - 把连续的数据分割成相同大小的数据块,把每段数据分别写入到不同对象上的方法
  - 参数
    - Object Size
    - Stripe Width
    - Stripe Count

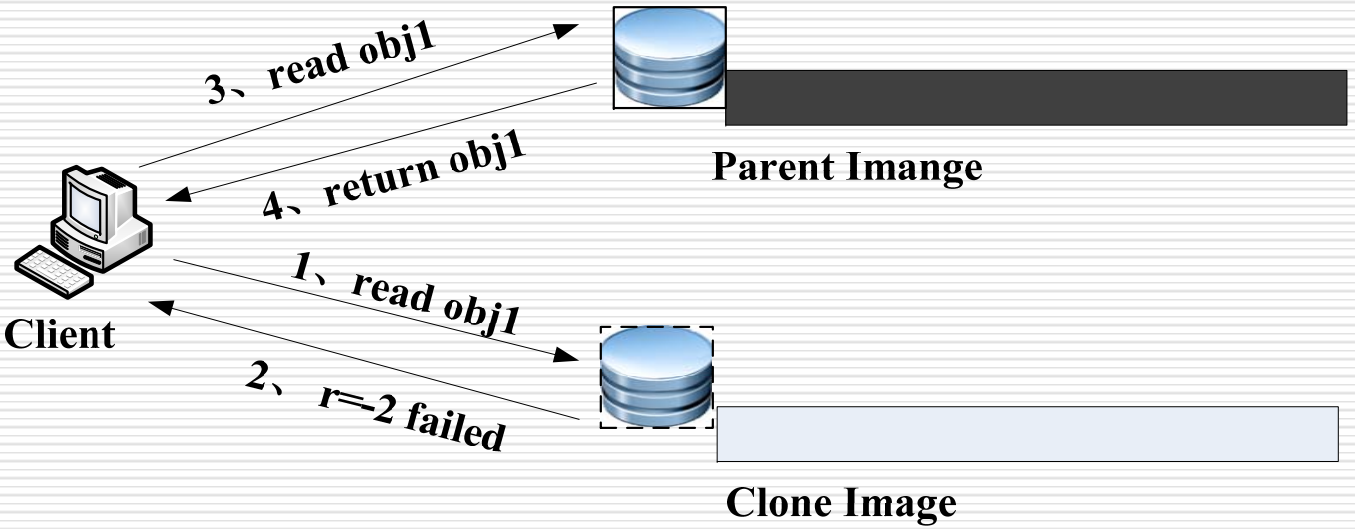


Cephfs Quota Support

- 动机
  - 为Cephphfs实现quota支持
  - quota针对目录设置，可限制目录下存放的文件容量及数量
  - Cephfs没有一个统一的UID/GID机制，传统的基于用户和组的配额很难使用
  - Cephfs一般跟应用配合使用，应用自己记录用户信息，将用户关联到对应的Cephfs目录
- Some tricks
  - 对于设置了quota的节点的元数据服务仅由一个mds进行
  - mds维护更新quota信息，推送给client进行检查
  - 从根目录到当前目录路径上所有设置了quota的节点都要进行quota检查
  - 处理用户挂载一个子目录做为根目录的情况

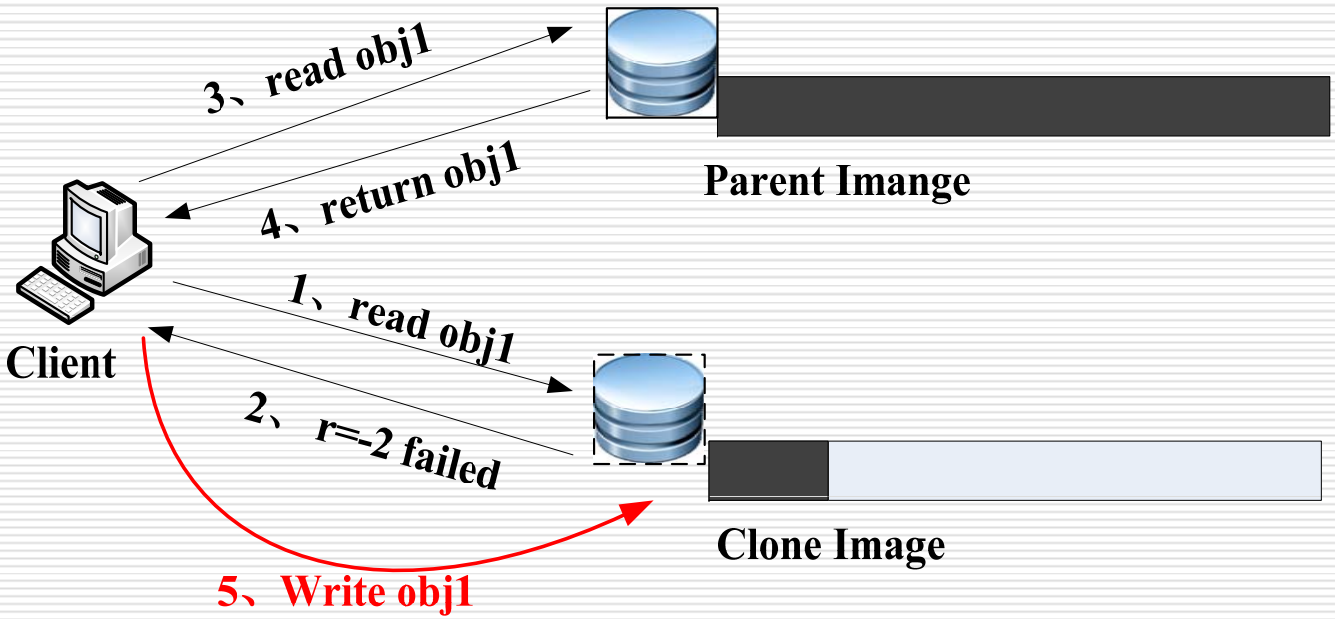
Rbd copy on read

- 动机
  - RBD克隆是基于传统的Copy on write实现，因此，只有修改过的数据会保存在克隆镜像，其它未修改的数据则从父镜像检索
  - 当读取对象在克隆镜像找不到时，从父镜像检索读取，而父镜像和克隆镜像可能存放于不同的host或者pool中，造成网络延时较大，且父镜像所在节点负载较重



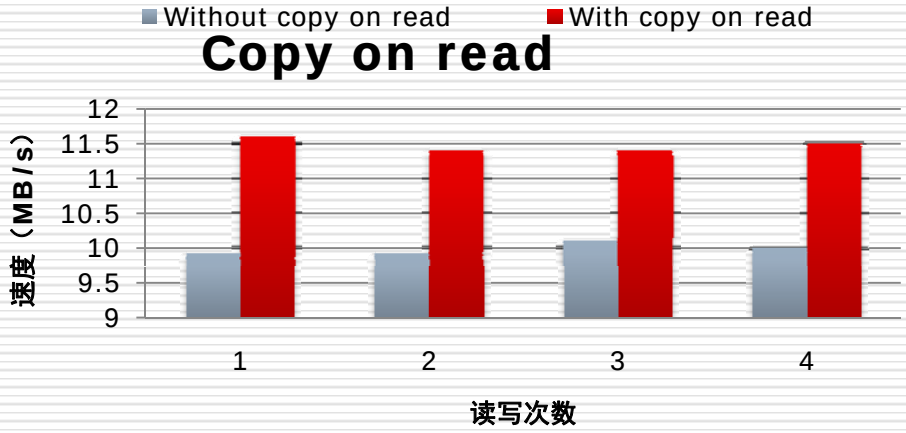
Rbd copy on read

- 方法
  - 读取镜像时异步的复制对象副本到克隆镜像
  - 下次读取时直接从克隆对象读取



Copy on read

- 优势
  - 提升BRD镜像中对象第二次读的性能
- 关键技术
  - 异步复制对象时的流程处理
- 评测



谢谢！