誉天教育
YU TIAN EDU

Linux云计算架构师涨薪班

RBD高级特性

学习目标

- RBD缓存机制
- RBD特性管理
- RBD快照管理
- RBD克隆管理
- RBD导入导出

RBD客户端缓存

- RBD客户端两种实现方式
  - librbd: RBD块设备在客户端由librbd通过用户空间的库实现，librbd就是利用librados与RBD进行交互的，librbd主要用于为虚拟机提供块设备，librbd无法使用Linux页面缓存，因此使用librbd的内存来进行缓存
  - krbd：使用原生Linux内核模块krbd进行挂载，内核驱动可利用 Linux 页缓存来提升性能
- 缓存模式
  - 直写（透写）缓存： 数据直接写入osd磁盘
  - 回写缓存: librbd 库将数据写入到服务器的本地缓存中，周期性刷盘到OSD（如主机故障有丢失数据的风险）
- RBD缓存
  - RBD缓存是使用客户端上的内存

RBD缓存参数

回写缓存:
考量两个值：未清空缓存字节数 U 和最大脏缓存字节数 M。
如果 U < M，则确认写入，否则在 数据写回到磁盘后确认，直到 U < M 为止

直写缓存:
将最大脏字节数设置为 0，以强制使用直写模式。数据在所有相关的 OSD 日志中写入并清空 时，Ceph 集群确认写入

参数	描述	默认值
rbd_cache	启用 RBD 缓存。值=true|false。	TRUE
rbd_cache_size	每个 RBD 镜像的缓存大小，以字节为单位。值=n。	32 MB
rbd_cache_max_dirty	每个 RBD 镜像允许的最大脏字节数。值=n。	24 MB
rbd_cache_target_dirty	每个 RBD 镜像开始抢先清空的脏字节数。值=n。	16 MB
rbd_cache_max_dirty_age	清空前的最大页面期限，以秒为单位。值=n。	1
rbd_cache_writethrough_until_flush	启动直写模式，直至执行第一次清空。值=true|false。	TRUE

RBD特性

名称	描述
layering	用于启用克隆的分层支持。
striping	用于提高性能的分条 v2 支持，由 librbd 提供支持。
exclusive-lock	独占锁定支持。
object-map	对象映射支持（需要 exclusive-lock）。
fast-diff	快速 diff 命令支持（需要 object-map 和 exclusive-lock）。
deep-flatten	扁平化 RBD 镜像的所有快照。
journaling	日志支持。
data-pool	EC 数据池支持。

RBD特性

特点	描述信息	ID
layering	是否支持克隆	1
striping	是否支持数据对象间的数据条带化，提升性能只支持librbd的客户端使用(内核态)	2
exclusive-lock	是否支持分布式排他锁机制以限制同时仅能有一个客户端访问当前 image	4
obiect-map	是否支持obiect位图，主要用于加速导入，导出及已用容量统计等操 作，依赖于exclusive-lock特件	8
fast-diff	是否支持快照间的快速比较操作，依赖于object-map特性	16
deep-filatten	是否支持克隆分离时解除在克隆image时创建的快照与其父image之间的关联关系	32
journaling	是否支持日志，即是否支持记录image的修改操作至日志对象: 依赖于exclusive-lock特性	64
data-pool	是否支持将image的数据对象存储于纠删码存储池，主要用于将image的元数据与数据放置于不同的存储池	128

RBD镜像特性管理

- 在RHEL7/Centos7中映射rbd可能会报错
  - RBD image feature set mismatch. You can disable features unsupported by the kernel with "rbd feature disable test object-map fast-diff deep-flatten"
  - 说明RBD启用了一些内核不支持的功能，需要关闭之后才能正常映射
- 启用RBD镜像特性
  - rbd feature enable rbd/test layering
- 禁用RBD镜像特性
  - rbd feature disable rbd/test layering

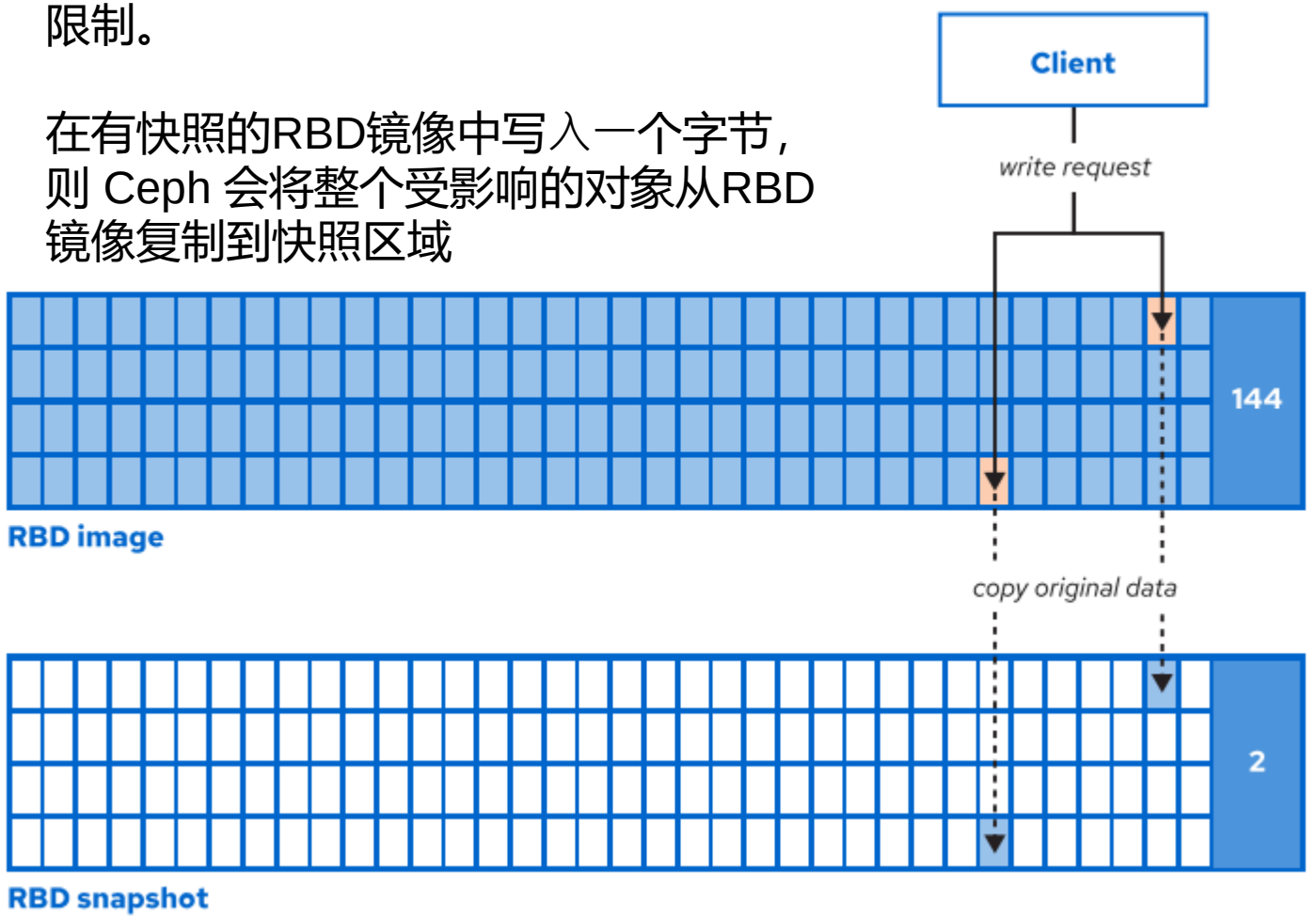
RBD快照

- RBD快照是创建于特定时间点的RBD镜像的只读副本
- RBD快照使用写时复制(COW)技术来最大程度减少所需的存储空间
- 在将写入 I/O 请求应用到 RBD 快照镜像前，集群会将原始数据复制到 I/O 操作所影响对象的 PG 中的另一区域
- 快照在创建时不会占用存储空间，但会随着所包含对象的变化 而增大。
- RBD 镜像支持增量快照

COW原理

快照 COW 步骤在对象级别上运行，不受对 RBD 镜像发出的写入 I/O 请求大小的限制。

在有快照的RBD镜像中写入一个字节，则 Ceph 会将整个受影响的对象从RBD镜像复制到快照区域



誉天教育
YU TIAN EDU

快照管理

- 创建快照
  - rbd snap create pool/image@firstsnap
  - 注意: 不要在文件系统未冻结时拍摄文件系统快照，因为这样会损坏快照的文件系统
  - fsfreeze -f mountpoint 冻结文件系统 fsfreeze -u mountpoint 解冻文件系统
- 查看快照
  - rbd snap ls pool/image
- 回滚快照
  - rbd snap rollback pool/image@firstsnap
- 删除快照
  - rbd snap rm pool/image@firstsnap

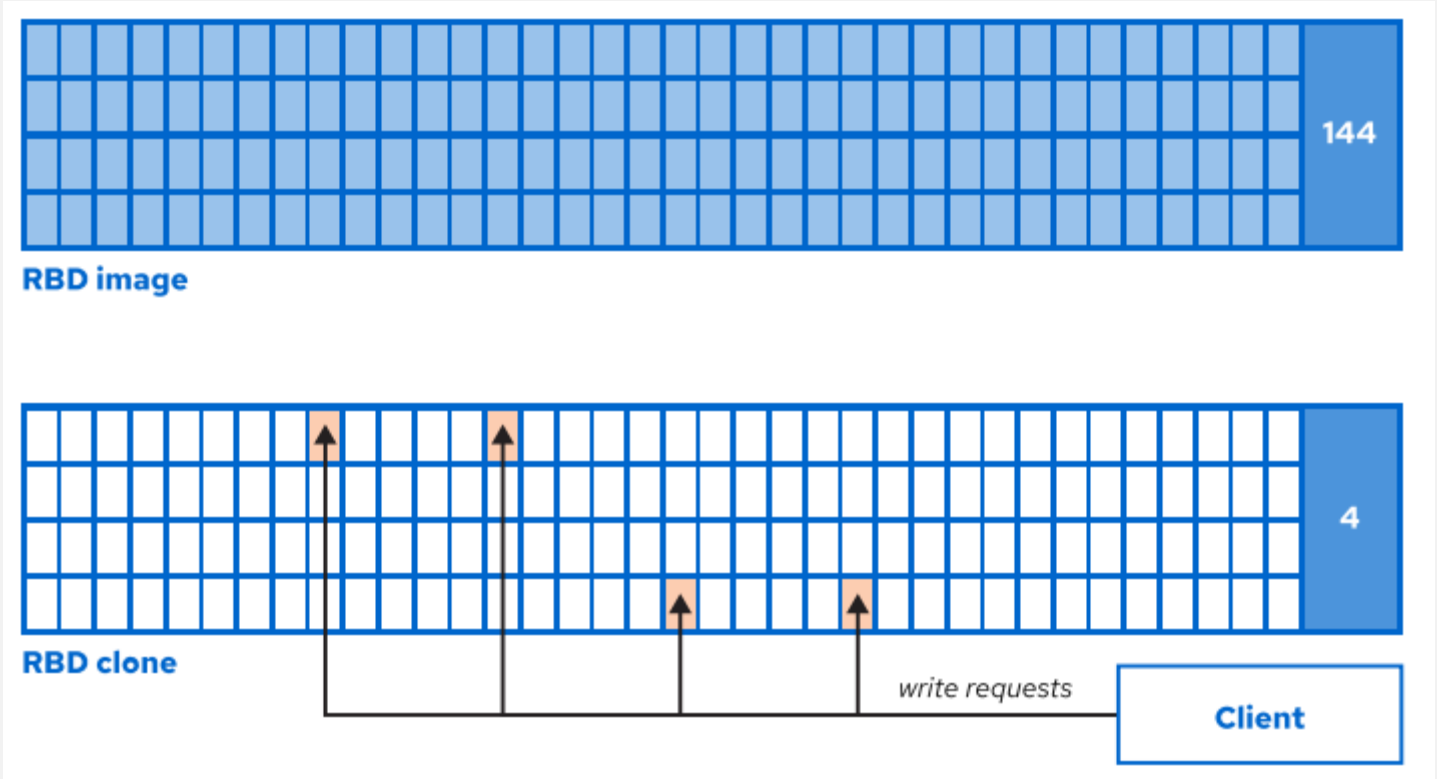
还原快照注意点

- 还原快照步骤：
  - 1.卸载文件系统
  - 2. 取消映射
  - 3. 还原快照
  - 4. 重新映射挂载查看
- 注意点:
  - Rolling back to snapshot: 0% complete...failed.

    rbd: rollback failed: (30) Read-only file system

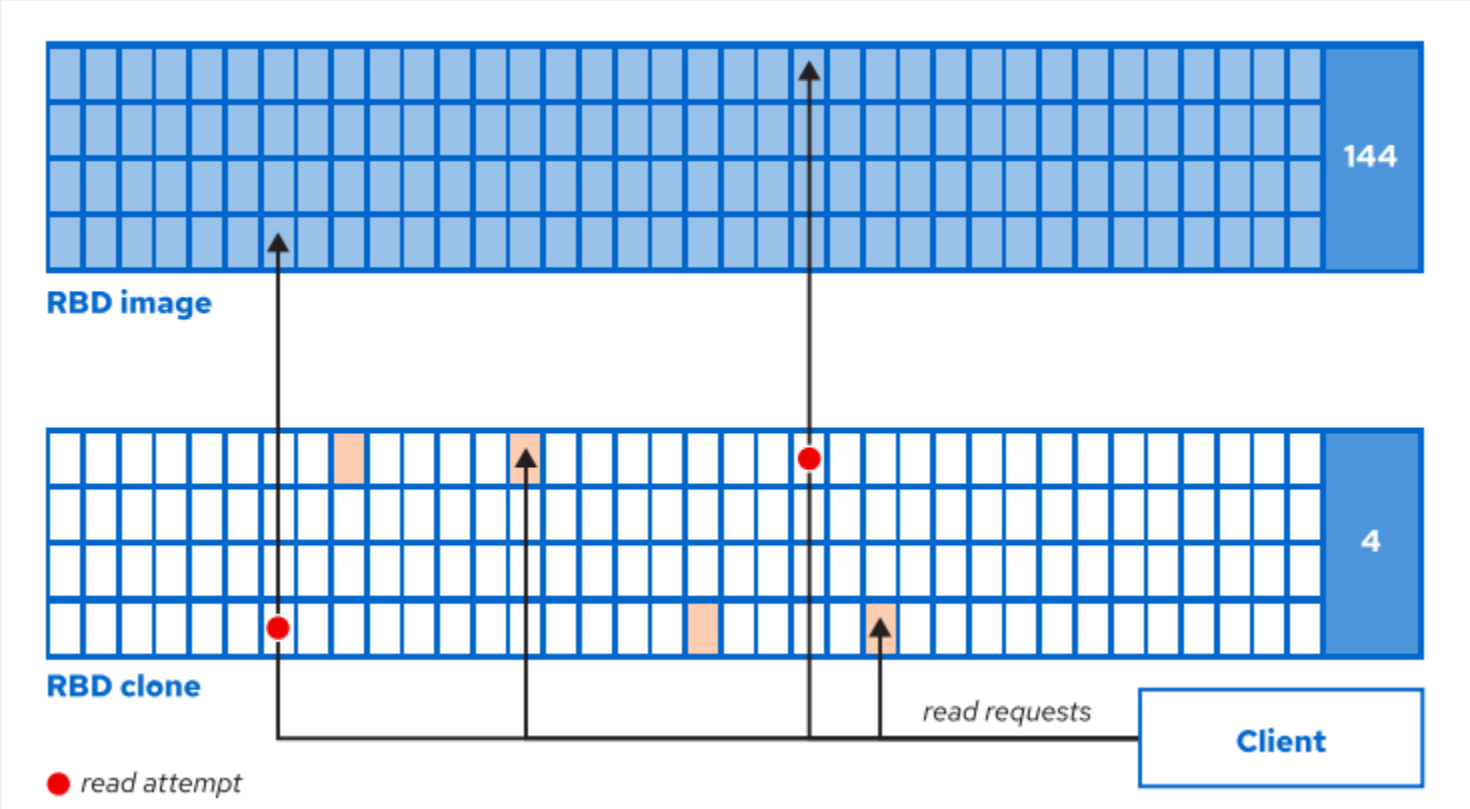
    因rbd默认开启独占锁，卷在被使用中其他进程无法操作卷，导致还原失败

RBD克隆



RBD 克隆是 RBD 镜像的可读写副本，它将受保护的 RBD 快照用作基础镜像

COR原理



父RBD快照和克隆相同的数据会直接从父快照读取 - 对于客户端而言读取效率低下

COR读加速: COR在第一次读取时将对象复制到克隆

克隆管理

- 创建快照
  - rbd snap create pool/image@snapshot
- 保护快照
  - rbd snap protect pool/image@snapshot
- 使用受保护创建克隆
  - rbd clone pool/imagename@snapshotname pool/clonename
- 列出克隆镜像
  - rbd children [pool-name/]image-name@snapshot-name
- 将克隆镜像转换为独立镜像
  - rbd flatten [pool-name/]child-image-name

RBD导入导出

- 利用RBD的导出与导入机制，可以在同一集群中或另一套集群中拥有完整访问性的 RBD 镜像副本
- RBD导入导出方式： 全量导入导出　增量导入导出
- RBD导入导出应用场景：
  - 利用实际的数据卷测试新版本
  - 利用实际的数据卷运行质量保障流程
  - 实施业务连续性方案
  - 将备份进程从生产块设备分离

全量导入导出

- 主集群全量导出

  将正在使用的RBD镜像卸载

  ```
  rbd unmap pool-name/images
  ```

  导出RBD镜像到文件

  ```
  rbd export  pool-name/images  file-name
  ```

- 从集群全量导入

  准备存储池

  ```
  ceph osd pool create  demo
  ```

  导入镜像到存储池

  ```
  rbd  import  file-name pool/images
  ```

增量导入导出

- 差异导出
  - 导出从创建镜像到第一次快照之间的差异数据
    - rbd export-diff pool/images@snap file-name
  - 导出镜像到第一次快照和第二次快照之间的差异数据
    - rbd export-diff --from-snap snap1 pool/images@snap2 file-name
- 差异导入
  - rbd import-diff file-name pool/images

誉天教育
YU TIAN EDU

Thank you