Ceph中国社区公开课

----普通人

www.ceph.org.cn

Agenda

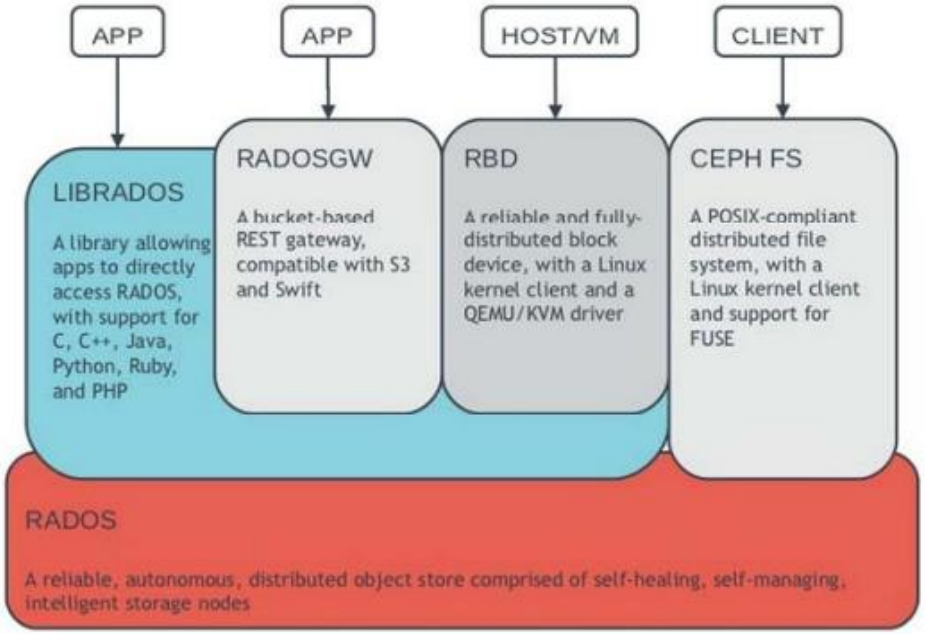
- Ceph的历程
- Ceph的架构
- Ceph的组件
- Ceph的应用
- Ceph的实战

Ceph的历程

- Ceph是Sage Weil 在加州大学专为博士论文设计的新一代自由软件分布式文件系统
- Ceph client included in Linux kernel since 2.6.34
- Supported by Openstack since Folsom
- Acquired by Redhat since 2015.4.30

Ceph架构

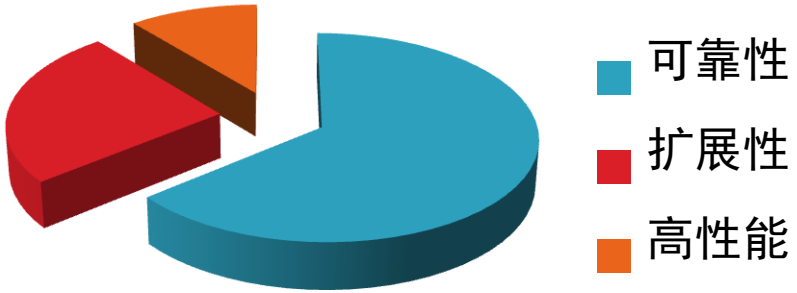
- Rados
  - Ceph的核心组件
  - 提供高可靠、高可扩展、高性能的分布式对象存储架构
  - 利用本地文件系统存储对象（ext4,xfs等）
- Client
  - RBD
  - Radosgw
  - Librados
  - Cephfs



Rados特征

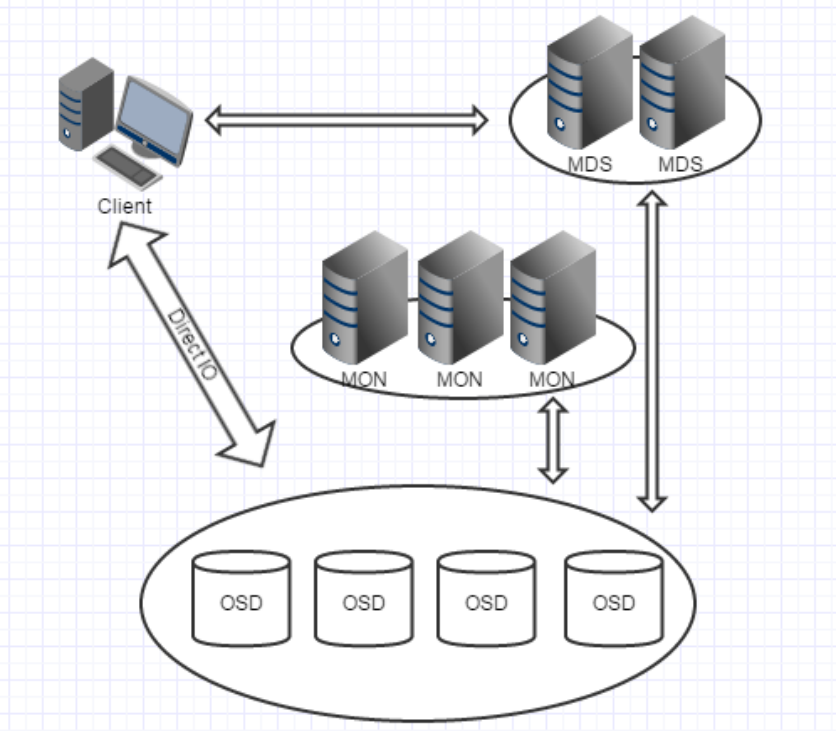
- 高可靠性
  - 多副本
  - 自动隔离失效节点
  - 数据自动恢复
- 高可扩展性
  - 数据分布式存储
  - 数据透明扩容
- 高性能
  - IO“聚合”
- CRUSH分布规则

特性比例



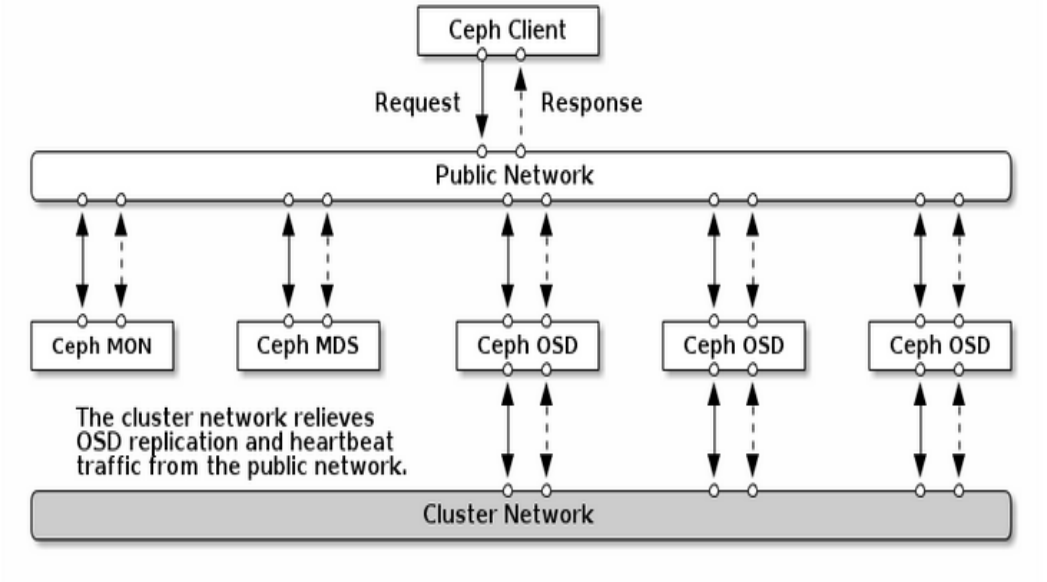
Ceph的组件

- OSD
  - 存储文件数据和元数据
- Monitor
  - 监视整个集群状态
  - 维护集群Map
- MDS
  - 缓存和同步元数据
  - 管理名字空间



Ceph的网络

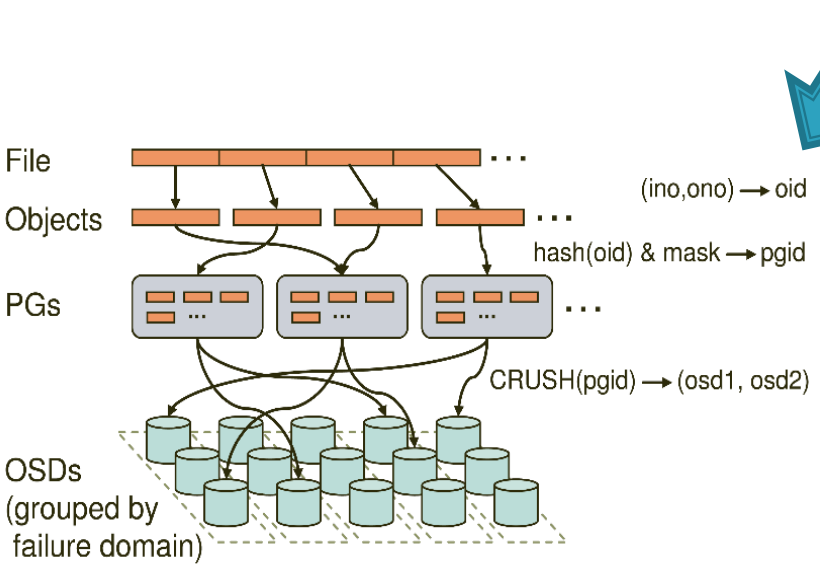
- Public network
  - Client ⟷ OSD
  - Client ⟷ MDS
  - Client ⟷ MON
- Cluster network
  - OSD ⟷ OSD

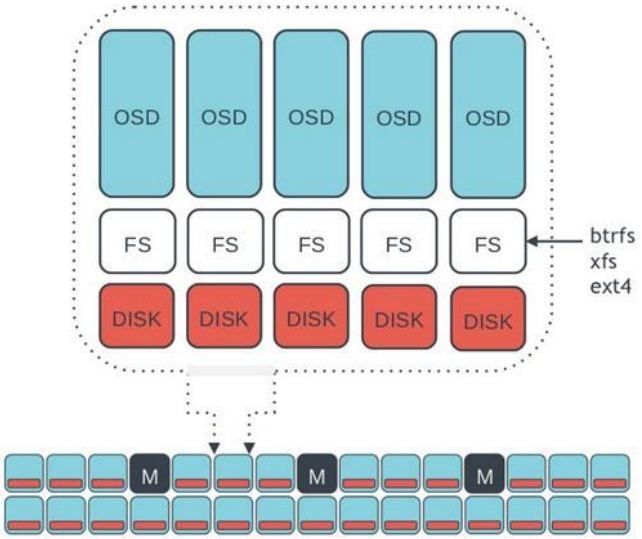


数据流向

Data→obj→PG→Pool→OSD

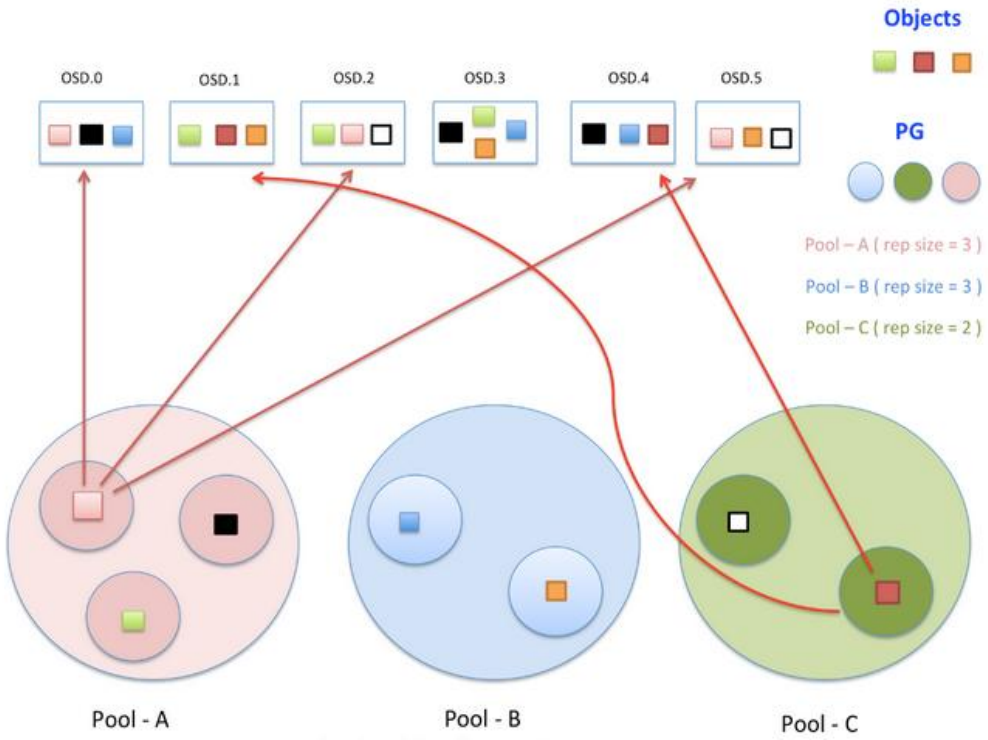
ID	Binary Data	Metadata
1234	01010101010101001101010010 01011000010101001101010010 01011000010101001101010010	name1 value1 name2 value2 nameN valueN





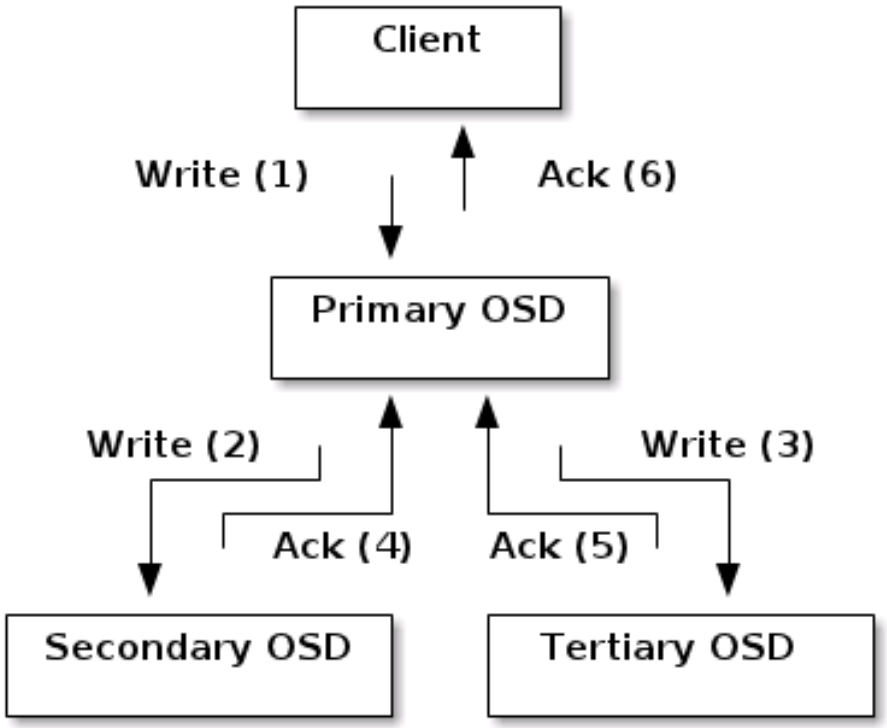
数据概念

- Object
- PG(place group)
- Pool



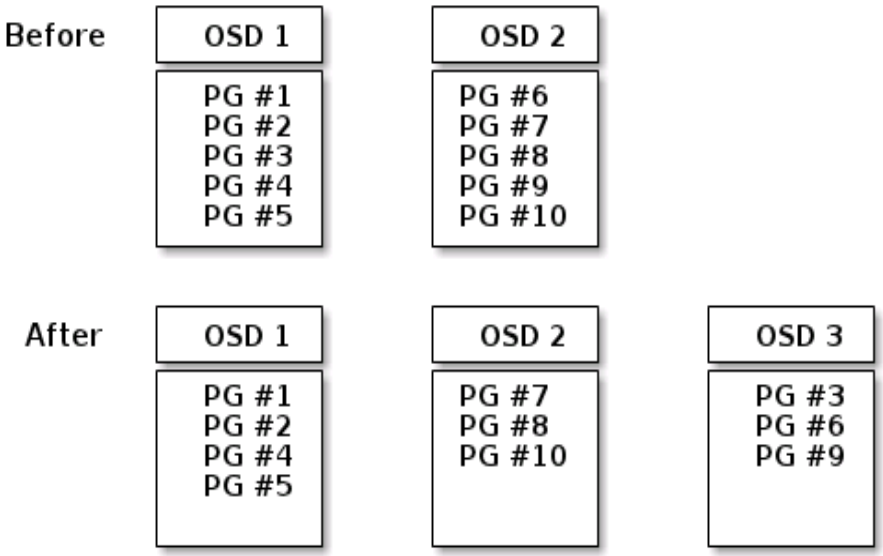
数据复制

- 所有读写都集中在Priamry OSD
- 数据同步自主完成
- 落盘才返回ack
- 数据的强一致性



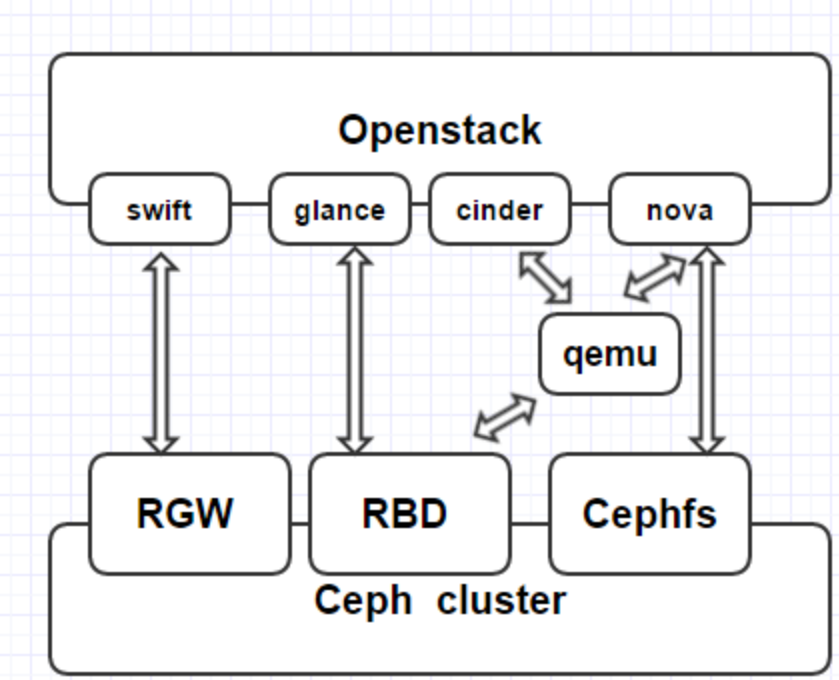
数据重分布

- 影响因素
  - OSD
  - OSD weight
  - OSD crush weight



Ceph应用

- 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
  - 替换swift
  - 网盘
- Cephfs
  - 提供共享的文件系统存储
  - 支持openstack的虚拟机迁移



Ceph实战

操刀实战

www.ceph.org.cn

部署工具

Ceph-deploy

部署mon

- 安装
  - Ceph-deploy install node1 node2 …
  - Ceph-deploy new   node1 node2 …
  - Ceph-deploy mon   create-initial
- 删除
  - Ceph-deploy mon destroy node1
- 添加
  - Ceph-deploy mon add [--address 192.168.1.10]

部署OSD

- 添加
  - Ceph-deploy osd create \
    - [--zap-disk] node:sda:[/dev/sdb]
- 删除
  - Ceph-deploy **error**
  - 手动
    - Service ceph-osd stop id=x
    - ceph-osd --flush-journal -i x
    - ceph osd out osd.x
    - ceph osd crush remove osd.x
    - ceph auth del osd.x
    - ceph osd rm x
    - Rm -rf /var/lib/ceph/osd/ceph-osd/

部署MDS

- 添加MDS（0.9以上需要自建pool并初始化）
  - ceph-deploy mds node1 node2 …
  - ceph osd pool create cephfs_data <pg_num>
    ceph osd pool create cephfs_metadata <pg_num>
  - ceph fs new <fs_name> <metadata> <data>

- 删除MDS
  - ceph-deploy mds node1 node2 …

Pool管理

- PG计算
  - (Target PGs per OSD) * (OSD #) * (%Data)
  - ----------
    Size
- 参考 http://ceph.com/pgcalc/
- 创建pool
  - osd pool create <poolname> <int[0-]> {<int[0-]>} {replicated|erasure}
- 修改/获取参数
  - osd pool set/get <poolname> size|min_size

“找”对象

- rados mkpool test
- rados lspools
- ceph osd lspools
- rados -p test put my-object my-object
- rados -p test stat my-object
- ceph osd map test my-object
- find /var/lib/ceph/osd/ceph-x/current/ -name

小技巧

- Tell
  - ceph tell osd.* injectargs "--rbd_default_format 2 "
- Admin socket
  - ceph daemon osd.1 config show | less
  - ceph daemon mon.ubuntu-ceph-06 config show | less

玩转块存储

- 建块
  - Rbd create –size 100 [--image-format 2] img
- 查看
  - Rbd info
- 变更
  - Resize
  - Rename
  - Copy
  - …..
- 挂载/卸载
  - Map /unmap

玩转块存储

- 导入/导出
  - Import/export
- 增量导入/导出
  - Import-diff/export-diff

玩转块存储

- 快照
  - Rbd snap create –image name --snap name
- 克隆
  - 1.快照
  - 2.保护
  - 3.克隆
- 填平
  - flatten