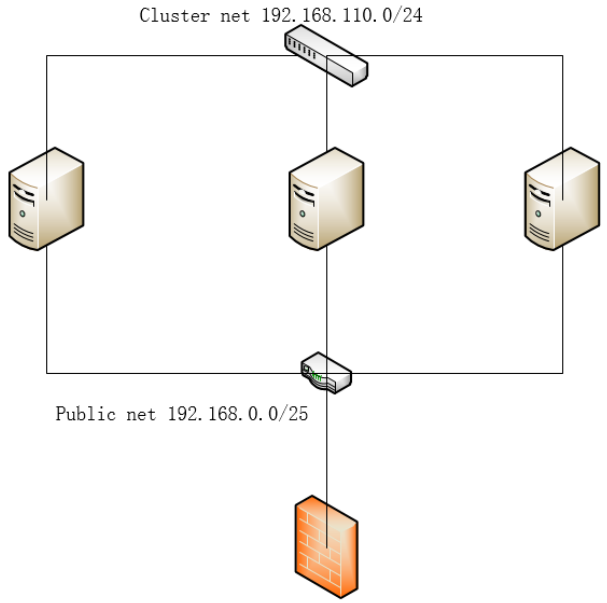
CEPH集群搭建与测试

目标测试：对象存储

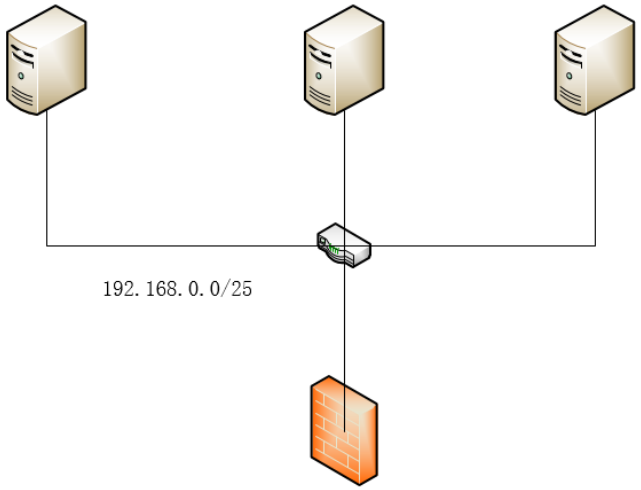
扩展实现：proxmox + openWrt + ceph , cloudstack + openWrt + ceph , k8s + ceph, 云存储

集群搭建拓扑

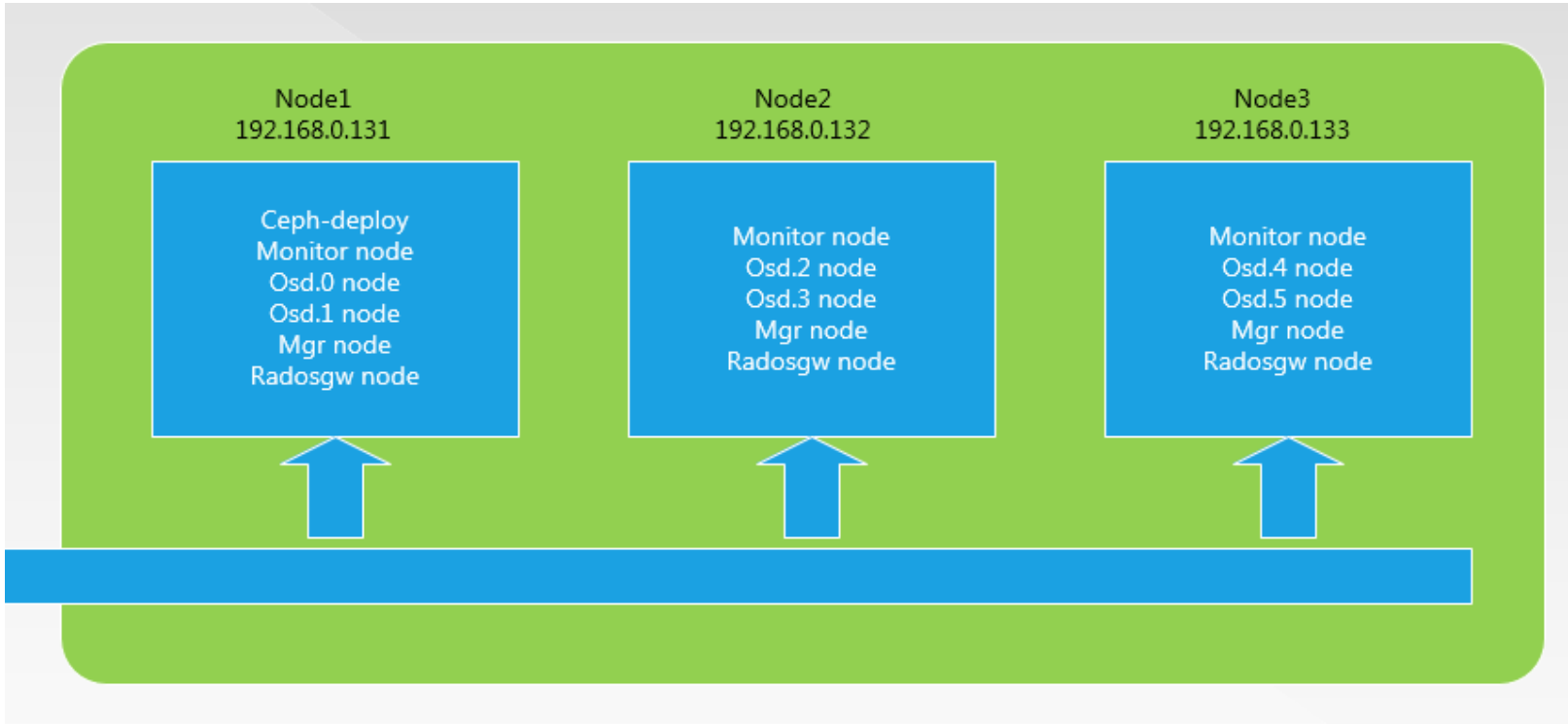
- 理论拓扑



- 实际拓扑



集群结构


Node1
192.168.0.131
Ceph-deploy
Monitor node
Osd.0 node
Osd.1 node
Mgr node
Radosgw node
Node2
192.168.0.132
Monitor node
Osd.2 node
Osd.3 node
Mgr node
Radosgw node
Node3
192.168.0.133
Monitor node
Osd.4 node
Osd.5 node
Mgr node
Radosgw node

集群结构介绍

物理结构

- 系统盘一个100G
- 数据盘两个200G
- 8G
- 1核双核心

结点结构

- OSD结点两个

  存储空间管理

- MON一个

  身份认证，存储结构

- MGR一个

  系统状态监控-dashboard

- RADOSGW一个

  S3接口，身份认证

存储结构

- 物理上支持跨地区，跨机房，跨机架搭建集群
- 对象由存储池，PG管理。默认情况下，认为存储池是个大文件系统，PG是文件夹。但是这个结构对外不公开，由CEPH内部维护
- 数据三备份，PG记录备份结点

集群使用

- Mgr集成dashboard功能，对外提供ceph集群的状态监控，可以了解集群的基本状态。但是mgr多实例后，选举功能不好用。
- 开源的nextcloud，owncloud等天然可以对接ceph，可以直接拿来做为云存储，简单做下优化即可使用。

对象网关接口

- 提供标准Amazson S3接口，支持创建Buckets，分区域管理对象数据。
- 基本的数据操作包括，文件上传，文件下载，文件删除（接口目前已经全部实现）
- 文件下载支持配置访问规则，默认对外不可访问。配置公开后，任何人，只要有访问URL既可以拿到数据
- 文件除了访问规则外，还支持URL有效时间设置，以及URL签名认证

用户认证

- 集群用户认证
- 对象网关认证

对象存储的WEB实现

- 对象网关配置

  用于配置WEB端如何连接访问CEPH

- 用户文件操作

  提供基本的文件操作

CEPH灾害测试

- 集群（三台）中任意一台机器关机

  集群进入degraded状态

  显示结点掉线，两个osd down

  1、机器关机后立刻重启(reboot)

  很快恢复health ok

  2、机器关机一段时间后，打开（poweroff + wait）

  存储空间丢失->由1.2T变成 800G

  OSD数据自动平衡

  直到机器打开后，存储空间恢复

  OSD数据自动平衡，完成后 health OK

- 集群（三台）全部关机

  1、三台机器时间差不大启动

  快速回复正常

  2、三台机器中某一台等待一段时间启动

  存储空间丢失->由1.2T变成800G

  OSD数据自动平衡

  直到机器打开后，存储空间恢复

  OSD数据自动平衡，完成后 health OK

- 复制node3虚拟机对集群的影响

同样结点的出现，会导致mon认证结点出现错误。Mon会下发指令，让出现冲突的osd.4和osd.5关闭多余的。

如果不是原来的osd.4和osd.5则会导致数据自动平衡。

认为关闭复制的虚拟机，重启原来node3上的osd，集群很快恢复正常

- 重复IP的影响

测试环境中没有受到任何影响

- 集群扩容操作

集群新增结点，会导致每个PG管理的空间变大，PG会出现自平衡。直到所有数据达到平衡状态。测试中使用了78000个样本，大约花了3个小时完成自平衡。

集群丢失结点，会导致每个PG管理的空间变下，PG会出现自平衡。直到所有数据达到平衡状态。测试中使用了78000个样本，大约花了3个小时完成自平衡。

自平衡会出现服务降级的现象，但是对于样本数据而言，无所谓了。本身就不是对外提供密集下载，或者密集上传的操作。

测试漏洞

- 无论是poweroff还是powerdown，实质上是提醒计算机可以关机了，而非突然断电了。这一点可以从关机时，cmd上的提示信息可以查看，在关机时，结点自动做了很多后台操作。
- 测试时，并没有同时执行数据写入操作，测试数据的丢失率，读写速度变化等