**Morning：**

1. Ceph概述

**-什么是Ceph？**

1）分布式文件系统：物理存储不一定连接在本地，而通过网络节点相连。

2）可支持PB级别的存储空间（将数据拆分n份，并发写入各网络节点）

3）对设备厂商无任何限制。

4）可提供块存储、文件系统（MDS）、对象存储

Help Doc：http://docs.ceph.org/start/intro

**-Ceph组件**

Mon+OSD：提供块存储

OSDS：存储服务器（越多性能越好，并发写入）

Monitors：集群监控服务器（供客户机访问，3台以上奇数）

MDSs：提供文件系统存储

RGW：提供对象存储

Client：ceph客户端

2、搭建实验环境

**步骤1：配置Yum、主机映射、无密码登陆**

1. 真实机/var/ftp/ceph挂载rhcs2.0-rhosp9-20161113-x86\_64.iso
2. node1搭建yum

baseurl=ftp://192.168.4.254/ceph/rhceph-2.0-rhel-7-x86\_64/MON

baseurl=ftp://192.168.4.254/ceph/rhceph-2.0-rhel-7-x86\_64/OSD

baseurl=ftp://192.168.4.254/ceph/rhceph-2.0-rhel-7-x86\_64/Tools

1. node1添加client、node1～3的主机映射记录
2. 实现node1与其他主机无密码连接
3. 同步yum文件以及主机映射记录hosts到其他主机

**步骤2：配置NTP、清除防火墙规则**

1. 真机作NTP服务端：allow 192.168.4.0/24
2. 各虚拟机NTP配置：server 192.168.4.254 iburst
3. 清除防火墙规则：iptables -F

**步骤3：准备存储磁盘**

node1、node2、node3各添加3块10G磁盘

3、部署ceph集群

**步骤1：安装软件**

1. 在配置无密码登陆的虚拟机安装ceph批量部署脚本

[root@node1 ~]# yum -y install ceph-deploy

1. 创建目录

[root@node1 ~]# mkdir ceph-cluster

[root@node1 ~]# cd ceph-cluster/

**步骤2：部署Ceph集群（只在node1操作）**

1）查看ceph-deploy、ceph-deploy mon指令的--help帮助

2）创建Monitors配置文件（必须为3台以上的奇数）

[root@node1 ceph-cluster]#ceph-deploy new node1 node2 node3

3）给所有节点安装软件包

[root@node1 ceph-cluster]#ceph-deploy install node1 node2 node3

#安装后systemctl status ceph[tab]，可以列出ceph相关服务

4）初始化所有节点的mon服务（主机名解析必须对）

[root@node1 ceph-cluster]# ceph-deploy mon create-initial

#相当于远程拷贝配置文件到/etc/ceph/下，并把mon服务启动

#/etc/ceph/ceph.client.admin.keyring，存放用户名+密码，用于识别并使用集群

常见错误及解决方法（非必要操作，有错误可以参考）：

如果提示如下错误信息：

[node1][ERROR ] admin\_socket: exception getting command descriptions: [Error 2] No such file or directory

解决方案如下（在node1操作）：

先检查自己的命令是否是在ceph-cluster目录下执行的！！！！如果时确认是在该目录下执行的create-initial命令，依然保存，可以使用如下方式修复。

[root@node1 ceph-cluster]# vim ceph.conf

#文件最后追加以下内容

public\_network = 192.168.4.0/24

修改后重新推送配置文件:

[root@node1 ceph-cluster]# ceph-deploy --overwrite-conf config push node1 node2 node3

**步骤3：创建OSD（只在node1操作）**

1. 准备journal缓存分区（相当于固态，可以不做）

[root@room12pc2 ~]#

pssh -i -h host.txt parted /dev/vdb mktable gpt

pssh -i -h host.txt parted /dev/vdb mkpart primary 1M 50%

pssh -i -h host.txt parted /dev/vdb mkpart primary 50% 100%

pssh -i -h host.txt chown ceph:ceph /dev/vdb1

pssh -i -h host.txt chown ceph:ceph /dev/vdb2

#对node1~3进行分区，ceph用户给用于缓存及journal盘

#ceph:ceph权限，重启后会丢失。可以用rc.local、udev规则做开机自启

1. 清空磁盘数据

[root@node1 ceph-cluster]#

ceph-deploy disk zap node1:vdc node1:vdd

ceph-deploy disk zap node2:vdc node2:vdd

ceph-deploy disk zap node3:vdc node3:vdd

#相当于远程执行命令/usr/sbin/partx -a /dev/vdc

1. 创建OSD存储空间（可以随时扩展）

[root@node1 ceph-cluster]#

ceph-deploy osd create node1:vdc:/dev/vdb1 node1:vdd:/dev/vdb2

ceph-deploy osd create node2:vdc:/dev/vdb1 node2:vdd:/dev/vdb2

ceph-deploy osd create node3:vdc:/dev/vdb1 node3:vdd:/dev/vdb2

#node1:vdc:/dev/vdb1=存储主机名:存储空间:缓存

#完成这一步后，磁盘自动分区、格式化、挂载

4）常见错误（非必须操作）

如提示run 'gatherkeys'，可以使用如下命令修复：

[root@node1 ~]# ceph-deploy gatherkeys node1 node2 node3

**步骤4：验证测试**

1）查看集群状态

[root@node1 ~]# ceph -s

[root@node1 ~]# ceph osd tree #拓展

2）常见错误（非必须操作）

health: HEALTH\_WARN

clock skew detected on node2, node3…

#表示时间不同步，解决后等3-5分钟health\_ok

如果还是失败，尝试以下命令，重启ceph服务：

[root@node1 ceph-cluster]# systemctl restart ceph\\*.service ceph\\*.target

#先开启mon的service、target服务，再开启osd的服务。

**Afternoon：**

1. Ceph块存储

**-块存储概述**

-块存储：RADOS block device（RBD）

-RBD驱动已经集成在Linux内核

-客户端可以随意创建



**步骤1：创建镜像**

1. 查看存储池

[root@node1 ~]# ceph osd lspools

0 rbd,

1. 创建镜像、查看镜像

[root@node1 ~]#

rbd create demo-image --image-feature layering --size 10G

rbd create rbd/image --image-feature layering --size 10G

#rbd/image，在共享池rbd创建块设备image，不写默认为rbd

#--image-feature layering，支持快照

[root@node1 ~]# rbd list #查看共享池

[root@node1 ~]# rbd info demo-image #查看块设备属性

**步骤2：动态调整镜像大小**

1. 缩小容量

[root@node1 ~]# rbd resize --size 7G image --allow-shrink

1. 扩容容量

[root@node1 ~]# rbd resize --size 15G image

[root@node1 ~]# rbd info image

**步骤3：通过KRBD访问**

1）集群主机访问

[root@node1 ~]# rbd map image #将image映射为本地磁盘

[root@node1 ~]# lsblk

rbd0 251:0 0 10G 0 disk

2）客户机访问

[root@client ~]# yum -y install ceph-common

scp 192.168.4.11:/etc/ceph/ceph.conf /etc/ceph/

scp 192.168.4.11:/etc/ceph/ceph.client.admin.keyring /etc/ceph/

#拷贝配置文件（否则不知道集群在哪）

#拷贝连接密钥（否则无连接权限）

[root@client ~]# rbd map demo-image #创建映射

[root@client ~]# rbd showmapped #查看映射块设备

3）格式化、挂载

#写入数据时，分散写入到OSDs下，同时复制成多份。

**步骤4：创建镜像快照**

1. 快照原理（COW写时复制）

创建快照时，以快照链接的方式记录；在文件发生修改前，将该文件的原数据拷贝到快照。（快照的大小跟修改内容的大小有关）

2）创建镜像快照

[root@node1 ~]# rbd snap create image --snap image-snap1

[root@node1 ~]# rbd snap ls image #查看快照

3）还原快照

[root@client ~]# umount /mnt   
 #先解除挂载，否则会报“超级坏块”错误

[root@node1 ~]# rbd snap rollback image --snap image-snap1

[root@client ~]# mount /dev/rbd0 /mnt/

[root@client ~]# ls /mnt

#客户端重新挂载分区

**步骤5：创建快照克隆**

1）克隆快照

[root@node1 ~]# rbd snap protect image --snap image-snap1

#保护快照，无法删除

[root@node1 ~]# rbd clone image --snap image-snap1 image-clone --image-feature layering

#将快照image-snap克隆一个新的imgae-clone镜像

1. 查看克隆镜像与父镜像快照的关系

[root@node1 ~]# rbd info image-clone

#含有父镜像快照信息，数据仍来自于快照链

[root@node1 ~]# rbd flatten image-clone

#与父镜像脱离关系（把数据复制一遍）

**步骤6：撤销、删除操作**

1）撤销客户端磁盘映射

[root@client ~]# umount /mnt

[root@client ~]# rbd unmap /dev/rbd/rbd/image

2）删除快照

[root@node1 ~]# rbd snap rm image --snap image-snap

3）删除RBD镜像

[root@node1 ~]# rbd rm image

如何解除快照保护？

编写创建OSD的快速配置脚本。

什么是PG？