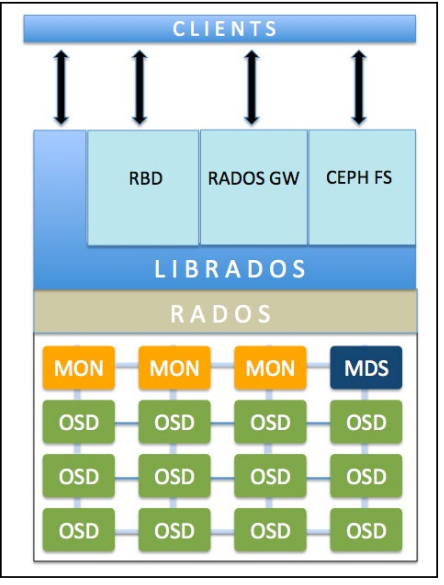
存储：

RAID：独立磁盘冗余阵列

性能、容错、空间

分布式存储：CEPH

1. 组成



1. **MON：监视器。MON通过保存一系列集群状态map来监视集群的组件。MON因为保存集群状态，要防止单点故障，所以需要多台；另外，MON需要是奇数，如果出现意见分岐，采用投票机制，少数服从多数。**
2. **OSD：对象存储设备。真正存储数据的组件。一般来说，每块参与存储的磁盘都需要一个OSD进程。**

**（3）MDS：元数据服务器。只有CephFS需要它。**

元数据：metadata，存储数据的数据。比如一本书内容是数据，那么书的作者、出版社、出版时间之类的信息就是元数据。

1. **RADOS：可靠自主分布式对象存储。它是ceph存储的基础，保证一切都以对象形式存储。**
2. RBD：RADOS块设备，提供块存储
3. CephFS：提供文件系统级别存储
4. RGW：RADOS网关，提供对象存储

存储分类：

块存储：提供硬盘，如iSCSI

文件级别存储：共享文件夹

对象存储：一切皆对象

http://storage.ctocio.com.cn/281/12110781\_2.shtml

**CEPH搭建**

CEPH环境准备

1. 准备6台虚拟机

主机名、IP地址

1. 在物理主机上配置名称解析

[root@room8pc16 nsd2018]# for i in {1..6}

> do

> echo -e "192.168.4.$i\tnode$i.tedu.cn\tnode$i" >> /etc/hosts

> done

3、提前将服务器的密钥保存，不需要ssh时回答yes

[root@room8pc16 nsd2018]# ssh-keyscan node{1..6} >> /root/.ssh/known\_hosts

4、实现免密登陆

[root@room8pc16 nsd2018]# for i in {1..6}

> do

> ssh-copy-id node$i

> done

5、配置yum源

[root@room8pc16 nsd2018]# mkdir /var/ftp/ceph/

[root@room8pc16 nsd2018]# vim /etc/fstab

/ISO/rhcs2.0-rhosp9-20161113-x86\_64.iso /var/ftp/ceph iso9660 defaults 0 0

[root@room8pc16 nsd2018]# mount -a

[root@room8pc16 nsd2018]# vim /tmp/server.repo

[rhel7.4]

name=rhel7.4

baseurl=ftp://192.168.4.254/rhel7.4

enabled=1

gpgcheck=0

[mon]

name=mon

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

enabled=1

gpgcheck=0

[osd]

name=osd

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

enabled=1

gpgcheck=0

[tools]

name=tools

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

enabled=1

gpgcheck=0

[root@room8pc16 nsd2018]# for vm in node{1..6}

> do

> scp /tmp/server.repo ${vm}:/etc/yum.repos.d/

> done

1. **配置node1节点为管理节点**
2. 配置名称解析

[root@node1 ~]# for i in {1..6}; do echo -e "192.168.4.$i\tnode$i.tedu.cn\tnode$i" >> /etc/hosts; done

1. **配置免密登陆**

[root@node1 ~]# ssh-keyscan node{1..6} >> /root/.ssh/known\_hosts

[root@node1 ~]# ssh-keygen -f /root/.ssh/id\_rsa -N ''

[root@node1 ~]# for i in {1..6}; do ssh-copy-id node$i; done

[root@node1 ~]# for vm in node{1..6}

> do

> scp /etc/hosts ${vm}:/etc/

> done

7**、NTP网络时间协议，基于UDP123端口。用于时间同步**

时区：地球一圈360度，经度每15度角一个时区，共24个时区。以英国格林威治这个城市所在纵切面为基准。北京在东八区。

夏季节约时间：夏令时。DST

Stratum：时间服务器的层级。

时间准确度：原子钟。

**8、配置node6为时间服务器**

（1）配置

[root@node6 ~]# yum install -y chrony

[root@node6 ~]# vim /etc/chrony.conf

server 0.centos.pool.ntp.org iburst

#server 1.centos.pool.ntp.org iburst

#server 2.centos.pool.ntp.org iburst

#server 3.centos.pool.ntp.org iburst

allow 192.168.4.0/24

local stratum 10

1. 启动服务

[root@node6 ~]# systemctl enable chronyd

[root@node6 ~]# systemctl restart chronyd

**将node1-5配置为NTP的客户端**

（1）配置

[root@node1 ~]# vim /etc/chrony.conf

#server 0.rhel.pool.ntp.org iburst

#server 1.rhel.pool.ntp.org iburst

#server 2.rhel.pool.ntp.org iburst

#server 3.rhel.pool.ntp.org iburst

server 192.168.4.6 iburst

[root@node1 ~]# systemctl restart chronyd

1. 测试

[root@node1 ~]# date -s "2018-7-13 12:00:00"

[root@node1 ~]# ntpdate 192.168.4.6

[root@node1 ~]# date

（3）同步其他主机

[root@node1 ~]# for i in {2..5}

> do

> scp /etc/chrony.conf node$i:/etc/

> done

[root@node1 ~]# for vm in node{2..5}

> do

> ssh $vm systemctl restart chronyd

> done

**9、为node1-3各添加3块10GB的磁盘**

可以在虚拟机不关机的情况下，直接添加硬盘

**安装ceph**

1. 在node1上安装部署软件

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

1. 创建ceph部署工具的工作目录

[root@node1 ~]# mkdir ceph-clu

1. 创建参与集群节点的配置文件(在目录里操作)

[root@node1 ceph-clu]# ceph-deploy new node{1..3}

[root@node1 ceph-clu]# ls

4、在3个节点上安装软件包

[root@node1 ceph-clu]# ceph-deploy install node{1..3}

1. 初始化mon服务

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

(创建各个节点的node)

**如果出现以下错误：**

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

**解决方案：**

[root@node1 ceph-clu]# vim ceph.conf 最下面加入行：

public\_network = 192.168.4.0/24

再执行以下命令：

[root@host1 ceoh-clu]# ceph-deploy --overwrite-conf config push node1 node2 node3

**偏方：**在 /var/lib/ceph下有一些文件不正确，将其删除重新跑就可以了

1. **把node1-3的vdb作为日志盘。Ext／xfs都是日志文件系统，一个分区分成日志区和数据区。为了更好的性能，vdb专门作为vdc和vdd的日志盘。**

[root@node1 ceph-clu]# for vm in node{1..3}

> do

> ssh $vm parted /dev/vdb mklabel gpt

> done

[root@node1 ceph-clu]# for vm in node{1..3}; do ssh $vm parted /dev/vdb mkpart primary 1M 50% ; done

[root@node1 ceph-clu]# for vm in node{1..3}; do ssh $vm parted /dev/vdb mkpart primary 50% 100% ; done

[root@node1 ceph-clu]# for vm in node{1..3}; do ssh ${vm} chown ceph.ceph /dev/vdb? ; done

1. **创建OSD设备**

[root@node1 ceph-clu]# for i in {1..3}

> do

> ceph-deploy disk zap node$i:vdc node$i:vdd

> done

[root@node1 ceph-clu]# for i in {1..3}

> do

> ceph-deploy osd create node$i:vdc:/dev/vdb1 node$i:vdd:/dev/vdb2

> done

1. **验证**

到第7步为止，ceph已经搭建完成。查看ceph状态

[root@node1 ceph-clu]# ceph -s 如果出现health HEALTH\_OK表示正常

问题排查：

重启时：

ERR：有时因为I/O接口争用到时加载较慢，等一回再看

若依旧不行，重启：

需要加装UDEV到VDB1/2上，开机自动识别。否则会在重启后ERR

WARN：time stew----》时间同步问题，需要每台机器ntpdate 192.168...

根据报错可能或mon节点没有加载完成

查看mon那个down掉了，去查看node 上的systemctl status mon@node1.service

Restart 一下

Pgs degraded----磁盘被降级（一个文件存到3个盘上，若一个损坏，则需要同步，此时显示degraded）

需要找到出现故障的磁盘，否则等待数据同步完成

此次不影响操作

1. 排错

https://www.zybuluo.com/dyj2017/note/920621

CEPH应用

1. 块存储：使用最多的一种方式
2. cephFS：了解，不建议在生产环境中使用，因为还不成熟
3. 对象存储：了解，使用亚马逊的s3

**使用RBD(Rados块设备)**

1. 查看存储池

[root@node1 ~]# ceph osd lspools

可以查看到0号镜像池，名字为rbd

1. 创建名为demo-img的镜像大小为10GB

[root@node1 ~]# rbd create demo-img --image-feature layering --size 10G

[root@node1 ~]# rbd list

[root@node1 ~]# rbd info demo-img

3、创建第2个镜像，名为image，指定它位于rbd池中

[root@node1 ~]# rbd create rbd/image（指定在rbd中创建一个image镜像） --image-feature layering --size 10G

**编写UDEV规则，使得vdb1和vdb2重启后，属主属组仍然是ceph**

[root@node1 ~]# vim /etc/udev/rules.d/90-cephdisk.rules

ACTION=="add", KERNEL=="vdb[12]", OWNER="ceph", GROUP="ceph"