Ceph

存储：

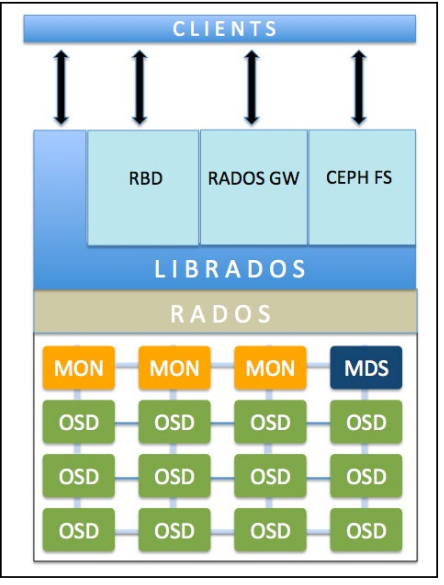
RAID：独立磁盘冗余阵列

性能、容错、空间

raid5性能不如raid1，要写三块硬盘，两块写数据，一块写校验，容错不如raid1，是raid0和1的折中

分布式存储：ceph

分布式文件系统：文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算机网络与节点相连



ceph组件

1、MON 监视器

通过保存一系列的集群状态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，如果出现意见分歧，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

1. MDS 元数据服务器，只有CepthFS需要它

元数据：metadata,存储数据的数据信息。比如一本书的内容是数据，那书的作者、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

它是ceph存储的基本，保证一切都以对象形式存储

1. RBD：RADOS块设备，提供块存储
2. CephFS 提供文件系统级别存储
3. RGW：RADOS网关，提供对象存储

存储分类：

1）块存储 ：提供硬盘，如ISCSI

2）文件级别存储 ： 共享文件夹 如NFS

3）对象存储 ： 一切皆对象

Object { oid data metadata attributes }

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

Ceph环境准备

1. 准备6台虚拟机

192.168.4.1-6

1. 在物理主机上配置域名解析

for i in {1..6}

do

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

done

1. 提前将服务器的密钥保存，不需要ssh再回答yes

ssh-keygen

ssh-keyscan node{1..6} > /root/.ssh/known\_hosts

4、 copy key到1..6上

for i in {1..6}

do

ssh-copy-id node$i

done

1. 配置三个ceph的yum源

vim /etc/fstab

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

vim /etc/yum.repos.d/dvd.repo

[dvd]

name=dvd

baseurl=http://192.168.4.254/rhel7

enable=1

gpgcheck=0

[mon]

name=mon

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

enable=1

gpgcheck=0

[osd]

name=dvd

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

enable=1

gpgcheck=0

[tools]

name=dvd

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

enable=1

gpgcheck=0

1. 配置node1为管理节点
2. 配置名称解析,主机名和域名要一致

for i in {1..6}

do

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

done

ssh-keygen -f /root/.ssh/id\_rsa -N '' //免交互生成密钥

ssh-keyscan node{1..6} > /root/.ssh/known\_hosts

for i in {1..6}

do

ssh-copy-id node$i

done

for i in {1..6}

do

scp /etc/hosts node$i:/etc/

done

1. 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

一、配置node6为时间服务器

1）配置

]# yum -y install chrony

]# 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

allow 192.168.4.0/24

1. 启动服务

]# systemctl restart chronyd

]# systemctl enable chronyd

1. 将node1-5配置为node6的客户端

]# 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

]# systemctl restart chronyd

]# systemctl enable chronyd

1. 测试时间

]# date -s '2008-08-08 12:00:00'

for i in {2..5}

do

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

done

for i in {2..5}

do

ssh node$i systemctl restart chronyd

done

二、为node1-3各加三块硬盘

alter + D

alter + G

alter + F

三、安装ceph

1、使用node1上安装部署软件

]# yum -y install ceph-deploy

1. 创建ceph部署工具的工作目录

]# mkdir ceph-clu

3、创建参与集群节点为的配置文件

]# ceph-deploy new node{1..3}

]# ls

1. 在三个节点上安装软件包

]# ceph-deploy install node{1..3}

1. 初始化mon服务

]# ceph-deploy mon create-initial

如果出现以下错误：No such file or directory

解决：

vim ceph.conf 最下面加入行：

public\_network=192.168.4.0/24

再执行以下命令

ceph-deploy --overwrite-conf config push node1 node2 node3

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

for i in {1..3}

do

ssh node$i parted /dev/vdb mktable gpt

done

for i in {1..3}

do

ssh node$i parted /dev/vdb mkpart primary 1M 50%

done

for i in {1..3}

do

ssh node$i parted /dev/vdb mkpart primary 50% 100%

done

for i in {1..3}

do

ssh node$i chown ceph:ceph /dev/vdb?

done

1. 创建OSD设备

for i in {1..3}

do

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

done

//vdb分成两个分区vdb1,vdb2,vdb1对应vdc的日志区，vdb2对应vdd的日志区

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

出果出现 health HEALTH\_OK代表正常

ceph -s （稍等点时间）

如果出现run gatherkeys错误

ceph-deploy gatherkeys node1 node2 node3

如果出现ERR

在所有节点，或仅在失败的节点重启服务

systemctl restart ceph\\*.service ceph\\*.target

9、排错

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

^1^2 把上一条命令的1换成2再执行

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，应用也不太多

使用RBD（Rados块设备）

1. 查看存储池 node1

#] ceph osd lspools

0 rbd 0号镜像池

1. 创建名为demo-img的镜像，大小为10GB

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

//rbd镜像名字

//rbd命令

#] rbd list

#] rbd info demo-image

1. 创建第2号镜像，名为image，指定它位于rbd池中

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

1. 将image镜像大小缩减为7G(不建议缩小)

#] rbd resize --size 7G image --allow-shrink

#] rbd info image

5、将image镜像扩容到15G

#] rbd resize --size 15G image

#] rbd info image

1. 将node6作为客户端，使用ceph创建的镜像作为存储设备
2. 安装客户端软件

#] yum -y install ceph-common

1. 拷贝相关文件

1]# scp /etc/ceph/ceph.conf node6:/etc/ceph

1]# scp /etc/ceph/ceph.client.admin.keyring node6:/etc/ceph

注：ceph.conf是配置文件，里面记录了ceph集群访问的方式和地址（否则不知道群集在哪）

ceph.client.admin.keyring是client.admin用户的密钥文件（否则无连接权限）

1. 映射image镜像到本地磁盘

6]# rbd map image

6]# lsblk

6]# rbd showmapped

4）本地应用，格式化，挂载

6]# mkfs.ext4 /dev/rbd0

6]# mount /dev/rbd0 /mnt

6]# echo test > /mnt/test.txt

快照

1. 查看image镜像的快照

6]# rbd snap ls image

1. 为image创建名为image-sn1的快照

6]# rbd snap create image --snap image-sn1

1. 模拟误删除操作，恢复数据
2. 删除

6]# rm -rf /mnt/test.txt

1. 卸载设备 （先卸载、再还原、再挂载）

6]# umount /mnt

1. 使用image-sn1还原快照

6]# rbd snap rollback image --snap image-sn1

镜像名 快照名

4）挂载

6]# mount /dev/rbd0 /mnt

6]# cat /mnt/test.txt

克隆快照

1. 克隆快照，首先要把快照保护起来，防止误删除之类的操作、

被保护的快照无法删除，要取消保护（unprotect）

6]# rbd snap protect image --snap image-sn1

1. 克隆image-sn1快照

6]# rbd clone image --snap image-sn1 image-cl1 --image-feature layering

3、查看克隆状态

6]# rbd info image-cl1

parent: rbd/image@image-sn1 //基于image-sn1快照的克隆

4、合并克隆文件

6]# rbd flatten image-cl1

6]# rbd info image-cl1 //没有parent了

1. 客户端撤销磁盘映射

6]# umount /mnt

6]# rbd showmapped

6]# rbd unmap /dev/rbd/rbd/image

1. 删除快照与镜像

删除快照（确保快照未被保护）

6]# rbd snap rm image --snap image-sn1

删除镜像

6]# rbd list

6]# rbd rm image

Ceph实战： (块存储)

安装KVM虚拟机，使用ceph存储提供的镜像作为硬盘

1. 创建名为vm1-image的镜像，大小10GB

#] rbd create vm1-image --image-feature layering --size 10G

#] rbd list

#] rbd info vm1-image //两种查看方式

#] qemu-img info rbd:rbd/vm1-image //两种查看方式

1. 将物理主机作为客户端，安装软件包，拷贝配置文件

#] yum -y install ceph-common

#] scp /etc/ceph/ceph.c\* 192.168.4.254:/etc/ceph

1. 正常创建一台KVM虚拟机，取名为myrhel7，向导结束之后，将其强制关机即可
2. 导出myrhel7虚拟机的声明文件，将虚拟机删掉

#] virsh dumpxml myrhel7 > /tmp/myrhel7.xml

1. 虚拟机使用ceph存储，需要认证，方式是虚拟机先生成secret,再将secret与ceph账户映射
2. 编写账号信息文件

#] vim /tmp/secret.xml

<secret ephemeral= 'no' private= 'no'>

<usage type='ceph'>

<name>clinet.admin secret</name>

</usage>

</secret>

1. 生成secret

#] virsh secret-define --file /tmp/secret.xml //生成完,secret.xml就没用了

#] virsh secret-list //查看secret

1. 将虚拟机软件的secret和ceph的管理员用户关联

1）查看管理员的密钥

cat /etc/ceph/ceph.client.admin.keyring

1. 关联secret和ceph的管理员

virsh secret-set-value

--secret 332d8894-1623-4e6e-a8e0-6e63c7d72563

--base64 AQAyS0hbWCJ9GhAAKctEPI4IxaKy0gLoitxCdg==

7、修改虚拟机的配置文件/tmp/myrhel7.xml，把管理员信息写到该文件中，并指定虚拟机磁盘使用ceph的镜像

vim /tmp/myrhel7.xml 查找/disk

<disk type='network' device='disk'>

<driver name='qemu' type='raw'/>

<auth username='admin'>

<secret type='ceph' uuid='332d8894-1623-4e6e-a8e0-6e63c7d72563'/>

</auth>

<source protocol='rbd' name='rbd/vm1-image'>

<host name='192.168.4.1' port='6789'/>

</source>

<target dev='vda' bus='virtio'/>

<address type='pci' domain='0x0000' bus='0x00' slot='0x07' function='0x0'/>

</disk>

1. 利用xml生成虚拟机

]# virsh define /tmp/myrhel7.xml

9、启动虚拟机时，不能直接安装，需要在虚拟机设置中连接光盘文件，并且设置启动选项，将光盘设置为第一启动介质

CephFS

它可以像NFS或SAMBA那样，提供共享文件夹，客户端通过挂载目录的方式使用ceph的存储

1. cephFS需要一台MDS元数据服务器node4
2. 在node4上安装软件包

]# yum -y install ceph-mds

1. 在node1上配置node4为mds服务器

1]# cd ceph-clu/

1]# ceph-deploy mds create node4

1. 将管理密钥同步到mds服务器

1]# ceph-deploy admin node4

1. 创建两个池，一个用于存储数据，名为：cephfs\_data ，一个用于存储元数据，名为：cephfs\_metadata

1]# ceph osd pool create cephfs\_data 128

1]# ceph osd pool create cephfs\_metadata 128

128表示PG的数目是128，PG是规置组，文件存到PG中，PG存在池中

6、创建名为myfs1的文件系统

1]# ceph fs new myfs1 cephfs\_metadata cephfs\_data

cephfs\_metadata cephfs\_data顺序不要写反，否则数据存储位置会错误

7、查看状态

1]# ceph fs ls

1]# ceph mds stat //查看mds状态

8、CephFS 客户端挂载

Linux内核支持ceph文件系统，不需要装软件

6]# cat /etc/ceph/ceph.client.admin.keyring

6]# mkdir /mnt/cephfs

6]# mount -t ceph 192.168.4.1:6789:/ /mnt/cephfs/ -o name=admin,secret=AQAyS0hbWCJ9GhAAKctEPI4IxaKy0gLoitxCdg==

6]# df -h /mnt/cephfs

Ceph对象存储

1. 使用ceph对象存储，需要RGW，即ceph网关，配置node5是RGW
2. 可以在管理节点上为node5进行安装和配置

1]# ceph-deploy install --rgw node5

3、将配置文件、密钥文件同步到node5

1]# ceph-deploy admin node5

1. 在node5上启动rgw服务

1]# ceph-deploy rgw create node5

1. rgw内建了一个名为civetweb的web服务器,既然是web服务，所以使用80端口更方便，为了应用简便，可以将默认的7480端口改为80端口

5]# vim /etc/ceph/ceph.conf 追加以下内容

[client.rgw.node5]

host = node5

rgw\_frontends = "civetweb port=80"

5]# systemctl restart ceph-radosgw\\*

6、验证端口是否已经修改成功

5]# curl 192.168.4.5

7、在node6上安装客户端软件

6]# rpm -ivh s3cmd-2.0.1-1.el7.noarch.rpm

1. 创建ceph对象存储用户

5]# radosgw-admin user create --uid='xyz' --display-name='xyz' //xyz自定义

注意access\_key和secret\_key

"keys": [

{

"user": "xyz",

"access\_key": "080CZA0390014B907Y27",

"secret\_key": "MV9O5PuMGNp8ALOywDjXXbkz9NuRBvBadUFLnD7Z"

}

],

1. 配置s3客户端(亚马逊)

6]# s3cmd --configure

Access Key: 080CZA0390014B907Y27

Secret Key: MV9O5PuMGNp8ALOywDjXXbkz9NuRBvBadUFLnD7Z

Default Region [US]: //直接回车

S3 Endpoint [s3.amazonaws.com]: 192.168.4.5

[%(bucket)s.s3.amazonaws.com]: %(bucket)s.192.168.4.5

Use HTTPS protocol [Yes]: No

Test access with supplied credentials? [Y/n]:y

Save settings? [y/N] :y

1. 客户端测试

6]# s3cmd ls //查看内容

6]# s3cmd mb s3://my\_dir //创建一个bucket,相当于是文件夹，名称要求x\_y格式

上传文件

6]# s3cmd put /etc/hosts s3://my\_dir

查看

6]# s3cmd ls s3://my\_dir/

下载文件

6]# s3cmd get s3://my\_dir/etc/hosts /tmp/haha

删除文件

6]# s3cmd del s3://my\_dir/etc/hosts

编写UDEV规则，使得vdb1和vdb2重启后，属主属组仍然是ceph

vim /etc/udev/rules.d/90-cephdisk.rules

ACTION=="add",

KERNEL=="vdb[12]",

OWNER="ceph",

GROUP="ceph"