1. **说明**

此文档默认防火墙已关闭，数据盘已挂载。

1. **配置免密**

在集群每台机器执行

|  |
| --- |
| ssh keygen  ssh-copy-id xx（xx集群其他节点ip） |

1. **挂载镜像**

vim /etc/yum.repos.d/CentOS-ISO.repo

|  |
| --- |
| [cdrom]  name=cdrom  baseurl=file:///mnt/cdrom  enabled=1  gpgcheck=0 |

mount -o loop -t iso9660 /root/CentOS-7-x86\_64-Everything-1804.iso /mnt/cdrom

1. **安装依赖库**

**用于脚本autogen.sh的库：**

yum install *libtool*

**用于configure配置的库：**

*yum install gcc-c++ snappy-devel libuuid-devel libblkid-devel libudev-devel expat-devel libcurl-deve fuse-devel libedit-devel libatomic\_ops-devel xfsprogs-devel boost boost-devel gperftools-devel*

以下6个库需要rpm安装，rpm包见压缩包：

*leveldb leveldb-devel cryptopp cryptopp-devel fcgi-devel fcgi*

rpm -ivh 包名 （例：rpm -ivh fcgi-2.4.0-25.el7.x86\_64.rpm）

1. **编译安装ceph**

./autogen.sh

./configure --with-radosgw --with-debug

make

make DESTDIR=/home/rhino/cfs install

mkdir -p /etc/ceph/

touch /etc/ceph/ceph.conf

在命令行执行ceph，运行时报错

ImportError: No module named rados

|  |
| --- |
| cp -r /usr/local/lib/python2.7/site-packages/\* /usr/lib64/python2.7/  Echo /usr/local/lib > /etc/ld.so.conf.d/ceph.conf  ldconfig |

1. **规划集群节点**

参考以下表格根据实际情况规划

|  |  |  |
| --- | --- | --- |
| **主机名** | **网络地址** | **节点类型** |
| rhino160 | 10.45.186.160 | MON,OSD0-2 |
| rhino161 | 10.45.186.161 | MDS,OSD3-5 |
| rhino162 | 10.45.186.162 | OSD6-8 |

替换文件/etc/ceph/ceph.conf（标黄部分根据实际情况修改）

|  |
| --- |
| [global]  fsid = 773ebff0-4e92-4e6d-ab79-e84cab895f60  public\_network = 10.45.186.0/24  auth cluster required = none  auth service required = none  auth client required = none  filestore xattr use omap = true  osd pool default size = 2  osd pool default min size = 1  osd pool default pg num = 128  osd pool default pgp num = 128  osd pool default crush rule = 0  osd crush chooseleaf type = 1  osd\_max\_write\_size= 1024  osd\_disk\_threads = 4  osd\_client\_message\_size\_cap = 2147483648  osd\_client\_message\_cap = 5000  filestore\_op\_threads = 8  mds\_log\_max\_expiring = 200  mds\_log\_max\_segments = 1000  mon\_allow\_pool\_delete = true  mon\_osd\_down\_out\_interval = 600  mon\_osd\_min\_down\_reporters = 3  osd\_map\_dedup = true  pid file=/home/rhino/cfs/var/run/ceph/$cluster-$name.pid  log\_file=/home/rhino/cfs/var/log/ceph/$cluster-$name.log  run\_dir=/home/rhino/cfs/var/run/ceph  erasure\_code\_dir=/home/rhino/cfs/lib/ceph/erasure-code  osd\_class\_dir=/home/rhino/cfs/lib/rados-classes  plugin\_dir=/home/rhino/cfs/lib/ceph  mon\_cluster\_log\_file=default=/home/rhino/cfs/var/log/ceph/$cluster.$channel.log  cluster=/home/rhino/cfs/var/log/ceph/$cluster.log  keyring = /home/rhino/cfs/etc/ceph/$cluster.$name.keyring  [mgr]  mgr\_module\_path=/root/cfs/lib/ceph/mgr  mgr\_modules = dashboard  [mon]  mon\_data = /home/rhino/cfs/var/lib/ceph/mon/$name  mon\_debug\_dump\_location = /home/rhino/cfs/var/log/ceph/$cluster-$name.tdump  keyring = /home/rhino/cfs/etc/ceph/keyring.$name  mon clock drift allowed = 200  mon clock drift warn backoff = 30  [mon.a]  host=rhino160  mon addr=10.45.186.160:6789  [mds]  mds\_data = /home/rhino/cfs/var/lib/ceph/mds/$name  keyring = /home/rhino/cfs/etc/ceph/keyring.$name  mds\_cache\_size = 10000000  mds\_sessionmap\_keys\_per\_op = 2048  max mds = 2  [mds.a]  host=rhino161  [osd]  osd mkfs type=xfs  osd mkfs options xfs = -f  osd mount options xfs = rw,noexec,nodev,user\_xattr,defaults,noatime,nodiratime,data=writeback,nobarrier  osd\_journal\_size = 20000  osd\_max\_write\_size= 1024  osd\_client\_message\_size\_cap = 2147483648  osd\_client\_message\_cap = 5000  osd\_op\_threads = 8  osd\_disk\_threads = 4  osd\_map\_cache\_size = 1024  osd\_map\_cache\_bl\_size = 128  filestore\_queue\_max\_ops = 25000  filestore\_queue\_max\_bytes = 1048576000  filestore\_queue\_committing\_max\_ops = 5000  filestore\_queue\_committing\_max\_bytes = 1048576000  filestore\_op\_threads = 32  journal\_max\_write\_bytes = 1048576000  journal\_queue\_max\_bytes = 10485760000  journal\_max\_write\_entries = 10000  journal\_queue\_max\_ops = 50000  ms\_dispatch\_throttle\_bytes = 1048576000  objecter\_infilght\_op\_bytes = 1048576000  objecter\_inflight\_ops =8192  ms\_rwthread\_stack\_bytes=10485760  osd\_hit\_set\_min\_size=5000  osd\_max\_pgls=4096  osd\_pg\_bits=24  osd\_pgp\_bits=24  keyring = /home/rhino/cfs/etc/ceph/keyring.$name  [osd.0]  host = rhino160  osd data = /data01/ceph  osd journal = /data01/journal  journal force aio = true  [osd.1]  host = rhino160  osd data = /data02/ceph  osd journal = /data02/journal  journal force aio = true  [osd.2]  host = rhino160  osd data = /data03/ceph  osd journal = /data03/journal  journal force aio = true  [osd.3]  host = rhino161  osd data = /data01/ceph  osd journal = /data01/journal  journal force aio = true  [osd.4]  host = rhino161  osd data = /data02/ceph  osd journal = /data02/journal  journal force aio = true  [osd.5]  host = rhino161  osd data = /data03/ceph  osd journal = /data03/journal  journal force aio = true  [osd.6]  host = rhino162  osd data = /data01/ceph  osd journal = /data01/journal  journal force aio = true  [osd.7]  host = rhino162  osd data = /data02/ceph  osd journal = /data02/journal  journal force aio = true  [osd.8]  host = rhino162  osd data = /data03/ceph  osd journal = /data03/journal  journal force aio = true  [client]  rbd\_cache = true  rbd\_cache\_size = 134217728  rbd\_cache\_max\_dirty = 100663296  rbd\_cache\_max\_dirty\_age = 5  rbd\_cache\_target\_dirty = 67108864  rbd\_cache\_writethrough\_until\_flush = true |

1. **安装MON**

在MON节点机器执行命令：

mkdir -p /etc/ceph

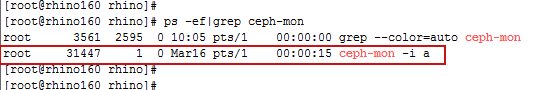
mkdir -p /home/rhino/cfs/var/lib/ceph/mon/mon.a

mkdir -p /home/rhino/cfs/var/run/ceph

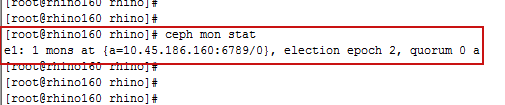
ceph-mon --mkfs -i a

ceph-mon -i a

查看进程：ps -ef|grep ceph-mon



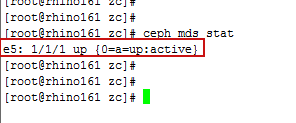
查看状态：ceph mon stat



1. **安装MDS**

ceph-mds -i a

查看状态：ceph mds stat



1. **安装OSD**
   1. **创建osd**

ceph osd create

ceph osd create

ceph osd create

注：每调用一次ceph osd create命令，osdid(osd节点的序号)就相应增加1。

删除osd

ceph osd out 6 （踢出集群）

ceph -w （立即重新平衡）

sudo systemctl stop ceph-osd@6 （停止）

ceph osd tree

ceph osd crush remove osd.6 （删除）

ceph auth del osd.6 （删除认证）

ceph osd rm 6

* 1. **初始化osd**

ceph-osd -i 0 --mkfs --mkkey

ceph-osd -i 1 --mkfs --mkkey

ceph-osd -i 2 --mkfs --mkkey

注：ceph-osd -i <osdid>

在主机rhino160执行3次ceph osd create命令，osdid为0~2

然后在arm161执行3次ceph osd create命令，osdid为3~5

在arm162执行3次ceph osd create命令后，osdid为6~8

* 1. **注册osd凭证**

ceph auth add osd.0 osd 'allow \*' mon 'allow profile osd' -i /etc/ceph/keyring.osd.0

ceph auth add osd.1 osd 'allow \*' mon 'allow profile osd' -i /etc/ceph/keyring.osd.1

ceph auth add osd.2 osd 'allow \*' mon 'allow profile osd' -i /etc/ceph/keyring.osd.2

注：ceph auth add <osd.id> osd 'allow \*' mon 'allow profile osd' -i /etc/ceph/<keyring.osd.id>

如果报错没有路径或者文件则手动创建

* 1. **添加host到CRUSH映射图**

ceph osd crush add-bucket rhino160 host

* 1. **将节点移到default树下**

ceph osd crush move rhino160 root=default

* 1. **添加OSD到host下**

ceph osd crush add osd.0 1.0 host=rhino160

ceph osd crush add osd.1 1.0 host=rhino160

ceph osd crush add osd.2 1.0 host=rhino160

...

...

* 1. **启动osd**

ceph-osd -i 0

ceph-osd -i 1

ceph-osd -i 2

...

...

1. **Ceph文件系统**
   1. **创建**

在MON上执行下面命令来创建存储池

ceph osd pool create cephfs\_data 512

ceph osd pool create cephfs\_metadata 512

创建文件系统

ceph fs new cephfs cephfs\_metadata cephfs\_data

查看文件系统

ceph fs ls

查看所有pool

rados df

查看指定库中所有object

rados -p cephfs\_data ls

* 1. **内核空间挂载（在客户端机器挂载）**

mkdir -p /cfs-kerl

mount -t ceph 10.45.186.160:6789:/ /cfs-kerl

* 1. **用户空间挂载（在客户端机器挂载）**

和上面属于两种不同的挂载方式，此方式需要安装yum install ceph-fuse

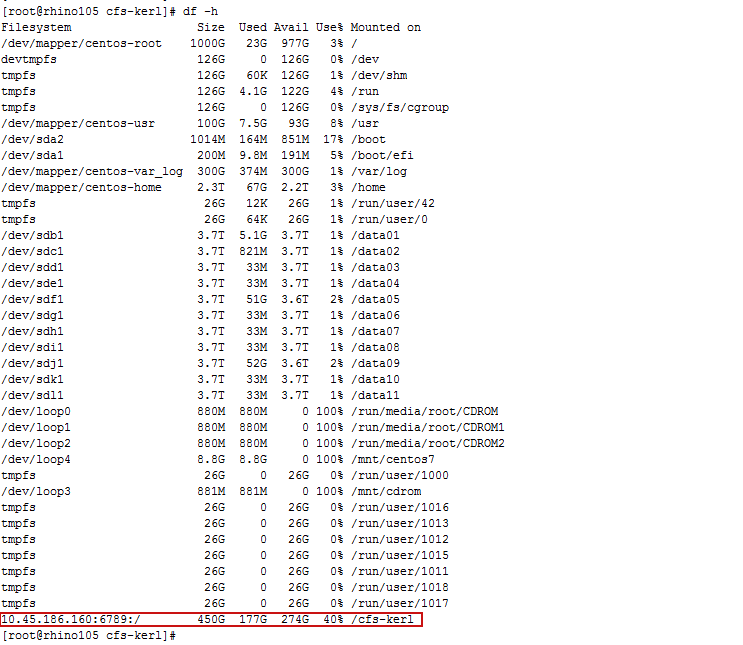
mkdir -p /cfs-user

ceph-fuse -m 10.45.186.160:6789 /cfs-user

rbd create zc\_image --image-feature layering --size 5g

* 1. **检查**

在客户端查看ceph文件系统是否挂载



在集群节点通过ceph health detail命令，如果ceph 状态为HEALTH\_OK，说明ceph集群系统部署成功

[root@hadoop102 iozone]# ceph health detail

HEALTH\_OK

1. **上传文件**

将本地文件上传到10.1创建的pool（ cephfs\_data）中，项目名为object1

rados -p cephfs\_data put object1 ~/zc/zc.log

查看存储池cephfs\_data中的项目

rados -p cephfs\_data ls

查看项目中的pg map信息

ceph osd map cephfs\_data object1



其中：

osdmap e110 map的版本号

pool 'cephfs\_data' (1) pool的名字和ID

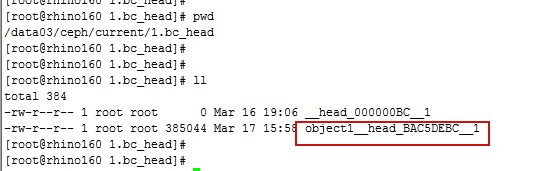
object 'object1' object的名字

pg 1.bac5debc (1.bc) pg number，即1.bc

up ([2], p2) acting ([2], p2) 即OSD.2

在osd找到上传的zc.log文件

cd /data03/ceph/current/1.bc\_head



1. **常用命令**

Ceph状态：ceph -s/status

集群健康状态：ceph -w/health

集群使用状态：ceph df

Mon状态：ceph mon stat

Osd状态：ceph osd stat

Pg（配置组）状态：ceph osd stat

列表PG ： ceph pg dump

列表Ceph存储池: ceph osd lspools

检查OSD的CRUSH map:ceph osd tree

列表集群的认证密钥: ceph auth list