Ceph安装部署文档

目录

[一：简介 1](#_Toc410298597)

[二：部署环境介绍 1](#_Toc410298598)

[三：集群配置准备工作 2](#_Toc410298599)

[四：安装ceph软件包 6](#_Toc410298600)

[五：安装ceph对象网关 9](#_Toc410298601)

[六、搭建ceph集群 10](#_Toc410298602)

[6.1、配置mon节点 10](#_Toc410298603)

[6.2：添加osd节点 13](#_Toc410298604)

[6.2.1、添加第一块osd节点 13](#_Toc410298605)

[6.3：添加元数据服务器 17](#_Toc410298606)

[6.3.1、添加第一个元数据服务器 17](#_Toc410298607)

[七：安装client端RBD、cephFS挂载演示 18](#_Toc410298608)

[7.1：客户端内核要求、安装装准备 18](#_Toc410298609)

[7.2、RBD方式挂载 19](#_Toc410298610)

[7.3、cephFS挂载 20](#_Toc410298611)

[八：检查ceph集群状态常用命令 23](#_Toc410298612)

[8.1、磁盘压力测试软件推荐 23](#_Toc410298613)

[九:故障排查 23](#_Toc410298614)

## 一：简介

Ceph 生态系统架构可以划分为四部分：

1. Clients：客户端（数据用户）

2. mds：Metadata server cluster，元数据服务器（缓存和同步分布式元数据）

3. osd：Object storage cluster，对象存储集群（将数据和元数据作为对象存储，执行其它关键职能）

4. mon：Cluster monitors，集群监视器（执行监视功能）

## 二：部署环境介绍

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 主机名 | 公网IP（eth0） 私网IP（eth1） 双网卡绑定 bond0 | 运行服务 | cpu | 内存 | 操作系统 内核 | 磁盘 | 备注 |
| client | 172.16.4.17 | Client | cpu: e5520\*2 5506/4核/8线程\*2 | 4g\*4 | CentOS6.6 3.18.2 |  |  |
| node01 | 172.16.2.27 | mon、mds、osd | cpu: e5520\*2 5506/4核/8线程\*2 | 4g\*4 | CentOS6.6 | 1T sas 7500转\*4 第一块装系统 |  |
| node02 | 172.16.2.28 | mon、mds、osd | cpu: e5520\*2 5506/4核/8线程\*2 | 4g\*4 | CentOS6.6 | 1T sas 7500转\*4 第一块装系统 |  |
| node03 | 172.16.2.29 | mon、mds、osd | cpu: e5520\*2 5506/4核/8线程\*2 | 4g\*4 | CentOS6.6 | 1T sas 7500转\*4 第一块装系统 |  |

## 三：集群配置准备工作

node01节点执行

3.1：创建工作文件夹，建立ip地址list，为文件传输做准备

mkdir /workspace/

[root@node01 ~]# cd /workspace/

[root@node01 workspace]# vi cephlist.txt 主机列表

client

node01

node02

node03

主机ssh登陆证书

ssh-keygen

cd .ssh/

ssh-copy-id 127.1

将配置完成的证书传输到其它服务器

sync –avp \* 172.17.2.27:/root/.ssh/

sync –avp \* 172.17.2.28:/root/.ssh/

sync –avp \* 172.17.2.29:/root/.ssh/

3.2：系统基础设置，建议整理为脚本，批量传输到所有集群主机执行

3.2.1.修改SELINUX模式

sed -i 's/SELINUX=enforcing/SELINUX=disabled/' /etc/sysconfig/selinux

sed -i 's/SELINUX=enforcing/SELINUX=disabled/' /etc/selinux/config

3.2.2.set max user processes

sed -i 's/1024/102400/' /etc/security/limits.d/90-nproc.conf

echo "\* - sigpending 256612" >> /etc/security/limits.d/90-nproc.conf

3.2.3.set ulimit

echo "ulimit -SHn 102400" >> /etc/rc.local

cat >> /etc/security/limits.conf << EOF

\* soft nofile 102400

\* hard nofile 102400

\* soft nproc 102400

\* - sigpending 256612

\* hard nproc 102400

EOF

#执行profile

source /etc/profile

3.3: 修改host文件

[root@node01 ~]# vi /etc/hosts

追加内容如下：

172.16.4.17 client

172.16.2.27 node01

172.16.2.28 node02

172.16.2.29 node03

172.16.4.240 ceph.ptmind.com

将host文件传输到其它主机

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;rsync -avp /etc/hosts $ip:/etc/;done

3.4：安装centos的yum源软件包 =>全部节点安装

3.4.1:升级基础包

rpm -Uvh http://mirrors.sohu.com/centos/6.7/os/x86\_64/Packages/python-iniparse-0.3.1-2.1.el6.noarch.rpm

rpm -Uvh http://mirrors.sohu.com/centos/6.7/os/x86\_64/Packages/yum-metadata-parser-1.1.2-16.el6.x86\_64.rpm

rpm -Uvh http://mirrors.sohu.com/centos/6.7/os/x86\_64/Packages/yum-3.2.29-69.el6.centos.noarch.rpm

rpm -Uvh http://mirrors.sohu.com/centos/6.7/os/x86\_64/Packages/yum-plugin-fastestmirror-1.1.30-30.el6.noarch.rpm

3.4.2：配置一个国内yum源

vi /etc/yum.repos.d/my.repo

详见批注

3.4.3：添加时间同步定时任务

vi /etc/crontab

3.4.4：时间同步,此处是我们的内部时间源

20 \*/1 \* \* \* root /usr/sbin/ntpdate 172.16.5.100 && /sbin/hwclock –w

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;rsync -avp --delete /etc/crontab $ip:/etc/;done

4:5：传输yum源文件到其它节点服务器

--delete 删除那些DST中SRC没有的文件

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;rsync -avp --delete /etc/yum.repos.d $ip:/etc/;done

3.5、更新yum源 （所有节点执行）

Screen

yum clean all &&yum update –y

3.6、添加ceph yum源或者自制yum源；

在所有节点安装ceph的所有的yum源 （所有节点执行）

3.6.1、安装软件包key

(1)、release.asc key

rpm --import 'https://ceph.com/git/?p=ceph.git;a=blob\_plain;f=keys/release.asc'

（2）、autobuild.asc key

rpm --import 'https://ceph.com/git/?p=ceph.git;a=blob\_plain;f=keys/autobuild.asc'

3.6.2、安装ceph附加包源ceph-extras.repo，设置priority=2，确保新的包（如qemu）优先级比标准包的高。

vi /etc/yum.repos.d/ceph-extras.repo

详见附件

#制作私有yum源下载软件包地址

wget -S -c -r -np -L <http://ceph.com/packages/ceph-extras/rpm/centos6/>

3.6.3、安装ceph源

vi /etc/yum.repos.d/ceph.repo

##制作私有yum源下载软件包地址

wget -S -c -r -np -L <http://ceph.com/rpm/rhel6/>

3.6.4、安装ceph的apache yum源

vi /etc/yum.repos.d/ceph-apache.repo

##制作私有yum源下载软件包地址

wget -S -c -r -np -L <http://gitbuilder.ceph.com/apache2-rpm-centos6-x86_64-basic/ref/master/>

3.6.5、添加ceph-fastcgi.repo file to /etc/yum.repos.d

vi /etc/yum.repos.d/mod\_fastcgi.repo

详见批注

##制作私有yum源下载软件包地址

wget -S -c -r -np -L <http://gitbuilder.ceph.com/mod_fastcgi-rpm-centos6-x86_64-basic/ref/master/>

3.7：如果制作内网yum源，完成上面给出的下载链接后，进行如下配置：

3.7.1：删除无用文件

cd /usr/share/nginx/html

find ./ -name index.html\* >1.txt

for i in $(cat 1.txt);do rm -rf $i;done

find ./ -name index.html\*

3.7.2：调整目录结构

下载至本地的软件包，目录结构根据实际情况作调整，

3.7.3：添加、修改nginx配置：

location / {

root /usr/share/nginx/html;

autoindex on;

3.7.4、修改yum源，repo文件，将yum请求的地址改为私有地址；

例如：

vi /etc/yum.repos.d/ceph-extras.repo

[ceph-extras-source]

name=Ceph Extras Sources

baseurl=http://172.16.2.240/ceph-extras/packages/ceph-extras/rpm/centos6/SRPMS

enabled=1

priority=2

gpgcheck=0

type=rpm-md

gpgkey=https://ceph.com/git/?p=ceph.git;a=blob\_plain;f=keys/release.asc

3.8、安装国内epel yum源

rpm -Uvh http://mirrors.sohu.com/fedora-epel/6/x86\_64/epel-release-6-8.noarch.rpm

备：

rpm -Uvh http://dl.fedoraproject.org/pub/epel/6/x86\_64/epel-release-6-8.noarch.rpm

官方推荐rpm安装ceph，安装下面插件：

3.9：安装 yum-plugin-priorities.

yum install yum-plugin-priorities -y

yum update –y

3.9.1：确认配置文件内容

cat /etc/yum/pluginconf.d/priorities.conf

[main]

enabled = 1

## 四：安装ceph软件包

4.1、安装ceph部署机 ，使用ceph-deploy方式安装部署

yum install -y ceph-deploy

4.2、安装ceph存储集群

（1）、安装ceph必备软件 ==all

yum install -y snappy leveldb gdisk python-argparse gperftools-libs

（2）、安装ceph软件 ==all

yum install -y ceph

4.3、安装ceph对象网关 ==all

（1）、安装apache fastcgi 需要yum install httpd mod\_fastcgi，安装之前先执行下面的操作

yum install -y yum-plugin-priorities

yum update -y

4.4、安装apache fastcgi ==all

yum install -y httpd mod\_fastcgi

4.4.1、确认 FastCGI module启用.

cat /etc/httpd/conf.d/fastcgi.conf|grep mod\_fastcgi.so

输出结果如下：

LoadModule fastcgi\_module modules/mod\_fastcgi.so

4.4.2、确认 Rewrite module 启用

cat /etc/httpd/conf/httpd.conf|grep mod\_rewrite.so

输出结果如下：

LoadModule rewrite\_module modules/mod\_rewrite.so

4.4.3、编辑配置文件 httpd.conf ,修改 httpd.conf ServerName为本机主机名

vim /etc/httpd/conf/httpd.conf

echo "ServerName $HOSTNAME" >> /etc/httpd/conf/httpd.conf

4.4.4、启动httpd进程

/etc/init.d/httpd restart

4.5、安装SSL (安装此步骤有报错）

yum install -y mod\_ssl openssl

4.5.1、创建证书

生成私钥.

openssl genrsa -out ca.key 2048

4.5.2、生成CSR文件

openssl req -new -key ca.key -out ca.csr

|  |  |
| --- | --- |
| ountry Name (2 letter code) | 使用国际标准组织(ISO)国码格式，填写2个字母的国家代号。中国请填写CN。 |
| State or Province Name (full name) | 省份，比如填写Shanghai |
| Locality Name (eg, city) | 城市，比如填写Shanghai |
| Organization Name (eg, company) | 组织单位，比如填写公司名称的拼音 |
| Organizational Unit Name (eg, section) | 比如填写IT Dept |
| Common Name (eg, your websites domain name): | 行使 SSL 加密的网站地址。请注意这里并不是单指您的域名，而是直接使用 SSL 的网站名称 例如:pay.abc.com。 一个网站这里定义是：  abc.com 是一个网站；  www.abc.com 是另外一个网站；  pay.abc.com 又是另外一个网站。 |
| Email Address | 邮件地址，可以不填 |
| A challenge password | 可以不填 |
| An optional company name | 可以不填 |

4.5.3、生成证书

openssl x509 -req -days 365 -in ca.csr -signkey ca.key -out ca.crt

4.5.4、copy生成证书至apache监听位置.

cp ca.crt /etc/pki/tls/certs

cp ca.key /etc/pki/tls/private/ca.key

cp ca.csr /etc/pki/tls/private/ca.csr

4.5.5、修改apache ssl配置文件 /etc/httpd/conf.d/ssl.conf.

vi +105 /etc/httpd/conf.d/ssl.conf

将

105 SSLCertificateFile /etc/pki/tls/certs/localhost.crt

改为：

105 SSLCertificateFile /etc/pki/tls/certs/ca.crt

将

112 SSLCertificateKeyFile /etc/pki/tls/private/localhost.key

改为

112 SSLCertificateKeyFile /etc/pki/tls/private/ca.key

4.5.6：将证书文件传到其它主机

--delete 删除那些DST中SRC没有的文件

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;rsync -avp --delete /etc/pki/tls/certs $ip:/etc/pki/tls/;done

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;rsync -avp --delete /etc/pki/tls/private $ip:/etc/pki/tls/;done

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;rsync -avp --delete /etc/httpd/conf.d/ssl.conf $ip:/etc/httpd/conf.d/ssl.conf;done

4.5.7、批量重启服务

for ip in $(cat /workspace/cephlist.txt);do echo -----$ip-----;ssh root@$ip /etc/init.d/httpd restart;done

官网步骤，写了host文件略过啦

（10）、Add Wildcard to DNS （The address of the DNS must also be specified in the Ceph configuration file with the rgw dns name = {hostname} setting.

address=/.ceph-node/192.168.0.1

## 五：安装ceph对象网关

5.1、基础软件包

yum install -y ceph-radosgw ceph

5.2、安装ceph对象网关代理

yum install -y radosgw-agent

5.3、为块存储安装虚拟化软件

5.3.1、如果之前有qemu模块先删除，确保之后安装的是最完善的

yum remove -y qemu-kvm qemu-kvm-tools qemu-img

5.3.2、卸载后重新安装qemu

yum install -y qemu-kvm qemu-kvm-tools qemu-img

5.4、安装qemu客户代理

yum install -y qemu-guest-agent

5.5、在所有节点安装ceph依赖的软件及模块

yum install \*argparse\* -y

yum install redhat-lsb -y

yum install xfs\* -y

############官方文档无此内容############

(4)、安装libvirt软件包

yum install -y libvirt

以上步骤参考配置连接

http://docs.ceph.com/docs/master/install/install-ceph-gateway/

配置 ceph对象网关

<http://docs.ceph.com/docs/master/radosgw/config/>

## 六、搭建ceph集群

<http://docs.ceph.com/docs/master/radosgw/config/>

### 6.1、配置mon节点

6.1.1、建立第一个mon节点

1、登录监控节点node01节点

ls /etc/ceph #查看ceph配置文件目录是否存在

2、创建ceph配置文件并配置ceph配置文件内的内容

3、创建一个ceph配置文件

touch /etc/ceph/ceph.conf

4、执行此命令可以得到一个唯一的标识，作为ceph集群ID

uuidgen

bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

fsid = bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

#此标识就是上面得到的，把此条命令加入ceph的配置文件

mon initial members = node01,node02,node03

#node01、node02、node03作为ceph集群的监控节点，把此条命令加入到ceph配置文件

mon host = 172.16.2.27, 172.16.2.28, 172.16.2.29

#监控节点的地址，把此条命令加入ceph的配置文件中

5、ceph配置文件预览

vi /etc/ceph/ceph.conf

[global]

fsid = bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

mon initial members = node01,node02,node03

mon host = 172.16.2.27, 172.16.2.28, 172.16.2.29

public network = 172.16.2.0/24

auth cluster required = cephx

auth service required = cephx

auth client required = cephx

osd journal size = 1024

filestore xattr use omap = true

osd pool default size = 3

osd pool default min size = 1

osd crush chooseleaf type = 1

osd\_mkfs\_type = xfs

max mds = 5

mds max file size = 100000000000000

mds cache size = 1000000

mon osd down out interval = 900

#设置osd节点down后900s，把此osd节点逐出ceph集群，把之前映射到此节点的数据映射到其他节点。

cluster\_network = 172.16.2.0/24

[mon]

mon clock drift allowed = .50

#把时钟偏移设置成0.5s，默认是0.05s,由于ceph集群中存在异构PC，导致时钟偏移总是大于0.05s，为了方便同步直接把时钟偏移设置成0.5s

6、在node01创建各种密钥

#为监控节点创建管理密钥

ceph-authtool --create-keyring /tmp/ceph.mon.keyring --gen-key -n mon. --cap mon 'allow \*'

#为ceph amin用户创建管理集群的密钥并赋予访问权限

ceph-authtool --create-keyring /etc/ceph/ceph.client.admin.keyring --gen-key -n client.admin --set-uid=0 --cap mon 'allow \*' --cap osd 'allow \*' --cap mds 'allow'

#添加client.admin key到 ceph.mon.keyring

ceph-authtool /tmp/ceph.mon.keyring --import-keyring /etc/ceph/ceph.client.admin.keyring

验证

cat /tmp/ceph.mon.keyring

7、在node01监控节点创建一个mon数据目录

mkdir -p /var/lib/ceph/mon/ceph-node01

8、在node01创建一个boot引导启动osd的key

mkdir -p /var/lib/ceph/bootstrap-osd/

ceph-authtool -C /var/lib/ceph/bootstrap-osd/ceph.keyring

9、在node01节点上初始化mon节点，

ceph-mon --mkfs -i node01 --keyring /tmp/ceph.mon.keyring

10、为了防止重新被安装创建一个空的done文件

touch /var/lib/ceph/mon/ceph-node01/done

11、创建一个空的初始化文件

touch /var/lib/ceph/mon/ceph-node01/sysvinit

12、启动ceph进程

/sbin/service ceph -c /etc/ceph/ceph.conf start mon.node01

10、查看asok mon状态

[root@node01 ~]# ceph --cluster=ceph --admin-daemon /var/run/ceph/ceph-mon.node01.asok mon\_status

6.1.2、添加其他mon（2）节点：

1、复制node01节点的/etc/ceph目录到node02

2、复制node01节点上的/var/lib/ceph/bootstrap-osd/ceph.keyring文件到node02

3、复制node01节点上的/tmp/ceph.mon.keyring

4、在node02节点上新建一个/var/lib/ceph/bootstrap-osd/目录

5、在node02节点上建立一个/var/lib/ceph/mon/ceph-node02目录

6、在node02节点上初始化mon节点

7、为了防止重新被安装创建一个空的done文件

8、创建一个空的初始化文件

9、启动ceph进程

scp /etc/ceph/\* node02:/etc/ceph/

scp /var/lib/ceph/bootstrap-osd/ceph.keyring node02:/var/lib/ceph/bootstrap-osd/

scp /tmp/ceph.mon.keyring node02:/tmp/

mkdir /var/lib/ceph/bootstrap-osd/

mkdir -p /var/lib/ceph/mon/ceph-node02

ceph-mon --mkfs -i node02 --keyring /tmp/ceph.mon.keyring

touch /var/lib/ceph/mon/ceph-node02/done

touch /var/lib/ceph/mon/ceph-node02/sysvinit

/sbin/service ceph -c /etc/ceph/ceph.conf start mon.node02

验证状态：

ceph --cluster=ceph --admin-daemon /var/run/ceph/ceph-mon.node02.asok mon\_status

6.1.3、添加其他mon（3）节点：

添加方法同 6.1.2 步骤，注意修改node名称

## 6.2：添加osd节点

### 6.2.1、添加第一块osd节点

1、创建一个OSD，生成一个osd number

[root@node01 ~]# ceph osd create

0

2、为osd节点创建一个osd目录

[root@node01 ~]# mkdir -p /var/lib/ceph/osd/ceph-0

3、格式化已准备好的osd硬盘（格式化为xfs格式）

[root@node01 ~]# mkfs.xfs -f /dev/sdb

4、创建osd挂载目录

[root@node01 ~]# mount /dev/sdb /var/lib/ceph/osd/ceph-0

[root@node01 ~]# mount -o remount,user\_xattr /var/lib/ceph/osd/ceph-0

查看挂载的情况

[root@node01 ~]# mount

/dev/sda2 on / type ext4 (rw)

proc on /proc type proc (rw)

/dev/sdb on /var/lib/ceph/osd/ceph-1 type xfs (rw,user\_xattr)

5、把上面的挂载信息写入分区表

[root@node01 ~]# vi /etc/fstab

/dev/sdb /var/lib/ceph/osd/ceph-0 xfs defaults 0 0

/dev/sdb /var/lib/ceph/osd/ceph-0 xfs remount,user\_xattr 0 0

6、初始化osd数据目录

[root@node01 ~]# ceph-osd -i 0 --mkfs --mkkey

6、注册osd的认证密钥

ceph auth add osd.[osd ID] osd 'allow \*' mon 'allow profile osd' -i /var/lib/ceph/osd/ceph-[osd ID]/keyring

[root@node01 ~]# ceph auth add osd.0 osd 'allow \*' mon 'allow profile osd' -i /var/lib/ceph/osd/ceph-0/keyring

查看认证节点

ceph auth list

osd.1

key: AQDwSsZUQMkeGRAAmYYw9CoUUdK73IjDSH/gnw==

caps: [mon] allow profile osd

7、为此osd节点创建一个crush map

[root@node01 ~]# ceph osd crush add-bucket node01 host

8、防放置 Node01 到根路径

[root@node01 ~]# ceph osd crush move node01 root=default

moved item id -2 name 'node01' to location {root=default} in crush map

9、放置osd.0到 bucket node01

[root@node01 ~]# ceph osd crush add osd.0 1.0 host=node01

add item id 0 name 'osd.0' weight 1 at location {host=node01} to crush map

10、创建一个初始化目录

[root@node01 ~]# touch /var/lib/ceph/osd/ceph-0/sysvinit

11、启动osd进程

/etc/init.d/ceph start osd.0

12、查看osd目录树

[root@node01 ~]# ceph osd tree

# id weight type name up/down reweight

-1 1 root default

-2 1 host node01

0 1 osd.0 up 1

6.2.2、添加第二个osd节点

[root@node02 ~]#

1、创建一个OSD，生成一个osd number

2、为osd节点创建一个osd目录

3、格式化已准备好的osd硬盘，并挂在到上一步创建的osd目录（格式化为xfs格式）

4、挂在目录

5、初始化osd数据目录

6、注册osd的认证密钥

7、为此osd节点创建一个crush map

8、Place the Ceph Node under the root default

9、放置osd.0到 bucket node01

10、创建一个初始化目录

11、启动osd进程 、对应参数如下

ceph osd create

mkdir -p /var/lib/ceph/osd/ceph-1

mkfs.xfs -f /dev/sdb

mount /dev/sdb /var/lib/ceph/osd/ceph-1

mount -o remount,user\_xattr /var/lib/ceph/osd/ceph-1

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

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

ceph osd crush add-bucket node02 host

ceph osd crush move node02 root=default

ceph osd crush add osd.1 1.0 host=node02

touch /var/lib/ceph/osd/ceph-1/sysvinit

/etc/init.d/ceph start osd.1

12、把上面的挂载信息写入分区表

vi /etc/fstab

/dev/sdb /var/lib/ceph/osd/ceph-1 xfs defaults 0 0

/dev/sdb /var/lib/ceph/osd/ceph-1 xfs remount,user\_xattr 0 0

验证：

mount

/dev/sdb on /var/lib/ceph/osd/ceph-1 type xfs (rw,user\_xattr)

13、查看osd目录树

[root@node02 ~]# ceph osd tree

# id weight type name up/down reweight

-1 2 root default

-2 1 host node01

0 1 osd.0 up 1

-3 1 host node02

1 1 osd.1 up 1

6.2.3、添加其它osd节点 ，参考6.2.1、6.2.2 步骤

添加node01 服务器上的其它磁盘为osd

添加node02 服务器上的其它磁盘为osd

添加node03 服务器上的其它磁盘为osd

验证：

mount

12、查看osd目录树

[root@node02 ~]# ceph osd tree

添加其它osd方法步骤同上

添加删除osd参考文档

http://docs.openfans.org/ceph/ceph4e2d658765876863/ceph-1/ceph-storage-cluster3010ceph5b5850a896c67fa43011/operations301064cd4f5c3011/adding-removing-osds3010589e52a0-52209664osd3011

## 6.3：添加元数据服务器

### 6.3.1、添加第一个元数据服务器

1、为mds元数据服务器创建一个目录

[root@node1 ~]# mkdir -p /var/lib/ceph/mds/ceph-node01

2、为bootstrap-mds客户端创建一个密钥 注：(如果下面的密钥在目录里已生成可以省略此步骤）

[root@node1 ~]# ceph-authtool --create-keyring /var/lib/ceph/bootstrap-mds/ceph.keyring --gen-key -n client.bootstrap-mds

3、在ceph auth库中创建bootstrap-mds客户端，赋予权限添加之前创建的密钥

注（查看ceph auth list 用户权限认证列表 如果已有client.bootstrap-mds此用户，此步骤可以省略）

[root@node1 ~]# ceph auth add client.bootstrap-mds mon 'allow profile bootstrap-mds' -i /var/lib/ceph/bootstrap-mds/ceph.keyring

added key for client.bootstrap-mds

4、在root家目录里创建ceph.bootstrap-mds.keyring文件

touch /root/ceph.bootstrap-mds.keyring

5、把keyring /var/lib/ceph/bootstrap-mds/ceph.keyring里的密钥导入家目录下的ceph.bootstrap-mds.keyring文件里

ceph-authtool --import-keyring /var/lib/ceph/bootstrap-mds/ceph.keyring ceph.bootstrap-mds.keyring

6、在ceph auth库中创建mds.node01用户，并赋予权限和创建密钥，密钥保存在/var/lib/ceph/mds/ceph-node01/keyring文件里

ceph --cluster ceph --name client.bootstrap-mds --keyring /var/lib/ceph/bootstrap-mds/ceph.keyring auth get-or-create mds.node01 osd 'allow rwx' mds 'allow' mon 'allow profile mds' -o /var/lib/ceph/mds/ceph-node01/keyring

7、为mds创建一个初始化文件用于启动使用（此文件为空文件）

[root@node1 ~]# touch /var/lib/ceph/mds/ceph-node01/sysvinit

8、为了防止重新被安装创建一个空的done文件

[root@node1 ~]# touch /var/lib/ceph/mds/ceph-node01/done

9、情况mds服务进程

[root@node1 ~]# service ceph start mds.node01

=== mds.node01 ===

Starting Ceph mds.node01 on node1...

starting mds.node01 at :/0

6.3.2：添加其它元数据服务器，参考6.3.1 步骤

添加第二个元数据服务器

添加第三个元数据服务器

1、查看osd目录树

ceph osd tree

查看集群状态

ceph -w

## 七：安装client端RBD、cephFS挂载演示

### 7.1：客户端内核要求、安装装准备

7.1.1、升级client端内核

1. 操作系统使用CentOS6.6，因为系统已经包含xfs的支持可以直接使用不需要再次编译。
2. 由于CentOS6.5系统的内核为2.6.32，因此要关闭硬盘的写入缓存，若高于此版本不需要关闭。
   * 1. #hdparm -W 0 /dev/sdb 0

7.1.2：升级系统内核

kernel 2.6.34以前的版本是没有Module rbd的，把系统内核版本升级到最新

rpm --import http://elrepo.org/RPM-GPG-KEY-elrepo.org

rpm -Uvh http://elrepo.org/elrepo-release-6-5.el6.elrepo.noarch.rpm

yum --enablerepo=elrepo-kernel install kernel-ml -y

安装完内核后修改/etc/grub.conf配置文件使

修改配置文件中的 Default=1 to Default=0

7.1.3、验证内核支持

[首先加载ceph.ko]

#modprobe -l|grep ceph

kernel/fs/ceph/ceph.ko

kernel/net/ceph/libceph.ko

#modprobe ceph

机器重启后生效 init 6

### 7.2、RBD方式挂载

7.2.1、安装ceph软件包

Yum install ceph –y

7.2.2、将node01配置文件copy client

rsync -avp node01:/etc/ceph/\* /etc/ceph/

7.2.3、新建一个ceph pool

ceph osd pool create {pool-name} {pg-num} [{pgp-num}]

[root@client ~]# ceph osd pool create kman 256

7.2.3、在pool中新建一个镜像

[root@client ~]# rbd create rbdtest-1 --size 80960 -p kman

7.2.4、把镜像映射到pool块设备中

[root@client ~]# rbd map rbdtest-1 -p kman

7.2.5、查看镜像映射map

[root@client ~]# rbd showmapped

id pool image snap device

0 kman rbdtest-1 - /dev/rbd0

7.2.6、格式化映射的设备块

[root@client ~]# mkfs.ext4 -m0 /dev/rbd0

参数：-m reserved-blocks-percentage

保留块百分比,-m:设置预留空间

默认是5%。也就是2G的空间，文件系统只是有2G\*95%

7.2.7、挂载新建的分区

[root@client ~]# mkdir /rbdtest

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

[root@client ~]# df -h

/dev/sda5 1.7T 4.8G 1.7T 1% /data

/dev/rbd0 78G 56M 78G 1% /rbdtest

7.2.8、进入新建的分区并dd测试性能

[root@client ~]# cd /rbdtest

7.2.9、其它参考命令

查看创建的池

ceph osd lspools

删除池

ceph osd pool delete kman kman --yes-i-really-really-mean-it

创建池参考资料

http://docs.openfans.org/ceph/ceph4e2d658765876863/ceph-1/ceph-storage-cluster3010ceph5b5850a896c67fa43011/operations301064cd4f5c3011/pools30106c603011

### 7.3、cephFS挂载

参考文档

<http://docs.openfans.org/ceph/ceph4e2d658765876863/ceph-1/copy_of_ceph-filesystem301065874ef67cfb7edf3011/manpage-mount-ceph3010mount-ceph5e2e52a998753011>

创建cephfs文件系统

对于一个刚创建的MDS服务，虽然服务是运行的，但是它的状态直到创建 pools 以及文件系统的时候才会变为Active.

[root@zeus-4-17 rbdtest]# ceph mds dump

dumped mdsmap epoch 1

epoch 1

flags 0

created 0.000000

modified 2015-01-26 21:49:35.368742

tableserver 0

root 0

session\_timeout 0

session\_autoclose 0

max\_file\_size 0

last\_failure 0

last\_failure\_osd\_epoch 0

compat compat={},rocompat={},incompat={}

max\_mds 0

in

up {}

failed

stopped

data\_pools

metadata\_pool 0

inline\_data disabled

7.3.1、创建cephfs 存储池

 ceph osd pool create {pool-name} {pg-num} [{pgp-num}]  
  {  
   Option implication:  
    {pool-name}--> 存储池名称，必须唯一。  
    {pg-num}--> 存储池拥有的归置组总数。  
    {pgp-num}--> 用于归置的归置组总数。

ceph osd pool create cephfs\_data 10

ceph osd pool create cephfs\_metadata 10

ceph fs new leadorfs cephfs\_metadata cephfs\_data

验证数据生成

[root@zeus-4-17 rbdtest]# ceph mds dump

dumped mdsmap epoch 5

epoch 5

flags 0

created 2015-01-27 16:10:16.789446

modified 2015-01-27 16:10:21.504363

tableserver 0

root 0

session\_timeout 60

session\_autoclose 300

max\_file\_size 100000000000000

last\_failure 0

last\_failure\_osd\_epoch 0

compat compat={},rocompat={},incompat={1=base v0.20,2=client writeable ranges,3=default file layouts on dirs,4=dir inode in separate object,5=mds uses versioned encoding,6=dirfrag is stored in omap,8=no anchor table}

max\_mds 5

in 0

up {0=4341}

failed

stopped

data\_pools 2

metadata\_pool 3

inline\_data disabled

4341: 172.16.2.29:6815/31531 'node03' mds.0.1 up:active seq 1493

7.3.2、客户端挂载

1、安装客户端挂载软件

yum install ceph-fuse -y

2、创建挂载目录

mkdir /mycephfs

3、模拟挂载

mount -t ceph node01:6789:/ /mycephfs -v -o name=admin,secret=AQCnrMZUaFsiCxAAXzM3aF9WjUBnwbN6PtvZEw==

返回结果：

parsing options: rw,name=admin,secret=AQCnrMZUaFsiCxAAXzM3aF9WjUBnwbN6PtvZEw==

4、证书来源位置

more /etc/ceph/ceph.client.admin.keyring #key证书来源

[client.admin]

key = AQCnrMZUaFsiCxAAXzM3aF9WjUBnwbN6PtvZEw==

auid = 0

caps mds = "allow"

caps mon = "allow \*"

caps osd = "allow \*"

5、另外一种命令挂载方式

[root@ceph-client ~]# mount -t ceph node01:6789:/ /mycephfs -v -o name=admin,secretfile=/etc/ceph/ceph.client.admin.keyring

6、若果有多个mon监控节点，可以挂载多可节点，保证了cephFS的安全行，当有一个节点down的时候不影响写入数据

[root@client ~]# mount.ceph node01,node02,node03:/ /mycephfs -v -o name=admin,secret= AQCnrMZUaFsiCxAAXzM3aF9WjUBnwbN6PtvZEw==

7、验证挂载信息：

df -h

Filesystem Size Used Avail Use% Mounted on

/dev/sda5 1.7T 4.8G 1.7T 1% /data

/dev/rbd0 78G 56M 78G 1% /rbdtest

172.16.2.27:6789:/ 8.2T 9.8G 8.2T 1% /mycephfs

8、把挂载的信息写到fstab里

[root@client ~]# vi /etc/fstab

172.16.2.27,172.16.2.28,172.16.2.29:/ /mycephfs ceph name=admin,secret= AQCnrMZUaFsiCxAAXzM3aF9WjUBnwbN6PtvZEw==,noatime 0

9、指定key文件的调用方式

sudo ceph-fuse -k ./ceph.client.admin.keyring -m 192.168.40.107:6789 ~/mycephfs

10、挂载后查看结果，注意观察类型

df -Th

11、取消挂载

取消挂载的操作如下:

sudo umount /mnt/mycephfs

## 八：检查ceph集群状态常用命令

#ceph health //ceph健康状态

#ceph status //ceph当前全部状态

#ceph -w //实时监控ceph状态及变化

#ceph osd dump //所有osd详细状态

#ceph osd tree //osd所在位置，及状态

#ceph quorum\_status //mon优先级状态

#ceph mon dump //mon节点状态

#ceph mds dump //mds详细状态

### 8.1、磁盘压力测试软件推荐

安装fio压测工具、Linux系统自带，yum即可安装

yum install fio –y

http://www.360doc.com/content/12/0809/11/2459\_229177650.shtml

http://www.gongzi.org/linux-io-test.html

## 九:故障排查

9.1、创建osd节点挂载磁盘报错，报错信息如下

[root@node01 ~]# mount -o user\_xattr /dev/sdb /var/lib/ceph/osd/ceph-0

mount: wrong fs type, bad option, bad superblock on /dev/sdb,

missing codepage or helper program, or other error

In some cases useful info is found in syslog - try

dmesg | tail or so

执行上面的命令会报错

解决的办法是：用下面的两条命令替代上面的一条命令。

[root@node01 ~]# mount /dev/sdb /var/lib/ceph/osd/ceph-0

[root@node01 ~]# mount -o remount,user\_xattr /var/lib/ceph/osd/ceph-0

9.2：执行ceph 命令报错

Traceback (most recent call last):

File "/usr/bin/ceph", line 63, in <module>

import rados

ImportError: No module named rados

解决： 注释自己安装python的环境变量

[root@vhadoop-18-14 ~]# more /etc/profile

#export PYTHON=/opt/python/

#export PATH=$PYTHON/bin:$PATH

#其它类似报错

Traceback (most recent call last):

File "/usr/local/bin/ceph", line 53, in <module>

import argparse

ImportError: No module named argparse

解决办法：

yum install python-pip

pip install argparse

pip install argparse --upgrade

/usr/local/bin/ceph osd create

Traceback (most recent call last):

File "/usr/local/bin/ceph", line 56, in <module>

import rados

ImportError: No module named rados cp -vf /usr/local/ceph/lib/python2.6/site-packages/\* /usr/lib64/python2.6

echo /usr/local/ceph/lib >/etc/ld.so.conf.d/ceph.conf

ldconfig

9.3：创建osd报错

[root@node01 ~]# ceph-osd -i 3 --mkfs --mkkey

2015-01-27 12:29:06.775251 7f4c5d342800 -1 OSD::mkfs: ObjectStore::mkfs failed with error -16

2015-01-27 12:29:06.775288 7f4c5d342800 -1 \*\* ERROR: error creating empty object store in /var/lib/ceph/osd/ceph-3: (16) Device or resource busy

排查解决：

[root@node01 ~]# mount 查看挂载分区

/dev/sda2 on / type ext4 (rw)

proc on /proc type proc (rw)

sysfs on /sys type sysfs (rw)

devpts on /dev/pts type devpts (rw,gid=5,mode=620)

tmpfs on /dev/shm type tmpfs (rw)

/dev/sda1 on /boot type ext4 (rw)

none on /proc/sys/fs/binfmt\_misc type binfmt\_misc (rw)

/dev/sdb on /var/lib/ceph/osd/ceph-0 type xfs (rw,user\_xattr)

此处的磁盘挂载出错啦，应该是一个新磁盘 /dev/sdc ,

/dev/sdb on /var/lib/ceph/osd/ceph-3 type xfs (rw,user\_xattr)

9.4：报错

在ceph auth库中创建mds.node01用户，并赋予权限和创建密钥，密钥保存在/var/lib/ceph/mds/ceph-node01/keyring文件里

root@node01 ~]# ^C

[root@node01 ~]# ceph --cluster ceph --name client.bootstrap-mds --keyring /var/lib/ceph/bootstrap-mds/ceph.keyring auth get-or-create mds.node01 osd 'allow rwx' mds 'allow' mon 'allow profile mds' -o /var/lib/ceph/mds/ceph-node01/keyring

2015-01-27 13:54:44.189535 7f79f9ec5700 0 librados: client.bootstrap-mds authentication error (1) Operation not permitted

Error connecting to cluster: PermissionError

解决办法：

确认key文件信息：

more /var/lib/ceph/bootstrap-mds/ceph.keyring

[client.bootstrap-mds]

key = AQDvRcZUALD4MxAAv4s4cMr6XraBt+Pw2Iowyg==

通过ceph auth list 命令查看审核信息，找到

client.bootstrap-mds

key: AQDvRcZUALD4MxAAv4s4cMr6XraBt+Pw2Iowyg==

caps: [mon] allow profile bootstrap-mds

确保/var/lib/ceph/bootstrap-mds/ceph.keyring 文件内容与ceph auth list 查看信息一致

9.5：报错

ceph -w

cluster bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

health HEALTH\_WARN too few pgs per osd (7 < min 20)

monmap e1: 3 mons at {node01=172.16.2.27:6789/0,node02=172.16.2.28:6789/0,node03=172.16.2.29:6789/0}, election epoch 8, quorum 0,1,2 node01,node02,node03

osdmap e73: 9 osds: 9 up, 9 in

pgmap v176: 64 pgs, 1 pools, 0 bytes data, 0 objects

9524 MB used, 8365 GB / 8374 GB avail

64 active+clean

解决办法：

解决办法, 需要修改pg\_num , pgp\_num .

先要获取pool name, 如下, 返回pool name : rbd.

ceph osd pool stats

pool rbd id 0

nothing is going on

修改pool对应的pg\_num和pgp\_num.

ceph osd pool set rbd pg\_num 128

specified pg\_num 128 <= current 128 #要大于当前128

ceph osd pool set rbd pg\_num 300

验证：

ceph -s

cluster bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

health HEALTH\_WARN too few pgs per osd (14 < min 20); pool rbd pg\_num 128 > pgp\_num 64

monmap e1: 3 mons at {node01=172.16.2.27:6789/0,node02=172.16.2.28:6789/0,node03=172.16.2.29:6789/0}, election epoch 8, quorum 0,1,2 node01,node02,node03

osdmap e75: 9 osds: 9 up, 9 in

pgmap v184: 128 pgs, 1 pools, 0 bytes data, 0 objects

9527 MB used, 8365 GB / 8374 GB avail

128 active+clean

同时还需要调整pgp\_num

[root@mon1 ~]# ceph osd pool set rbd pgp\_num 300

ceph -s

cluster bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

health HEALTH\_WARN too few pgs per osd (14 < min 20)

monmap e1: 3 mons at {node01=172.16.2.27:6789/0,node02=172.16.2.28:6789/0,node03=172.16.2.29:6789/0}, election epoch 8, quorum 0,1,2 node01,node02,node03

osdmap e78: 9 osds: 9 up, 9 in

pgmap v225: 128 pgs, 1 pools, 0 bytes data, 0 objects

9529 MB used, 8365 GB / 8374 GB avail

128 active+clean

最终验证：

[root@node02 ~]# ceph -s

cluster bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

health HEALTH\_OK

monmap e1: 3 mons at {node01=172.16.2.27:6789/0,node02=172.16.2.28:6789/0,node03=172.16.2.29:6789/0}, election epoch 8, quorum 0,1,2 node01,node02,node03

osdmap e83: 9 osds: 9 up, 9 in

pgmap v265: 300 pgs, 1 pools, 0 bytes data, 0 objects

9538 MB used, 8365 GB / 8374 GB avail

300 active+clean

9.6: 报错

[root@node01 ~]# ceph -w

cluster bc5ec6c4-3cc9-48ad-94d9-cee3cd75952a

health HEALTH\_WARN pool cephfs\_data has too few pgs