**Ceph集群部署**

**目录**

[1 环境介绍 2](#_Toc403045798)

[2 集群规划 2](#_Toc403045799)

[3 安装配置 3](#_Toc403045800)

[3.1 所有ceph节点 3](#_Toc403045801)

[修改机器名 3](#_Toc403045802)

[添加主机名解析 3](#_Toc403045803)

[磁盘大小 3](#_Toc403045804)

[时区时间 3](#_Toc403045805)

[sudoers权限： 4](#_Toc403045806)

[创建ceph用户 4](#_Toc403045807)

[3.2 ceph-deploy节点 4](#_Toc403045808)

[添加ceph源的key 4](#_Toc403045809)

[添加软件源 4](#_Toc403045810)

[更新源 4](#_Toc403045811)

[安装ceph-deploy 4](#_Toc403045812)

[添加ceph用户ssh无密码认证 5](#_Toc403045813)

[3.3 部署ceph集群 5](#_Toc403045814)

[创建集群文件目录 5](#_Toc403045815)

[修改ceph.conf配置 6](#_Toc403045816)

[安装软件包 6](#_Toc403045817)

[添加Monitors 6](#_Toc403045818)

[添加Osds 7](#_Toc403045819)

[添加MDSs 7](#_Toc403045820)

[4 创建存储池Pool 8](#_Toc403045821)

[创建存储池 8](#_Toc403045822)

[设置存储池副本数 8](#_Toc403045823)

[查看所有存储池 8](#_Toc403045824)

[5 常见集群维护 8](#_Toc403045825)

[6 附录 9](#_Toc403045826)

# 1 环境介绍

使用3台机器建设ceph集群，以下是集群信息：

|  |  |  |
| --- | --- | --- |
| **IP地址** | **主机名** | **备注** |
| 192.168.203.11 | ceph1 |  |
| 192.168.203.12 | ceph2 |  |
| 192.168.203.15 | ceph3 |  |

192.168.203.11 ceph1

192.168.203.12 ceph2

192.168.203.15 ceph3

每台机器都有第二块硬盘/dev/sdb,用来存贮ceph数据，

Admin node,monitor node 这两个节点可以在一起。

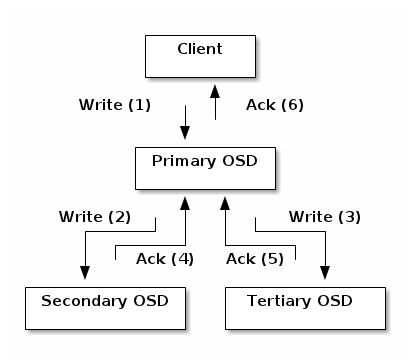
Object storage node (osd)

**OSDs、Monitor、MDS概念介绍：**

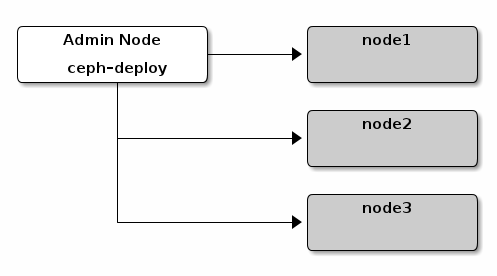
<http://ceph.com/docs/v0.80.5/start/intro/>

**Replication副本概念及Ceph架构介绍**：

<http://ceph.com/docs/v0.80.5/architecture/>



# 2 集群规划



|  |  |  |
| --- | --- | --- |
| **节点** | **角色** | **备注** |
| ceph1 | ceph-deploy、mon、osd、mds |  |
| ceph2 | mon、osd、mds |  |
| ceph3 | mon、osd、mds |  |

**注：**

1 ceph-deploy节点没有单独部署，ceph-deploy部署在ceph1节点上，ceph1节点上/home/ceph/my-cluster目录下放置的是集群相关的配置文件，用于自动部署其他节点及推送配置文件，不能轻易更改文件内容。

2 每个节点上都部署了ceph monitor（mon）

# 3 安装配置

## 3.1 所有ceph节点

### 修改机器名

# vi /etc/hostname

分别修改3台节点的主机名为：ceph1、ceph2、ceph3

修改主机名后，使之生效：

立即生效：执行命令hostname -F /etc/hostname，然后断开shell后重新登陆。

永久生效：重启服务器

### 添加主机名解析

在每个ceph节点上互相添加主机名解析，并保持一致：

# vi /etc/hosts

192.168.203.11 ceph1

192.168.203.12 ceph2

192.168.203.15 ceph3

### 磁盘大小

在每个节点上，检查所有osd磁盘的容量大小一定要保持一致

# fdisk -l

### 时区时间

查看每个节点时区、时间是否一致：

# date

如果时区不一致：

# dpkg-reconfigure tzdata

执行后会弹出重新选择时区的选项，然后选择“Asia→Shanghai”

### sudoers权限：

使ceph用户获取root权限并且免密码认证：

# echo "ceph ALL = (root) NOPASSWD:ALL" | sudo tee /etc/sudoers.d/ceph

# chmod 0440 /etc/sudoers.d/ceph

### 创建ceph用户

如果系统中已经有ceph用户，则不用执行这部分操作

如果系统中没有ceph用户，则创建ceph用户：

# useradd -m -d /home/ceph -k /etc/skel -s /bin/bash ceph

设置ceph用户密码：

# passwd ceph

## 3.2 ceph-deploy节点

### 添加ceph源的key

# wget -q -O- 'https://ceph.com/git/?p=ceph.git;a=blob\_plain;f=keys/release.asc' | sudo apt-key add -

### 添加软件源

# echo deb http://ceph.com/debian-firefly/ $(lsb\_release -sc) main | sudo tee /etc/apt/sources.list.d/ceph.list

### 更新源

# apt-get update

### 安装ceph-deploy

# apt-get install ceph-deploy

# ceph-deploy --version

### 添加ceph用户ssh无密码认证

**切换到ceph用户，然后执行：**

# sudo -l ceph

$ ssh-keygen -q -t rsa

$ vi ~/.ssh/config

新建~/.ssh/config文件并添加以下内容：

StrictHostKeyChecking no

Host ceph1

Hostname ceph1

User ceph

Host ceph2

Hostname ceph2

User ceph

Host ceph3

Hostname ceph3

User ceph

$ ssh-copy-id ceph@ceph1

$ ssh-copy-id ceph@ceph2

$ ssh-copy-id ceph@ceph3

(上面三条命令分别输入ceph用户的密码即完成无密码认证)

## 3.3 部署ceph集群

以下操作是用ceph用户登陆到ceph-deploy节点上（本文是ceph1节点）

### 创建集群文件目录

$ mkdir my-cluster

$ cd my-cluster

$ ceph-deploy new ceph{1..3}

$ ls

ceph.conf ceph.log ceph.mon.keyring

### 修改ceph.conf配置

$ vi my-cluster/ceph.conf

注释以下三行：

#auth\_cluster\_required = cephx

#auth\_service\_required = cephx

#auth\_client\_required = cephx

添加以下配置：

# Add

auth\_supported = none

# add public/cluster network

public\_network = 192.168.0.0/16

cluster\_network = 192.168.0.0/16

# special jornal size

osd\_journal\_size = 5120

# clock skew detected

mon\_clock\_drift\_allowed = 1

# Set Default Replication

osd\_pool\_default\_size = 2

### 安装软件包

$ ceph-deploy install ceph1

$ ceph-deploy install ceph2

$ ceph-deploy install ceph3

### 添加Monitors

$ ceph-deploy --overwrite-conf on create-initial

(注:如果报: [ceph\_deploy][ERROR ] KeyNotFoundError: Could not find keyring file: /etc/ceph/ceph.client.admin.keyring on host ceph1，则重启ceph相关服务: sudo restart ceph-all、sudo restart ceph-osd-all)

/etc/ceph目录下会出现两个文件：(在新加的ceph节点,Monitors不需要在安装,只需要拷贝配置文件即可)

$ ls /etc/ceph/

ceph.client.admin.keyring ceph.conf

$ find /var/lib/ceph/ -name "\*.keyring"

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

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

$ sudo ceph -s

cluster fe17a037-5d43-4690-9150-fa43062ddfb3

health HEALTH\_ERR 192 pgs stuck inactive; 192 pgs stuck unclean; no osds;

……

……

查看上面的Cluster集群ID是否与配置文件的一致，用以判断是否安装成功

### 添加Osds

查看每个节点可以作为osd的磁盘：

$ ceph-deploy disk list ceph1

$ ceph-deploy disk list ceph2

$ ceph-deploy disk list ceph3

zap（擦除）每节点的磁盘分区：(注意磁盘sdb不能有分区，如果有则需要删除)

$ ceph-deploy disk zap ceph1:sdb

$ ceph-deploy disk zap ceph2:sdb

$ ceph-deploy disk zap ceph3:sdb

创建osd：

$ ceph-deploy osd create ceph1:sdb

$ ceph-deploy osd create ceph2:sdb

$ ceph-deploy osd create ceph3:sdb

创建好所有osd后，执行ceph -s，如果时区和时间都一致的话，应该得到正确的输出：

$ ceph -s

cluster fe17a037-5d43-4690-9150-fa43062ddfb3

health HEALTH\_OK

monmap e3: 3 mons at {ceph1=192.168.203.11:6789/0,ceph2=192.168.203.12:6789/0,ceph3=192.168.203.15:6789/0}, election epoch 30, quorum 0,1,2 ceph1,ceph2,ceph3

mdsmap e11: 1/1/1 up {0=ceph1=up:active}, 2 up:standby

osdmap e22: 3 osds: 3 up, 3 in

pgmap v79: 192 pgs, 3 pools, 1884 bytes data, 20 objects

108 MB used, 1381 GB / 1381 GB avail

192 active+clean

如果不是health HEALTH\_OK，则说明有问题。

检查创建的osd是否正常

ceph osd tree

### 添加MDSs

$ ceph-deploy mds create ceph1

$ ceph-deploy mds create ceph2

$ ceph-deploy mds create ceph3

$ ceph -s

输出结果中会多出mds的信息：

mdsmap e11: 1/1/1 up {0=ceph1=up:active}, 2 up:standby

# 4 创建存储池Pool

## 创建存储池

$ ceph osd pool create volumes 150

$ ceph osd pool create images 150

注：150是存储池的PG数量，如何计算请见下面解释：

Here’s we created a storage pool called ‘datastore’ with 150 placement groups. The number of placement groups is based on **100 x the number of OSD’s / the number of replicas** you want to maintain. I want two copies of the data (so if a server fails no data is lost), so **3 x 100 / 2 = 150**.

## 设置存储池副本数

Ceph集群默认副本数为2个，假如设置为3个副本数：

$ ceph osd pool set volumes size 2

$ ceph osd pool set images size 2

## 查看所有存储池

Ceph集群创建好后，有3个默认的存储池，分别为data、metadata、rbd

$ rados lspools

data

metadata

rbd

volumes

images

# 5 常见集群维护

比如Mons、Osds等添加/删除操作：

<https://wiki.deimos.fr/Ceph_:_performance,_reliability_and_scalability_storage_solution#health_HEALTH_WARN_clock_skew_detected>

Ceph官方的维护文档：

<http://ceph.com/docs/v0.80.5/rados/operations/>

ceph.conf配置文件中，osd相关的配置选项：

<http://ceph.com/docs/argonaut/config-ref/osd-config/>

ceph.conf配置文件中，osd相关的配置选项：

<http://ceph.com/docs/argonaut/config-ref/mon-config/>

# 6 附录