Ubuntu下Ceph集群的安装

1. 安装概述

1.1 摘要

本章介绍Ceph分布式存储系统的安装及软硬件环境要求。

1.2 软件环境

本文采用的软件环境

操作系统：

Distributor ID:	Ubuntu
Description:	Ubuntu 14.04.4 LTS
Release:	14.04
Codename:	trusty

Ceph版本：	V10.2.2 jewel LTS

Jewel要求的环境：

Distro	Release	Code Name	Kernel	Notes	Testing
CentOS	7	N/A	linux-3.10.0		B, I, C
Debian	8.0	Jessie	linux-3.16.0	1, 2	B, I
Fedora	22	N/A	linux-3.14.0		B, I
RHEL	7	Maipo	linux-3.10.0		B, I
Ubuntu	14.04	Trusty Tahr	linux-3.13.0		B, I, C

1.3 硬件环境

4台虚拟机，硬件配置如下：

cpu信息：

```
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                2
On-line CPU(s) list:   0,1
Thread(s) per core:    1
Core(s) per socket:    1
Socket(s):             2
NUMA node(s):          1
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 42
Stepping:              1
CPU MHz:               1899.998
BogoMIPS:              3799.99
Hypervisor vendor:     KVM
Virtualization type:   full
L1d cache:             32K
L1i cache:             32K
L2 cache:              4096K
NUMA node0 CPU(s):     0,1
```

内存信息：2G

硬盘信息：

```
Filesystem      Size  Used Avail Use% Mounted on
udev            990M   12K  990M   1% /dev
tmpfs           201M  1.0M  200M   1% /run
/dev/vda1        52G   14G   36G  28% /
none            4.0K     0  4.0K   0% /sys/fs/cgroup
none            5.0M     0  5.0M   0% /run/lock
none           1001M     0 1001M   0% /run/shm
none            100M     0  100M   0% /run/user
```

1.4 安装准备

如果你安装的是Ceph的Firefly这个版本，可以用公司的mirrors.zte.com.cn这个源来快速安装Ceph，这个版本以上的在安装Ceph集群时需要保证机器能连接外网，而且把源换成mirrors.163.com，这个源是比较快的。具体怎么连接外网，可以参考公司的dev社区的这篇帖子：“公司内部使用linux外部镜像源安装软件，如docker等；以centos为例”地址：http://dev.zte.com.cn/topic/view/7148。
确保以上内容操作完成，才可以参考下面章节安装Ceph集群。

2. Ceph集群部署结构

该手册参考ceph官网的快速安装来部署ceph集群环境。部署结构图如下：

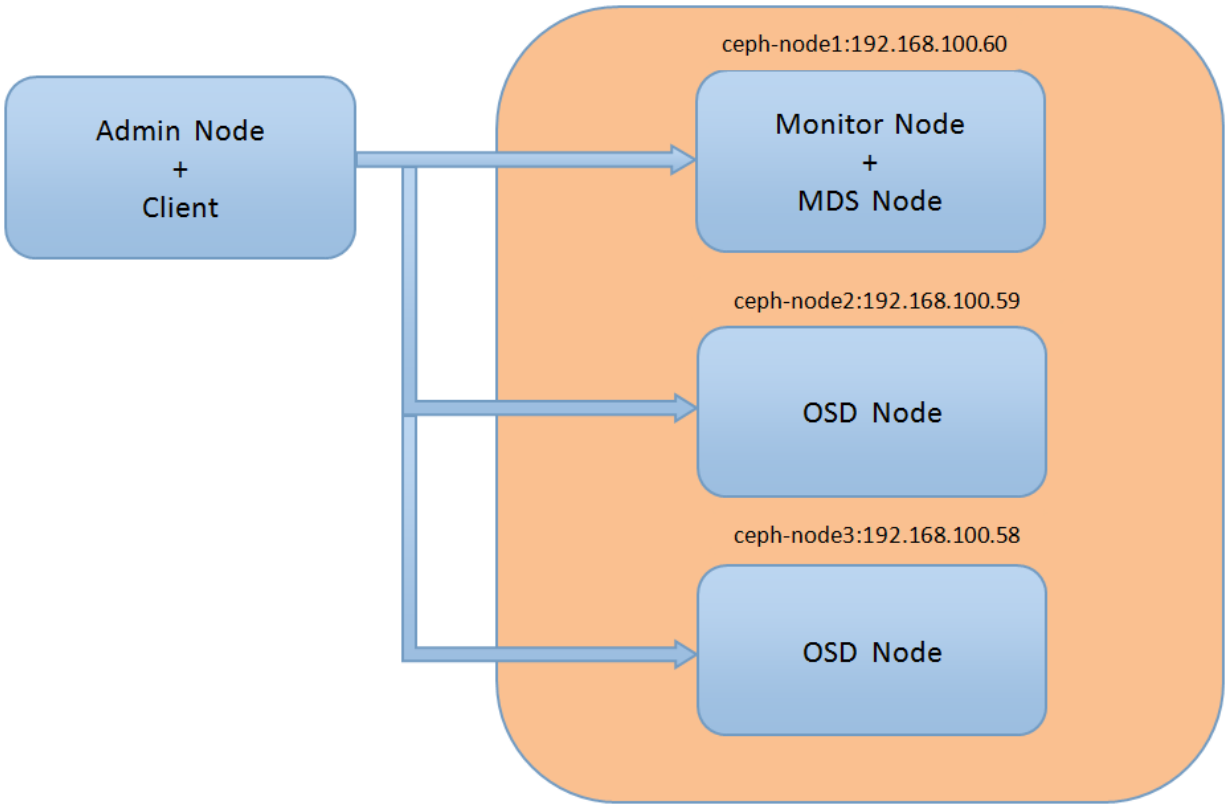


图4.1 ceph集群部署结构图

本文后续会针对该部署结构图的各个节点进行介绍

3 快速安装

快速安装有三个阶段：预检、存储集群和Ceph客户端。本文是针对ubuntu版本的linux系统进行安装Ceph集群，在其他版本的linux系统安装Ceph本文不涉及。

预检：在部署 Ceph 存储集群之前，需要对 Ceph 客户端和Ceph 节点进行一些基本的配置和检查。

存储集群：完成预检之后，你就可以开始部署 Ceph 存储集群了。

Ceph客户端：大多数 Ceph 用户不会直接往 Ceph 存储集群里存储对象，他们通常会使用 Ceph 块设备、 Ceph 文件系统、或 Ceph 对象存储这三大功能中的一个或多个。

3.1 预检

按照图4.1的部署结构安装一个 ceph-deploy 管理节点和一个三节点的Ceph 存储集群来研究 Ceph 的基本特性。这篇预检会帮你准备一个 ceph-deploy 管理节点、以及三个Ceph 节点（或虚拟机），以此构成 Ceph 存储集群。

3.1.1 安装Ceph部署工具（ceph-deploy）

把 Ceph 仓库添加到 ceph-deploy 管理节点，然后安装 ceph-deploy 。

1、 添加 release key ：

```
wget -q -O- 'https://download.ceph.com/keys/release.asc' | sudo apt-key add –
```

2、 添加Ceph软件包源，用Ceph稳定版，本文使用Ceph的最新版本jewel，例如：

```
echo deb http://download.ceph.com/debian-ceph-jewel/ $(lsb_release -sc) main | sudo tee /etc/apt/sources.list.d/ceph.list
```

3、 更新你的仓库，并安装 ceph-deploy ：

```
sudo apt-get update && sudo apt-get install ceph-deploy
```

3.1.2 Ceph节点的基本配置

你的管理节点必须能够通过 SSH 无密码地访问各 Ceph
节点。如果 ceph-deploy 以某个普通用户登录，那么这个用户必须有无密码使用 sudo 的权限。

1、 安装NTP

我们建议在所有 Ceph 节点上安装 NTP 服务（特别是 Ceph Monitor 节点），以免因时钟漂移导致故障。

```
sudo apt-get install ntp
```

确保在各 Ceph 节点上启动了 NTP 服务，并且要使用同一个 NTP 服务器。

2、 安装SSH服务器

一般情况，公司的机器或者虚拟机都是安装了ssh服务器的，如果没有安装，参考下面安装一下，在所有 Ceph 节点上执行如下步骤：

在各 Ceph 节点安装 SSH 服务器（如果还没有）：

```
sudo apt-get install openssh-server
```

确保所有 Ceph 节点上的 SSH 服务器都在运行。

3、 允许无密码SSH登录

官网推荐：ceph-deploy 工具必须以普通用户（不要用ceph作为用户名，已作为ceph的保留用户）登录 Ceph
节点，且此用户拥有无密码使用 sudo 的权限，因为它需要在安装软件及配置文件的过程中，不必输入密码。但由于公司的代理设置问题，只有root用户才能从163的源下载软件，普通用户无法下载（该问题暂时没找到解决办法），所以本文全程都是使用root用户安装，可能在安装过程中遇到”Permission
Denied”此类问题，解决办法是，用chown命令修改所属组为相应的用户（一般是ceph用户，该用户是ceph自动生成的用户）。

正因为 ceph-deploy 不支持输入密码，你必须在管理节点上生成 SSH 密钥并把其公钥分发到各 Ceph 节点。

生成 SSH 密钥对，提示 “Enter passphrase” 时，直接回车，口令即为空：

```
ssh-keygen


Generating public/private key pair.
Enter file in which to save the key (/ceph-admin/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /ceph-admin/.ssh/id_rsa.
Your public key has been saved in /ceph-admin/.ssh/id_rsa.pub.
```

把公钥拷贝到各 Ceph 节点，把下列命令中的 {username} 替换成新创建的用户名，本文用root。

```
ssh-copy-id {username}@node1
ssh-copy-id {username}@node2
ssh-copy-id {username}@node3
```

4、（推荐做法）修改 ceph-deploy 管理节点上的 ~/.ssh/config 文件，这样 ceph-deploy 就能用你所建的用户名登录 Ceph 节点了，而无需每次执行 ceph-deploy 都要指定 --username {username} 。这样做同时也简化了 ssh 和 scp 的用法。把 {username} 替换成你创建的用户名，本文采用root。

```
Host node1
   Hostname node1
   User {username}
Host node2
   Hostname node2
   User {username}
Host node3
   Hostname node3
   User {username}
```

例如，我的/root/.ssh目录下的config内容如下：

```
Host node1
   Hostname node1
   User root
Host node2
   Hostname node2
   User root
Host node2
   Hostname node2
   User root
```

用 ping 短主机名（ hostname -s ）的方式确认网络联通性，你可能需要修改/etc/hosts文件，

```
/etc/hosts:
192.168.100.25  node1
192.168.100.26  node2
192.168.100.27  node3
```

3.2 建立存储集群

预检完成后，本章节会通过ceph-deploy工具从管理节点（Admin Node）建立一个Ceph集群，该集群按照第四章中的部署结构图来探索Ceph的功能，包括三个节点：1个Monitor节点（也是MDS节点，cephfs会用到）和2个OSD节点

3.2.1 创建集群（部署Monitor节点）

为获得最佳体验，先在管理节点上创建一个目录，用于保存 ceph-deploy 生成的配置文件和密钥对。

```
mkdir my-cluster
cd my-cluster
```

在管理节点上，进入刚创建的放置配置文件的目录，用 ceph-deploy 执行如下步骤。

1、 创建集群

```
ceph-deploy new {initial-monitor-node(s)}
```

例如：

```
ceph-deploy new node1
```

在当前目录下用 ls 和 cat 检查 ceph-deploy 的输出，应该有一个 Ceph 配置文件(ceph.conf)、一个 monitor 密钥环(ceph.mon.keyring)和一个日志文件(ceph-deploy-ceph.log)。

2、 把 Ceph 配置文件里的默认副本数从 3 改成 2 ，这样只有两个 OSD 也可以达到 active + clean 状态。把下面这行加入 [global] 段：

```
osd pool default size = 2
```

3、 如果你有多个网卡，可以把 public network 写入 Ceph 配置文件的 [global] 段下

```
public network = {ip-address}/{netmask}
```

例如：

```
public network =192.168.100.27/24
```

4、 安装Ceph

```
ceph-deploy install {ceph-node} [{ceph-node} ...]
```

例如：

```
ceph-deploy install node1 node2 node3
```

ceph-deploy 将在各节点安装 Ceph 。 注：如果你执行过 ceph-deploy purge ，你必须重新执行这一步来安装 Ceph 。

5、 配置初始 monitor(s)、并收集所有密钥：

```
ceph-deploy mon create-initial
```

完成上述操作后，当前目录里应该会出现这些密钥环：

- {cluster-name}.client.admin.keyring
- {cluster-name}.bootstrap-osd.keyring
- {cluster-name}.bootstrap-mds.keyring
- {cluster-name}.bootstrap-rgw.keyring

3.2.2 部署OSD节点

为了快速地安装，这篇快速入门把目录而非整个硬盘用于 OSD 守护进程。如何为 OSD 及其日志使用独立硬盘或分区，本手册不做深入讲解。登录到 Ceph 节点（即部署图中的ceph-node2）、并给 OSD 守护进程创建一个目录。

1、 添加两个OSD

```
ssh ceph-node2
sudo mkdir /var/local/osd0
exit

ssh ceph-node3
sudo mkdir /var/local/osd1
exit
```

然后，从管理节点执行 ceph-deploy 来准备 OSD 。

```
ceph-deploy osd prepare {ceph-node}:/path/to/directory
```

例如：

```
ceph-deploy osd prepare ceph-node2:/var/local/osd0  ceph-node3:/var/local/osd1
```

最后，激活 OSD 。

```
ceph-deploy osd activate {ceph-node}:/path/to/directory
```

例如：

```
ceph-deploy osd activate ceph-node2:/var/local/osd0  ceph-node3:/var/local/osd1
```

注：激活 OSD时,遇到"filestore(/var/local/osd0) mkfs: write_version_stamp() failed: (13) Permission denied",导致无法部署。原因：创建/var/local/osd0时，使用的是root用户，而后续访问这个目录的用户是ceph系统默认的ceph用户，所以无权限,。解决办法：通过chown -R ceph:ceph /var/local/osd0后部署成功。

2、 用 ceph-deploy 把配置文件和 admin 密钥拷贝到管理节点和 Ceph 节点，这样你每次执行 Ceph 命令行时就无需指定 monitor 地址和ceph.client.admin.keyring 了。

```
ceph-deploy admin  {ceph-node}
```

例如：

```
ceph-deploy admin ceph-node1 ceph-node2 ceph-node3
```

3、 确保你对 ceph.client.admin.keyring 有正确的操作权限。

```
sudo chmod +r /etc/ceph/ceph.client.admin.keyring
```

4、 检查集群的健康状况。

```
ceph health
```

等 peering 完成后，集群应该达到 active + clean 状态(可以用ceph –s 命令查看)，，用 ceph-deploy 部署完成后它会自动启动集群。ceph health查看显示“HEALTH_OK”，或执行ceph -s 会出现类似如下信息，说明ceph集群正常启动。

```
cluster aa70be96-884f-41cf-9164-5b1e2bf753f8
     health HEALTH_OK
     monmap e1: 1 mons at {skangchao-4=192.168.100.27:6789/0}
            election epoch 5, quorum 0 skangchao-4
      fsmap e10: 1/1/1 up {0=skangchao-4=up:active}
     osdmap e26: 2 osds: 2 up, 2 in
            flags sortbitwise
      pgmap v52208: 320 pgs, 3 pools, 6252 MB data, 1593 objects
            27631 MB used, 71579 MB / 102 GB avail
                 320 active+clean
```

注：执行ceph health时，可能显示** ERROR: osd init failed: (36) File name too long，原因：jewel官方不建议使用ext4文件系统，该文件系统不支持ceph过长的object名字

解决办法：在每个ceph节点的/etc/ceph/ceph.conf下添加两句：

```
osd max object name len = 256
osd max object namespace len = 64
```

然后，重启ceph-all服务，在用ceph health查看ceph集群的健康状况，可用下面命令重启

```
sudo stop ceph-all
sudo start ceph-all
```

5、 如果在某些地方碰到麻烦，想从头再来，可以用下列命令清除配置：

```
ceph-deploy purgedata {ceph-node} [{ceph-node}]
ceph-deploy forgetkeys
```

用下列命令可以连 Ceph 安装包一起清除：

```
ceph-deploy purge {ceph-node} [{ceph-node}]
```

如果执行了 purge ，你必须重新安装 Ceph 。

3.2.3 部署MDS节点（可选）

如果使用CephFs访问ceph集群，则至少需要一个元数据服务器才能使用 CephFS
（块存储和对象存储的方式不需要安装MDS进程），执行下列命令创建元数据服务器：

```
ceph-deploy mds create {ceph-node}
```

例如：

```
ceph-deploy mds create node1
```

Note：当前生产环境下的 Ceph
只能运行一个元数据服务器。你可以配置多个，但现在Ceph官方还不会为多个元数据服务器的集群提供商业支持。

3.3 Ceph客户端安装

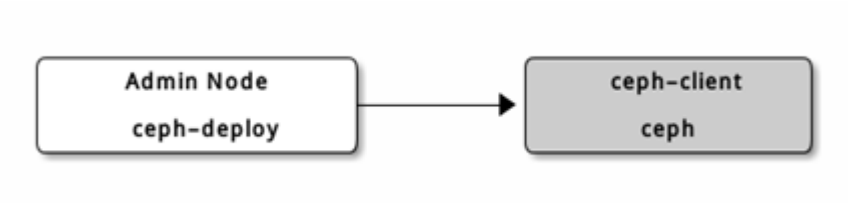
Ceph 独一无二地用统一的系统提供了对象、块、和文件存储功能，用户只需根据业务需求，安装对应的客户端，就可以使用该ceph存储集群了。 Ceph 客户端包括三种服务接口，有：

- 块设备： Ceph 块设备（也叫 RBD ）服务提供了大小可调、精炼、支持快照和克隆的块设备。为提供高性能， Ceph 把块设备条带化到整个集群。 Ceph 同时支持内核对象（ KO ） 和 QEMU 管理程序直接使用``librbd``——避免了内核对象在虚拟系统上的开销。
- 对象存储： Ceph 对象存储（也叫 RGW ）服务提供了 RESTful 风格的 API ，它与 Amazon S3 和 OpenStack Swift 兼容。
- 文件系统： Ceph 文件系统（ CephFS ）服务提供了兼容 POSIX 的文件系统，可以直接 mount 或挂载为用户空间文件系统（ FUSE ）。

本文只针对块设备（RBD）和文件系统（cephfs）两种客户端访问方式，介绍其安装方法。

3.3.1 块设备快速入门

确保 Ceph 存储集群处于 active + clean 状态，这样才能使用 Ceph
块设备。通过管理节点的ceph-deploy工具将ceph部署到ceph-client客户端。



你可以在虚拟机上运行 ceph-client 节点，但是不能在与 Ceph 存储集群（除非它们也用 VM ）相同的物理节点上执行下列步骤。

1、 安装Ceph

在管理节点上，通过 ceph-deploy 把 Ceph 安装到 ceph-client 节点。

```
ceph-deploy install ceph-client
```

在管理节点上，用 ceph-deploy 把 Ceph 配置文件和 ceph.client.admin.keyring 拷贝到 ceph-client 。

```
ceph-deploy admin ceph-client
```

ceph-deploy 工具会把密钥环复制到 /etc/ceph 目录，要确保此密钥环文件有读权限（如 sudo chmod +r /etc/ceph/ceph.client.admin.keyring ）

2、 配置块设备

在 ceph-client 节点上创建一个块设备 image 。

```
rbd create {image-name} --size 4096 [-m {mon-IP}] [-k /path/to/ceph.client.admin.keyring]
```

-m，-k参数可选，如果在/etc/ceph/目录下有ceph.conf和ceph.client.admin.keyring这两个文件，这两个参数可以省去；在创建image时，一般情况下是需要先创建存储池（pool）,再创建image。如果不指定存储池，创建的image是放在ceph默认的rbd存储池中的。

例如，下面image为foo的镜像是rbd池所有，

```
rbd create foo --size 4096 -m 192.168.100.27 -k /etc/ceph/ceph.client.admin.keyring
```

要在 swimmingpool 这个存储池中创建一个名为 bar 、大小为 1GB 的映像，不指定monitor和用户密钥环，执行：

```
rbd create --size 1024 swimmingpool/bar
```

在 ceph-client 节点上，把 image 映射为块设备。

```
rbd map {image-name} --name client.admin [-m {mon-IP}] [-k /path/to/ceph.client.admin.keyring]
```

例如，

```
rbd map foo –name client.admin -m192.168.100.27 -k /etc/ceph/ceph.client.admin.keyring
```

注：经过实践，上述命令在ceph的Jewel版本下执行报错：“RBD image feature set mismatch. You can disable features unsupported by the kernel with "rbd feature disable".”，原因：jewel 的image-format默认是使用的2，且默认全部开启所有功能（包括：layering, exclusive-lock, object-map, fast-diff,
deep-flatten），3.11内核以下的linux不支持某些特性导致。本文实践时的内核版本是4.2.0-27-generic，但也出现的以上问题，所以如果你的内核版本低于3.11，可以升级内核得以解决，或者高于3.11还存在该问题，可以使用一下命令，开启一些必要的功能，不要全部开启，

```
rbd create mypool/foo --size 10G --image-format 2 --image-feature  layering
```

在 ceph-client 节点上，创建文件系统后就可以使用块设备了，此命令可能耗时较长。

```
mkfs.ext4 -m0 /dev/rbd/rbd/foo
```

在 ceph-client 节点上挂载此文件系统。

```
mkdir /mnt/ceph-block-device
mount /dev/rbd/rbd/foo /mnt/ceph-block-device
cd /mnt/ceph-block-device
```

通过以上两步，ceph客户端就安装好了，你的应用应该部署到该ceph-client机器上，通过该ceph-client和ceph集群进行交互

3.3.2 CEPH 文件系统快速入门

1、 准备工作

确认你使用了合适的内核版本

```
lsb_release -a
uname -r
```

在管理节点上，通过 ceph-deploy 把 Ceph 安装到 ceph-client 节点上。

```
ceph-deploy install ceph-client
```

在管理节点上，用 ceph-deploy 把 Ceph 配置文件和 ceph.client.admin.keyring 拷贝到 ceph-client 。

```
ceph-deploy admin ceph-client
```

ceph-deploy 工具会把密钥环复制到 /etc/ceph 目录，要确保此密钥环文件有读权限（如 sudo chmod +r /etc/ceph/ceph.client.admin.keyring ）

确保 Ceph 存储集群在运行，且处于 active + clean 状态。同时，确保至少有一个 Ceph 元数据服务器在运行。

```
ceph –s
```

查看mds状态

```
Initctl list | grep ceph
```

显示：ceph-mds (ceph/skangchao-4) start/running, process 12964，说明mds已经启动成功。

2、 创建文件系统

虽然已创建了元数据服务器，但如果你没有创建存储池和文件系统，它是不会变为活动状态的。

```
ceph osd pool create cephfs_data <pg_num>
ceph osd pool create cephfs_metadata <pg_num>
ceph fs new <fs_name> cephfs_metadata cephfs_data
```

cephfs_data和cephfs_metadata
是存储池（pool）的名字，cephfs需要两个存储池，前者放置用户数据，后者放置元数据；pg_num是存储池的属性，即放置组的个数，pg_num的值和ceph的性能有一定关系，ceph fs new 命令负责将两个池格式化为ceph文件系统。

下面是pg_num的几个常用值

- 少于 5 个 OSD 时可把 pg_num 设置为 128
- OSD 数量在 5 到 10 个时，可把 pg_num 设置为 512
- OSD 数量在 10 到 50 个时，可把 pg_num 设置为 4096

例如，

```
ceph osd pool create cephfs_data 100
ceph osd pool create cephfs_metadata 100
ceph fs new paas_cephfs cephfs_metadata cephfs_data
```

3、 创建密钥文件

Ceph 存储集群默认启用认证，你应该有个包含密钥的配置文件（但不是密钥环本身）。用下述方法获取某一用户的密钥：

在密钥环文件中找到与某用户对应的密钥，例如：

```
cat ceph.client.admin.keyring
```

找到用于挂载 Ceph 文件系统的用户，复制其密钥。大概看起来如下所示：

```
[client.admin]
  key = AQCj2YpRiAe6CxAA7/ETt7Hcl9IyxyYciVs47w==
```

新建一个文件/path/to/admin.secret，将以上的密钥粘贴进去，像这样，

```
AQCj2YpRiAe6CxAA7/ETt7Hcl9IyxyYciVs47w==
```

保存文件，确保此文件对用户有合适的权限，但对其他用户不可见。

4、 挂载文件系统

挂载文件系统有两种方式：内核驱动和FUSE（用户空间文件系统），可以选择一种方式进行挂载。

（1） 把 Ceph FS 挂载为内核驱动。

```
mkdir /mnt/mycephfs
mount -t ceph {ip-address-of-monitor}:6789:/ /mnt/mycephfs
```

但是，Ceph 存储集群默认需要认证，所以挂载时需要指定用户名 name 和创建密钥文件一节中创建的密钥文件 secretfile ，例如：

```
mount -t ceph 192.168.100.27:6789:/ /mnt/mycephfs -o name=admin,secretfile=/path/to/admin.secret
```

注：以上是ceph官方推荐的做法，这样做可以确保密钥安全，但是经实践后，执行以上命令后，会报错：

```
mount: wrong fs type, bad option, bad superblock on 192.168.100.27:6789:/,
missing codepage or helper program, or other error
(for several filesystems (e.g. nfs, cifs) you might
need a /sbin/mount.<type> helper program)
In some cases useful info is found in syslog - try
dmesg | tail or so
```

把命令修改成下面这样，执行顺利成功，这种方式的缺点是密钥会留在Bash中，不是很安全。

```
mount -t ceph 192.168.100.27:6789:/ /mnt/mycephfs -o name=admin,secret=AQCj2YpRiAe6CxAA7/ETt7Hcl9IyxyYciVs47w==
```

注：从管理节点而非服务器节点挂载 Ceph FS 文件系统

（2） 用户空间文件系统（FUSE）

把 Ceph FS 挂载为用户空间文件系统（ FUSE ）。

```
mkdir ~/mycephfs
ceph-fuse -m {ip-address-of-monitor}:6789 ~/mycephfs
```

Ceph 存储集群默认要求认证，需指定相应的密钥环文件，除非它在默认位置（即 /etc/ceph ）：

```
ceph-fuse -k ./ceph.client.admin.keyring -m 192.168.0.1:6789 ~/mycephfs
```