有关ceph搭建与各种问题解决办法

V1.0

修改记录

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本号 | 拟制人/  修改人 | 拟制日期/  修改日期 | 更改理由 | 主要更改内容  （写要点即可） |
| V1.0 | 纪成 | 2016-09-19 | 初稿 | 初稿 |
|  |  |  |  |  |
|  |  |  |  |  |

目 录

[1 背景 4](#_Toc450899513)

[2 解决思路 5](#_Toc450899514)

[2.1 软件需求 5](#_Toc450899515)

[2.2 操作步骤 5](#_Toc450899516)

[2.2.1 代码库未下载到本地 5](#_Toc450899517)

[2.2.2 代码库已下载到本地 6](#_Toc450899518)

[2.2.3 后期更新子目录 6](#_Toc450899519)

[2.2.4 移除Spare-checkout功能 6](#_Toc450899520)

[2.2.5 指定排除的目录 7](#_Toc450899521)

[2.2.6 忽略指定的目录或文件，不做版本管控 8](#_Toc450899522)

[3 实践情况 9](#_Toc450899523)

[4 效果评价 9](#_Toc450899524)

[5 推广建议 9](#_Toc450899525)

[6 参考资料 9](#_Toc450899526)

【摘要】

到了5G时代之后，我们的网管是一个个基于Paas平台部署的微服务，和曾经基于一台主机（刀片）有很大的区别。在曾经虽然有多机部署的情况，可是对于分布式文件系统的需求，还仅限于非常底层与初级的文件与文件之前的互传与共享。

如今，我们微服务与微服务之间是完全无状态且独立的，每次Paas平台启动一个微服务都有可能变换到不同的主机上面去，如此一来，使用一个成熟且稳定的分布式文件系统，变得非常必要。

本文主要针对公司内网的情况下，如何部署一个分布式文件系统—ceph。并将部署过程中，可能遇到的问题一并总结给出，为将来再次部署给出一盏明灯。

【关键词】

微服务 分布式文件系统 Paas平台

# 准备工作

Ceph是一个开源、可靠性高、管理简单的分布式文件系统，本文主要内容是如何进行ceph的快速简单的搭建与搭建问题解决办法，如果想了解ceph的详细的介绍请移步：<http://docs.ceph.org.cn/start/intro/。>

要搭建一个最简单的ceph分布式文件系统集群，我们至少需要三台主机，我自己使用的基础三台主机配置如下：

节点一（node1）：8G内存、500G硬盘

节点二（node2）：4G内存、320G硬盘

节点三（node3）：4G内存、320G硬盘

因为ceph只能运行在unix类型的操作系统之上，所以我将所有主机都安装上了ubuntu14.04，Linux内核版本都是3.19.0-25-generic。具体安装Ubuntu、配置访问公司外网、安装ftp、安装ssh请移步我的wiki文章：<http://wiki.zte.com.cn/pages/viewpage.action?pageId=24189328>

要保证三台主机都能够连接到公司外网，并且保证三台主机的ssh连接都可使用

# 配置ssh

## 配置

由于后期使用ceph的管理工具来管理各个节点的主机，所以要保证在管理节点上面要能够无密码的访问各个其他节点，因为ceph管理工具是不支持输入密码的。要实现无密码ssh访问各个节点，主要就是使用ssh秘钥对进行分发。本文都是使用root用户直接访问与登陆的，所以保证当前用户切换到了root。

首先我们直接生成ssh秘钥：

**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.**

过程中会要你输入密码，都不要输入，直接回车。

接下来我们要保证能使用短主机名直接访问远程主机。主要就是ping node2或者ping node3这种能够ping通。主要在/etc/hosts文件中添加如下几条：

| **/etc/hosts:**  **192.168.100.25  node1**  **192.168.100.26  node2**  **192.168.100.27  node3** |
| --- |

其中前面是IP地址，后面是主机名。

最后我们向各个节点分发我们的ssh秘钥，在终端中依次执行下面命令：

**ssh-copy-id {username}@node1**

**ssh-copy-id {username}@node2**

**ssh-copy-id {username}@node3**

其中的username我们这里使用的是root，@符号后面是我们短主机名。

（推荐做法）修改管理节点上的 ~/.ssh/config 文件，这样 ceph管理工具就能用你所建的用户名登录每个节点了，而无需每次执行 ceph管理命令都要指定 --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 root**

**Host node2**

**Hostname node2**

**User root**

**Host node2**

**Hostname node2**

**User root**

## 可能遇到的问题

在配置ssh过程中，我们主要注意一个问题，也是我初期遇到的。就是要知道这里所说的主机名，类似于node1、node2等，是真实的主机名，要修改一定要保证你主机真实名称与这里使用的主机名一直，不要仅仅配置了hosts里面能够ping通的短名称，而真实主机名和这里使用的名称不一致，如此一来回导致后面使用ceph的管理工具进行ceph操作的时候出现各种各样的问题，当时令我最头疼的就是进行monitor主机初始化的时候，一直报错，后来发现是主机名不对。

Ubuntu更改主机名也是很简单的，修改/etc/hostname文件里面的名称。（里面就一行，很容易修改）改完之后使用reboot命令，重启计算机即可。

# 安装ceph管理工具ceph-deploy

## 安装

首先添加release key：

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

然后再添加ceph的软件包源，使用ceph稳定版替换下面的{ceph-stable-release}:

echo deb http://download.ceph.com/debian-{ceph-stable-release}/ $(lsb\_release -sc) main | sudo tee /etc/apt/sources.list.d/ceph.list

最后更新仓库，并安装ceph-deploy：

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

## 可能遇到的问题

这里我们可能遇到的问题，主要就是内部网络访问外部网络的问题。我刚开始添加release key的时候怎么都添加不成功，连接不上。Wget这个命令是一个Ubuntu内置的一个下载工具，-q参数是不提示下载过程系统。如果下不下来，我们可以直接从https://download.ceph.com/keys/release.asc这个网址直接下载，然后在终端中进入到下载目录下面，使用apt-key add release.asc进行添加key。

# 配置NTP

## NTP服务安装与配置

NTP服务是保证几台机器上面时钟能够同步的一个工具，我们这里使用NTP服务主要是能够避免时钟偏移的情况出现。在官方文档与以往的wiki文档中都只是教怎么安装，并没有介绍如何使用，其实使用NTP是非常容易的。下面直接使用apt-get进行安装：

| **sudo apt-get install ntp** |
| --- |

在每一台节点机器上面都安装NTP。然后在每一台机器上面都配置一下etc/ntp.conf这个文件，主要添加如下几个配置：

Server 10.9.177.33//后面的ip是我自己的，主要配置上游ip

Server 127.127.1.0

Fudge 127.127.1.0 stratum 10

Restrict 10.9.177.33

Restrict 10.9.0.0 mask 255.255.255.128 nomodify//同样ip段和子网掩码是我自己的

配置之后，重启node1节点上面的NTP服务（我的node1节点ip是10.9.177.33），命令是service ntp restart。然后登陆到每一个其他节点主机上面，使用

ntpdate 10.9.177.33这条命令，与node1上面的时间进行对齐。如此一来，就不会再后面出现时钟偏移的问题了。

## 可能出现的问题

这个地方，我要尤为强调的是，如果不在这里配置ntp服务器，不进行对齐时间操作，由于我们公司使用的是代理的方式访问外网的，那就导致我们普通安装Ubuntu系统的时候并不能连接到外网，最终导致安装之后时间是错乱的，而且即使后面配置了公司的代理，我们也要登陆之后才能够访问外网，导致很多时候，也是获取不到网络时间的。那么这里使用ntp对齐时间是非常必要的。因为后面ceph的几大关键性的配置环境，如果几台机器的时间不一致，会直接导致配置不成功，这也是我后面阻塞好几天的“血泪”史，所以在这个最开始的环境，一定要配置好NTP服务，并进行对时！

# 部署monitor节点，

## 部署monitor

Monitor节点，在ceph体系中可以被看做一个“接口人”角色，任何后面的各种面向客户端的操作，都是通过这个monitor节点进行的，其中包括key的分发、块设备的挂载、文件系统的挂载等，所以这个节点非常关键，一般一个成熟的ceph集群中是拥有几个monitor节点的，在此我们就配置一个monitor。

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。

**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)。

 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**

对每个节点上，安装ceph：

**ceph-deploy install node1 node2 node3**

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

## 可能遇到的问题

这个地方遇到的问题主要集中在初始化各个节点的ceph这个环节上面。我自己几乎就没有成功执行过ceph-deploy install node1 node2 node3这个命令的，原因很简单，主要就是网络连接太慢，我们网络代理不给力等等的网络网速问题，解决主要办法就是登陆到各个主机上面，直接使用apt-get install ceph这个命令进行安装。并且在使用ceph-deploy install这个命令上面，会列出他会下载什么工具，我们一并都可以登陆到具体的主机上面，使用apt-get install这个命令进行本地安装部署，这样一来会可靠些，成功率大大的提升。

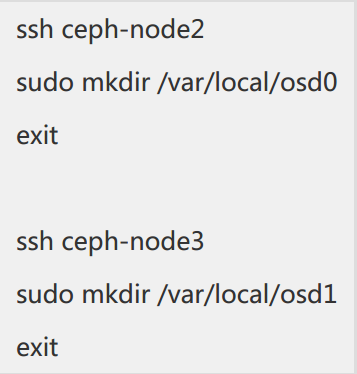
另外，在ceph-deploy mon create-initial这个名初始化monitor节点的时候，我也是阻塞了很久，最后发现是因为ping通的主机名，和真实的主机名并不一致，导致的这个问题，这个我在前面已经强调过了。

# 部署OSD节点

## 部署OSD

OSD节点，在整个ceph的集群中，充当一个“支撑”的作用，OSD节点越多，整个ceph的集群效率会越佳。OSD节点的部署是可以使用整个磁盘分区来进行的，此次我们部署的为了快速与学习使用，使用了单个目录来进行的部署，如果使用磁盘来部署的话，可以参考官方中文文档。

我们要添加两个OSD节点，首先分别登陆两个要充当OSD节点的主机，创建文件夹：



然后在管理节点上面执行ceph-deploy来准备OSD



最后激活两个OSD节点



要注意的是，我们这里使用的root用户创建的OSD节点上面的文件夹，可是真正ceph使用却是ceph自己创建的用户，我们要注意创建文件夹之后要把文件夹的所属用户组改成ceph可用的用户组，否则部署OSD节点的过程之后可能会报错，命令如下：



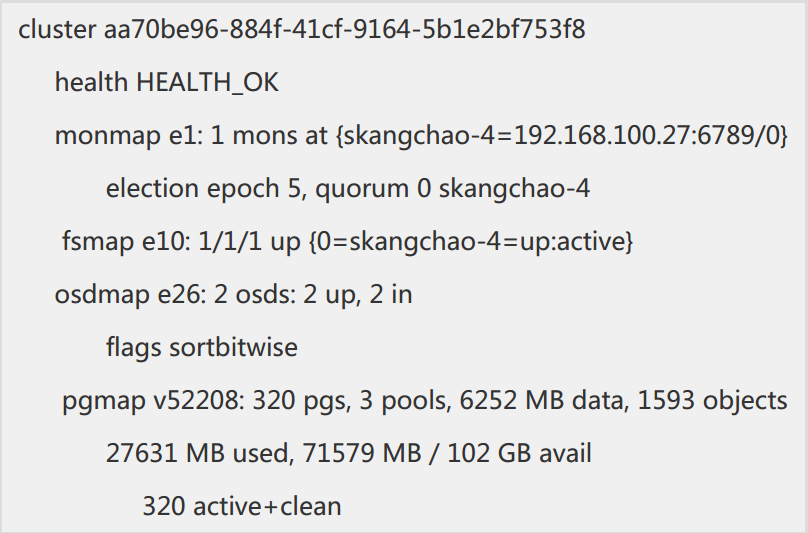
我们也可以把管理节点上面的ceph秘钥拷贝到两个OSD节点上面，这样在后续的操作就无需指定秘钥了：



至此，我们的集群就已经建立完全了，后续的其实都是使用集群的方式了。我们可以通过如下方式查看集群的健康状态：

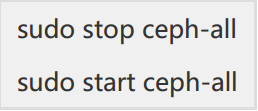


如果正常的情况会显示HEALTH\_OK：



## 可能遇到的问题

## 这个环节可能是我们会碰到最多问题的环节，我们wiki上面现成的文档，只是介绍了整个部署的流程，也是我一直觉得不厚道的地方。因为这个东西，并不是按照流程走就能够完事的，整个过程会遇到种种，你想都想象不到的问题，一度会把人逼入“绝境”！所以我觉得不仅仅要总结整个部署的流程，还要将遇到的问题一并总结出来，才能算是一个完整的技术文档。例如在这个部署OSD环节， 激活我们OSD节点之后很难看到HEALTH\_OK的这个健康状态，我当时就为了这个一度抓狂。但是不要气馁，直接细心的去翻看ceph的日志，记住，日志日志，日志是很关键解决问题的途径，里面记录了出错问题的详细。Ceph日志所在目录是/var/log/ceph/下面。如果健康状态不是OK，首先看看当前管理节点的日志，再看看几个OSD节点的日志，里面都有记录。我当时看了日志才发现，里面记录了什么请求超时什么的，我立马反应过来是因为几台电脑时钟都不一致照成的，然后去配置了NTP服务器。最后记住两个重启当前节点机器上面ceph服务的命令：



# 部署cephfs

## 部署MDS

Cephfs是与ceph集群交互的一种方式，叫做文件系统方式。如果不涉及大量的IO操作，并且要能够实时共享读写的话，这种方式是最佳的。其他的方式还有块模式、对象存储模式。项目上推荐使用cephfs这种，并且以内核方式挂载到客户端，效率为佳。

如果要使用cephfs，我们首先就要部署一个元数据服务器（MDS），而且官方只支持部署一个MDS节点。部署MDS节点也很简单，在管理机器节点上面只用执行下面一个命令就行：



在这里我们使用node1这个主机作为MDS节点。

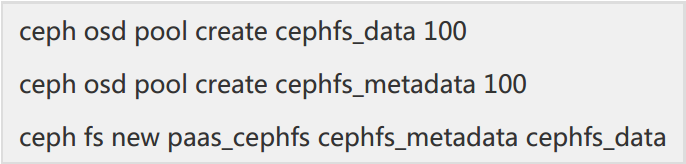
## 部署cephfs

创建文件系统之前，要有几个先决条件：

1. 客户端机器上面要保证安装ceph这个包，可以使用ceph-deploy安装，也可以登录到客户端机器上面使用apt-get安装。
2. 必须要保证集群是在HEALTH\_OK状态下。
3. 使用下面命令将秘钥拷贝到客户机上面：



然后就是要创建存储池与集群的文件系统了，在管理节点上执行下面命令：



其中cephfs\_data、cephfs\_metadata是两个存储池，paas\_cephfs是文件系统

然后登录到客户端机器上面，在/etc/ceph/目录下面找到ceph.client.admin.keyring这个秘钥文件将里面的秘钥复制到单独的一个文件夹，例如/home/admin.secret，大致秘钥是这个样子的：



最后，我们以内核的方式将ceph的文件系统挂载到客户端机器上面，过程命令如下面：

创建一个挂载的目录



最后的path/to/admin.secret换成我们刚刚的/home/admin.secret，里面的ip地址换成我们创建的monitor节点的ip地址就可以，如此一来就完成了以内核方式挂载cephfs。

## 可能会碰到的问题

这个环境，我自己碰到的问题并不是很多，其中说是可能在将秘钥存成一个文件的时候，会挂载不成功的问题我也是没有出现的。如果要是不成功，可以将上面的挂载命令直接以明文的方式给出秘钥：



# 效果评价

虽然是一个学习测试使用的集群，使用的机器也不是很好，不过如今能过正常使用，为团队提供支撑，也算是达到了预想中的效果。

# 推广建议

无

# 参考资料

无。