Ceph集群部署

系统基础环境设定

测试环境说明

测试使用的Ceph存储集群可由一个MON主机及两个以上的OSD机组成，这些主机可以是物理服务器，也可以运行于vmware、virtualbox或kvm等虚拟化平台上的虚拟机，甚至是公有云上的VPS主机。

本测试环境将由stor01、stor02、stor03和ceph-admin四个独立的主机组成，其中stor01、stor02和stor03是为Ceph存储集群节点，它们分别作为MON节点和OSD节点，各自拥有专用于存储数据的磁盘设备/dev/vdb和/dev/vdc，操作系统环境均为CentOS 7.5 1804。而ceph-admin主机是为管理节点，用于部署ceph-deploy。

主机地址	主机名称	主机角色
172.20.0.59	ceph-admin.magedu.com	admin
172.20.0.55	stor01.magedu.com	mon, osd, mgr, mds
172.20.0.56	stor02.magedu.com	mon, osd, mgr
172.20.0.57	stor03.magedu.com	mon, osd, rgw

此外，各主机需要预设的系统环境如下：

- 借助于NTP服务设定各节点时间精确同步；
- 通过DNS完成各节点的主机名称解析，测试环境主机数量较少时也可以使用hosts文件进行；
- 关闭各节点的iptables或firewalld服务，并确保它们被禁止随系统引导过程启动；
- 各节点禁用SELinux；

设定时钟同步

若节点可直接访问互联网，直接启动chronyd系统服务，并设定其随系统引导而启动。

```
~]# systemctl start chronyd.service
~]# systemctl enable chronyd.service
```

不过，建议用户配置使用本地的的时间服务器，在节点数量众多时尤其如此。存在可用的本地时间服务器时，修改节点的/etc/crhony.conf配置文件，并将时间服务器指向相应的主机即可，配置格式如下：

server CHRONY-SERVER-NAME-OR-IP iburst

主机名称解析

出于简化配置步骤的目的，本测试环境使用hosts文件进行各节点名称解析，文件内容如下所示：

172.20.0.55 stor01.magedu.com stor01 mon01 mds01 172.20.0.56 stor02.magedu.com stor02 mon02 mgr01 172.20.0.57 stor03.magedu.com stor03 mon03 mgr02 172.20.0.59 ceph-admin.magedu.com ceph-admin

关闭iptables或firewalld服务

在CentOS7上，iptables或firewalld服务通常只会安装并启动一种，在不确认具体启动状态的前提下，这里通过同时关闭并禁用二者即可简单达到设定目标。

```
~]# systemctl stop firewalld.service
~]# systemctl stop iptables.service
~]# systemctl disable firewalld.service
~]# systemctl disable iptables.service
```

关闭并禁用SELinux

若当前启用了SELinux，则需要编辑/etc/sysconfig/selinux文件,禁用SELinux，并临时设置其当前状态为permissive：

```
~]# sed -i 's@^\(SELINUX=\).*@\1disabled@' /etc/sysconfig/selinux
~]# setenforce 0
```

准备部署Ceph集群

准备yum仓库配置文件

Ceph官方的仓库路径为http://download.ceph.com/，目前主流版本相关的程序包都在提供，包括kraken、luminous和mimic等，它们分别位于rpm-mimic等一类的目录中。直接安装程序包即可生成相关的yum仓库相关的配置文件，程序包位于相关版本的noarch目录下，例如rpm-mimic/el7/noarch/ceph-release-1-1.el7.noarch.rpm是为负责生成适用于部署mimic版本Ceph的yum仓库配置文件，因此直接在线安装此程序包，也能直接生成yum仓库的相关配置。

在ceph-admin节点上，使用如下命令即可安装生成mimic版本相关的yum仓库配置。

```
~]# rpm -ivh https://mirrors.aliyun.com/ceph/rpm-mimic/el7/noarch/ceph-release-1-1.el7.noarch.rpm
```

创建部署Ceph的特定用户账号

部署工具ceph-deploy 必须以普通用户登录到Ceph集群的各目标节点，且此用户需要拥有无密码使用sudo命令的权限，以便在安装软件及生成配置文件的过程中无需中断配置过程。不过，较新版的ceph-deploy也支持用 ”--username“ 选项提供可无密码使用sudo命令的用户名（包括 root ，但不建议这样做）。

另外，使用”ceph-deploy --username {username} “命令时，指定的用户需要能够通过SSH协议自动认证并连接到各Ceph节点，以免ceph-deploy命令在配置中途需要用户输入密码。

在各Ceph各节点创建新用户

首先需要在各节点以管理员的身份创建一个专用于ceph-deploy的特定用户账号，例如cephadm（建议不要使用ceph），并为其设置认证密码（例如magedu）：

```
~]# useradd cephadm
~]# echo "magedu" | passwd --stdin cephadm
```

而后，确保这些节点上新创建的用户cephadm都有无密码运行sudo命令的权限。

```
~]# echo "cephadm ALL = (root) NOPASSWD:ALL" | sudo tee /etc/sudoers.d/cephadm
~]# chmod 0440 /etc/sudoers.d/cephadm
```

配置用户基于密钥的ssh认证

ceph-deploy命令不支持运行中途的密码输入，因此，必须在管理节点（ceph-admin.magedu.com）上生成SSH密钥并将其公钥分发至Ceph集群的各节点上。下面直接以cephadm用户的身份生成SSH密钥对：

```
~]$ ssh-keygen -t rsa -P ""
```

而后即可把公钥拷贝到各Ceph节点：

```
~]$ ssh-copy-id -i .ssh/id_rsa.pub cephadm@stor01.magedu.com
~]$ ssh-copy-id -i .ssh/id_rsa.pub cephadm@stor02.magedu.com
~]$ ssh-copy-id -i .ssh/id_rsa.pub cephadm@stor03.magedu.com
```

另外，为了后续操作之便，建议修改管理节点上cephadm用户的 ~/.ssh/config 文件，设定其访问Ceph集群各节点时默认使用的用户名为，从而避免每次执行ceph-deploy命令时都要指定 使用”--username“选项设置使用的用户名。文件内容示例如下所示：

```
Host stor01
   Hostname stor01.magedu.com
   User cephadm
Host stor02
   Hostname stor02.magedu.com
   User cephadm
Host stor03
   Hostname stor03.magedu.com
   User cephadm
```

在管理节点安装ceph-deploy

Ceph存储集群的部署的过程可通过管理节点使用ceph-deploy全程进行，这里首先在管理节点安装ceph-deploy及其依赖到的程序包：

```
[root@ceph-admin ~]# yum update
[root@ceph-admin ~]# yum install ceph-deploy python-setuptools python2-subprocess32
```

部署RADOS存储集群

初始化RADOS集群

1. 首先在管理节点上以cephadm用户创建集群相关的配置文件目录：

```
~]$ mkdir ceph-cluster
~]$ cd ceph-cluster
```

2. 初始化第一个MON节点，准备创建集群：

初始化第一个MON节点的命令格式为”ceph-deploy new {initial-monitor-node(s)}“，本示例中，stor01即为第一个MON节点名称，其名称必须与节点当前实际使用的主机名称保存一致。运行如下命令即可生成初始配置：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy new stor01
```

3. 编辑生成ceph.conf配置文件，在[global]配置段中设置Ceph集群面向客户端通信时使用的IP地址所在的网络，即公网网络地址：

public network = 172.20.0.0/16

4. 安装Ceph集群

ceph-deploy命令能够以远程的方式连入Ceph集群各节点完成程序包安装等操作，命令格式如下：

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

因此，若要将stor01、stor02和stor03配置为Ceph集群节点，则执行如下命令即可：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy install stor01 stor02 stor03
```

提示：若需要在集群各节点独立安装ceph程序包，其方法如下：

```
~]#  yum install -y https://download.ceph.com/rpm-mimic/el7/noarch/ceph-release-1-
0.el7.noarch.rpm
~]# yum install ceph ceph-radosgw
```

5. 配置初始MON节点，并收集所有密钥：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy mon create-initial
```

6. 把配置文件和admin密钥拷贝Ceph集群各节点，以免得每次执行”ceph“命令行时不得不明确指定MON节点地址和ceph.client.admin.keyring：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy admin stor01 stor02 stor03
```

而后在Ceph集群中需要运行ceph命令的的节点上（或所有节点上）以root用户的身份设定用户cephadm能够读取/etc/ceph/ceph.client.admin.keyring文件：

```
~]$ setfacl -m u:cephadm:r /etc/ceph/ceph.client.admin.keyring
```

7. 配置Manager节点，启动ceph-mgr进程（仅Luminious+版本）：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy mgr create stor01
```

8. 在Ceph集群内的节点上以cephadm用户的身份运行如下命令，测试集群的健康状态：

```
[cephadm@stor01 ~]$ ceph health
HEALTH_OK
```

```
[cephadm@stor01 ~]$ ceph -s
  cluster:
    id:     fc5b806d-3b43-41f1-974a-c07468b9d9ff
    health: HEALTH_OK

  services:
    mon: 1 daemons, quorum stor01
    mgr: stor01(active)
    osd: 0 osds: 0 up, 0 in

  data:
    pools:   0 pools, 0 pgs
    objects: 0  objects, 0 B
    usage:   0 B used, 0 B / 0 B avail
    pgs:
```

向RADOS集群添加OSD

列出并擦净磁盘

“ceph-deploy disk”命令可以检查并列出OSD节点上所有可用的磁盘的相关信息：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy disk list stor01 stor02 stor03
```

而后，在管理节点上使用ceph-deploy命令擦除计划专用于OSD磁盘上的所有分区表和数据以便用于OSD，命令格式为”ceph-deploy disk zap {osd-server-name} {disk-name}“，需要注意的是此步会清除目标设备上的所有数据。下面分别擦净stor01、stor02和stor03上用于OSD的一个磁盘设备vdb：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy disk zap stor01 /dev/vdb
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy disk zap stor02 /dev/vdb
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy disk zap stor03 /dev/vdb
```

提示：若设备上此前有数据，则可能需要在相应节点上以root用户使用“ceph-volume lvm zap --destroy {DEVICE}”命令进行；

添加OSD

早期版本的ceph-deploy命令支持在将添加OSD的过程分为两个步骤：准备OSD和激活OSD，但新版本中，此种操作方式已经被废除，添加OSD的步骤只能由命令”ceph-deploy osd create {node} --data {data-disk}“一次完成，默认使用的存储引擎为bluestore。

如下命令即可分别把stor01、stor02和stor03上的设备vdb添加为OSD：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy osd create stor01 --data /dev/vdb
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy osd create stor02 --data /dev/vdb
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy osd create stor03 --data /dev/vdb
```

而后可使用”ceph-deploy osd list”命令列出指定节点上的OSD：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy osd list stor01 stor02 stor03
```

事实上，管理员也可以使用ceph命令查看OSD的相关信息：

```
~]$ ceph osd stat
3 osds: 3 up, 3 in; epoch: e15
```

或者使用如下命令了解相关的信息：

```
~]$ ceph osd dump
~]$ ceph osd ls
```

从RADOS集群中移除OSD的方法

Ceph集群中的一个OSD通常对应于一个设备，且运行于专用的守护进程。在某OSD设备出现故障，或管理员出于管理之需确实要移除特定的OSD设备时，需要先停止相关的守护进程，而后再进行移除操作。对于Luminous及其之后的版本来说，停止和移除命令的格式分别如下所示：

1. 停用设备：ceph osd out {osd-num}
2. 停止进程：sudo systemctl stop ceph-osd@{osd-num}
3. 移除设备：ceph osd purge {id} --yes-i-really-mean-it

若类似如下的OSD的配置信息存在于ceph.conf配置文件中，管理员在删除OSD之后手动将其删除。

> [osd.1] host = {hostname}

不过，对于Luminous之前的版本来说，管理员需要依次手动执行如下步骤删除OSD设备：

1. 于CRUSH运行图中移除设备：ceph osd crush remove {name}
2. 移除OSD的认证key：ceph auth del osd.{osd-num}
3. 最后移除OSD设备：ceph osd rm {osd-num}

测试上传/下载数据对象

存取数据时，客户端必须首先连接至RADOS集群上某存储池，而后根据对象名称由相关的CRUSH规则完成数据对象寻址。于是，为了测试集群的数据存取功能，这里首先创建一个用于测试的存储池mypool，并设定其PG数量为16个。

```
~]$ ceph osd pool create mypool 16
pool 'mypool' created
```

而后即可将测试文件上传至存储池中，例如下面的“rados put”命令将/etc/issue文件上传至mypool存储池，对象名称依然保留为文件名issue，而“rados ls”命令则可以列出指定存储池中的数据对象。

```
~]$ rados put issue /etc/issue --pool=mypool
~]$ rados ls --pool=mypool
issue
```

而“ceph osd map”命令可以获取到存储池中数据对象的具体位置信息：

```
~]$ ceph osd map mypool issue
osdmap e26 pool 'mypool' (1) object 'issue' -> pg 1.651f88da (1.a) -> up ([2,1,0], p2)
acting ([2,1,0], p2)
```

删除数据对象，“rados rm”命令是较为常用的一种方式：

```
~]$ rados rm issue --pool=mypool
```

删除存储池命令存在数据丢失的风险，Ceph于是默认禁止此类操作。管理员需要在ceph.conf配置文件中启用支持删除存储池的操作后，方可使用类似如下命令删除存储池。

```
~]$ ceph osd pool rm mypool mypool --yes-i-really-really-mean-it
```

扩展Ceph集群

扩展监视器节点

Ceph存储集群需要至少运行一个Ceph Monitor和一个Ceph Manager，生产环境中，为了实现高可用性，Ceph存储集群通常运行多个监视器，以免单监视器整个存储集群崩溃。Ceph使用Paxos算法，该算法需要半数以上的监视器（大于n/2，其中n为总监视器数量）才能形成法定人数。尽管此非必需，但奇数个监视器往往更好。

“ceph-deploy mon add {ceph-nodes}”命令可以一次添加一个监视器节点到集群中。例如，下面的命令可以将集群中的stor02和stor03也运行为监视器节点：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy mon add stor02
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy mon add stor03
```

设置完成后，可以在ceph客户端上查看监视器及法定人数的相关状态：

~]$ ceph quorum_status --format json-pretty

{ "election_epoch": 12, "quorum": [0, 1, 2], "quorum_names": ["stor01", "stor02", "stor03"], "quorum_leader_name": "stor01", "monmap": { "epoch": 3, ... }, "mons": [{ "rank": 0, "name": "stor01", "addr": "172.20.0.55:6789/0", "public_addr": "172.20.0.55:6789/0" }, { "rank": 1, "name": "stor02", "addr": "172.20.0.56:6789/0", "public_addr": "172.20.0.56:6789/0" }, { "rank": 2, "name": "stor03", "addr": "172.20.0.57:6789/0", "public_addr": "172.20.0.57:6789/0" }] }}

扩展Manager节点

Ceph Manager守护进程以“Active/Standby”模式运行，部署其它ceph-mgr守护程序可确保在Active节点或其上的ceph-mgr守护进程故障时，其中的一个Standby实例可以在不中断服务的情况下接管其任务。

“ceph-deploy mgr create {new-manager-nodes}”命令可以一次添加多个Manager节点。下面的命令可以将stor02节点作为备用的Manager运行：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy mgr create stor02
```

添加完成后，“ceph -s”命令的services一段中会输出相关信息：

~]$ ceph -s cluster: id: fc5b806d-3b43-41f1-974a-c07468b9d9ff health: HEALTH_OK

services: mon: 3 daemons, quorum stor01,stor02,stor03 **mgr: stor01(active), standbys: stor02** osd: 3 osds: 3 up, 3 in

.....

Ceph存储集群的访问接口

Ceph块设备接口（RBD）

Ceph块设备，也称为RADOS块设备（简称RBD），是一种基于RADOS存储系统支持超配（thin-provisioned）、可伸缩的条带化数据存储系统，它通过librbd库与OSD进行交互。RBD为KVM等虚拟化技术和云OS（如OpenStack和CloudStack）提供高性能和无限可扩展性的存储后端，这些系统依赖于libvirt和QEMU实用程序与RBD进行集成。

客户端基于librbd库即可将RADOS存储集群用作块设备，不过，用于rbd的存储池需要事先启用rbd功能并进行初始化。例如，下面的命令创建一个名为rbddata的存储池，在启用rbd功能后对其进行初始化：

```
~]$ ceph osd pool create rbddata 64
~]$ ceph osd pool application enable rbddata rbd
~]$ rbd pool init -p rbddata
```

不过，rbd存储池并不能直接用于块设备，而是需要事先在其中按需创建映像（image），并把映像文件作为块设备使用。rbd命令可用于创建、查看及删除块设备相在的映像（image），以及克隆映像、创建快照、将映像回滚到快照和查看快照等管理操作。例如，下面的命令能够创建一个名为img1的映像：

```
~]$ rbd create img1 --size 1024 --pool rbddata
```

映像的相关的信息则可以使用“rbd info”命令获取：

```
~]$ rbd --image img1 --pool rbddata info
rbd image 'img1':
        size 1 GiB in 256 objects
        order 22 (4 MiB objects)
        id: 11616b8b4567
        block_name_prefix: rbd_data.11616b8b4567
        format: 2
        features: layering, exclusive-lock, object-map, fast-diff, deep-flatten
        op_features:
        flags:
        create_timestamp: Tue Dec 11 17:20:23 2018
```

在客户端主机上，用户通过内核级的rbd驱动识别相关的设备，即可对其进行分区、创建文件系统并挂载使用。

启用radosgw接口

RGW并非必须的接口，仅在需要用到与S3和Swift兼容的RESTful接口时才需要部署RGW实例，相关的命令为“ceph-deploy rgw create {gateway-node}”。例如，下面的命令用于把stor03部署为rgw主机：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy rgw create stor03
```

添加完成后，“ceph -s”命令的services一段中会输出相关信息：

~]$ ceph -s cluster: id: fc5b806d-3b43-41f1-974a-c07468b9d9ff health: HEALTH_OK

services: mon: 3 daemons, quorum stor01,stor02,stor03 mgr: stor01(active), standbys: stor02 osd: 3 osds: 3 up, 3 in **rgw: 1 daemon active**

……

默认情况下，RGW实例监听于TCP协议的7480端口7480，需要算定时，可以通过在运行RGW的节点上编辑其主配置文件ceph.conf进行修改，相关参数如下所示：

```
[client]
rgw_frontends = "civetweb port=8080"
```

而后需要重启相关的服务，命令格式为“systemctl restart ceph-radosgw@rgw.{node-name}”，例如重启stor03上的RGW，可以以root用户运行如下命令：

```
~]# systemctl status ceph-radosgw@rgw.stor03
```

RGW会在rados集群上生成包括如下存储池的一系列存储池：

```
~]$ ceph osd pool ls
.rgw.root
default.rgw.control
default.rgw.meta
default.rgw.log
```

RGW提供的是REST接口，客户端通过http与其进行交互，完成数据的增删改查等管理操作。

启用文件系统（CephFS）接口

CephFS需要至少运行一个元数据服务器（MDS）守护进程（ceph-mds），此进程管理与CephFS上存储的文件相关的元数据，并协调对Ceph存储集群的访问。因此，若要使用CephFS接口，需要在存储集群中至少部署一个MDS实例。“ceph-deploy mds create {ceph-node}”命令可以完成此功能，例如，在stor01上启用MDS：

```
[cephadm@ceph-admin ceph-cluster]$ ceph-deploy mds create stor01
```

查看MDS的相关状态可以发现，刚添加的MDS处于Standby模式：

```
~]$ ceph mds stat
, 1 up:standby
```

使用CephFS之前需要事先于集群中创建一个文件系统，并为其分别指定元数据和数据相关的存储池。下面创建一个名为cephfs的文件系统用于测试，它使用cephfs-metadata为元数据存储池，使用cephfs-data为数据存储池：

```
~]$ ceph osd pool create cephfs-metadata 64
~]$ ceph osd pool create cephfs-data 64
~]$ ceph fs new cephfs cephfs-metadata cephfs-data
```

而后即可使用如下命令“ceph fs status ”查看文件系统的相关状态，例如：

```
~]$ ceph fs status cephfs
```

此时，MDS的状态已经发生了改变：

```
 ~]$ ceph mds stat
cephfs-1/1/1 up  {0=stor01=up:active}
```

随后，客户端通过内核中的cephfs文件系统接口即可挂载使用cephfs文件系统，或者通过FUSE接口与文件系统进行交互。