Ceph集群部署

1.1 概述

无论那种方式部署k8s，需要为其准备存储，在选型中本地存储不可跨node，NFS共享存储不好做高可用，因此选型Ceph来为k8s提供存储类。Ceph是一种为优秀的性能、可靠性和可扩展性而设计的统一的、分布式文件系统。Ceph是一个开源的分布式文件系统。因为它还支持块存储、对象存储，所以很自然的被用做云计算框架openstack或cloudstack整个存储后端。当然也可以单独作为存储，例如部署一套集群作为对象存储、SAN存储、NAS存储等。可以作为k8s的存储类，来方便容器持久化存储。

1.2 支持格式

- 对象存储：即radosgw,兼容S3接口。通过rest api上传、下载文件。
- 文件系统：posix接口。可以将ceph集群看做一个共享文件系统挂载到本地。
- 块存储：即rbd。有kernel rbd和librbd两种使用方式。支持快照、克隆。相当于一块硬盘挂到本地，用法和用途和硬盘一样。比如在OpenStack项目里，Ceph的块设备存储可以对接OpenStack的后端存储

1.3 优势

- 统一存储：虽然ceph底层是一个分布式文件系统，但由于在上层开发了支持对象和块的接口
- 高扩展性：扩容方便、容量大。能够管理上千台服务器、EB级的容量。
- 高可靠性：支持多份强一致性副本，EC。副本能够垮主机、机架、机房、数据中心存放。所以安全可靠。存储节点可以自管理、自动修复。无单点故障，容错性强。
- 高性能：因为是多个副本，因此在读写操作时候能够做到高度并行化。理论上，节点越多，整个集群的IOPS和吞吐量越高。另外一点ceph客户端读写数据直接与存储设备(osd) 交互。

1.4 核心组件

- Ceph OSDs:Ceph OSD 守护进程（ Ceph OSD ）的功能是存储数据，处理数据的复制、恢复、回填、再均衡，并通过检查其他OSD 守护进程的心跳来向 Ceph Monitors 提供一些监控信息。当 Ceph 存储集群设定为有2个副本时，至少需要2个 OSD 守护进程，集群才能达到 active+clean 状态（ Ceph 默认有3个副本，但你可以调整副本数）。
- Monitors: Ceph Monitor维护着展示集群状态的各种图表，包括监视器图、 OSD 图、归置组（ PG ）图、和 CRUSH 图。 Ceph 保存着发生在Monitors 、 OSD 和 PG上的每一次状态变更的历史信息（称为 epoch ）。
- MDSs: Ceph 元数据服务器（ MDS ）为 Ceph 文件系统存储元数据（也就是说，Ceph 块设备和 Ceph 对象存储不使用MDS ）。元数据服务器使得 POSIX 文件系统的用户们，可以在不对 Ceph 存储集群造成负担的前提下，执行诸如 ls、find 等基本命令。

二 安装部署

2.1 主机信息

主机名	操作系统	配置	K8S组件	CEPH组件	私网IP	SSH端口	用户名密码
k8s-master	CentOS 7.4 64bit	4C8G + 500G硬盘		admin,osd, mon	172.16.60.2	2001/22	root/uWWKWnjySO7Zocuh
k8s-node01	CentOS 7.4 64bit	4C8G + 500G硬盘		osd, mon	172.16.60.3	2002/22	root/IZ5IReaUBz3QOkLh
k8s-node02	CentOS	4C8G +		osd, mon	172.16.60.	2003/22	root/nUM

	7.4 64bit	500G硬盘			4		Flg9a4zpz DMcE

2.2 磁盘准备

需要在三台主机创建磁盘,并挂载到主机的/var/local/osd{0,1,2}

```
[root@master ~]# mkfs.xfs /dev/vdc
[root@master ~]# mkdir -p /var/local/osd0
[root@master ~]# mount /dev/vdc /var/local/osd0/

[root@node01 ~]# mkfs.xfs /dev/vdc
[root@node01 ~]# mkdir -p /var/local/osd1
[root@node01 ~]# mount /dev/vdc /var/local/osd1/

[root@node02 ~]# mkfs.xfs /dev/vdc
[root@node02 ~]# mkdir -p /var/local/osd2
[root@node02 ~]# mount /dev/vdc /var/local/osd2/

将磁盘添加进入fstab中，确保开机自动挂载


```

2.3 配置各主机hosts文件

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
172.16.60.2 k8s-master
172.16.60.3 k8s-node01
172.16.60.4 k8s-node02

```

2.4 管理节点ssh免密钥登录node1/node2

```
[root@master ~]# ssh-keygen -t rsa
[root@master ~]# ssh-copy-id -i /root/.ssh/id_rsa.pub root@node01
[root@master ~]# ssh-copy-id -i /root/.ssh/id_rsa.pub root@node02

```

2.5 master节点安装ceph-deploy工具

```
# 各节点均更新ceph的yum源
vim /etc/yum.repos.d/ceph.repo

[ceph]
name=ceph
baseurl=http://mirrors.aliyun.com/ceph/rpm-jewel/el7/x86_64/
gpgcheck=0
priority =1
[ceph-noarch]
name=cephnoarch
baseurl=http://mirrors.aliyun.com/ceph/rpm-jewel/el7/noarch/
gpgcheck=0
priority =1
[ceph-source]
name=Ceph source packages
baseurl=http://mirrors.aliyun.com/ceph/rpm-jewel/el7/SRPMS
gpgcheck=0
priority=1

# 安装ceph-deploy工具
yum clean all && yum makecache
yum -y install ceph-deploy

```

2.6 创建monitor服务

创建monitor服务,指定master节点的hostname

```
[root@master ~]# mkdir /etc/ceph && cd /etc/ceph
[root@master ceph]# ceph-deploy new k8s-master
[root@master ceph]# ll
total 12
-rw-r--r-- 1 root root  195 Sep  3 10:56 ceph.conf
-rw-r--r-- 1 root root 2915 Sep  3 10:56 ceph-deploy-ceph.log
-rw------- 1 root root   73 Sep  3 10:56 ceph.mon.keyring


[root@master ceph]# cat ceph.conf
[global]
```

```
fsid = 5b9eb8d2-1c12-4f6d-ae9c-85078795794b
mon_initial_members = master
mon_host = 172.16.60.2
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
osd_pool_default_size = 2

配置文件的默认副本数从3改成2，这样只有两个osd也能达到active+clean状态，把下面这行加入到[globa

```

2.7 所有节点安装ceph

```
# 各节点安装软件包
yum -y install yum-plugin-priorities epel-release
# master节点利用ceph-deply 部署ceph

[root@master ceph]# ceph-deploy install k8s-master k8s-node01 k8s-node02

[root@master ceph]# ceph --version
ceph version 10.2.11 (e4b061b47f07f583c92a050d9e84b1813a35671e)

yum install -y yum-utils && yum-config-manager --add-repo https://dl.fedoraprojec
```

2.8 部署相关服务

```
# 安装ceph monitor
[root@master ceph]# ceph-deploy mon create k8s-master

# 收集节点的keyring文件
[root@master ceph]# ceph-deploy  gatherkeys k8s-master

# 创建osd
[root@master ceph]# ceph-deploy osd prepare k8s-master:/var/local/osd0 k8s-node01

# 权限修改
[root@master ceph]# chmod 777 -R /var/local/osd{0..2}

[root@master ceph]# chmod 777 -R /var/local/osd{0..2}/*
```

```
13
14 # 激活osd
15 [root@master ceph]# ceph-deploy osd activate k8s-master:/var/local/osd0 k8s-node0
16
17 # 查看状态
18 [root@master ceph]# ceph-deploy osd list k8s-master k8s-node01 k8s-node02
19
```

2.9 统一配置

用ceph-deploy把配置文件和admin密钥拷贝到所有节点，这样每次执行Ceph命令行时就无需指定monitor地址和ceph.client.admin.keyring了

```
1  [root@master ceph]# ceph-deploy admin k8s-master k8s-node01 k8s-node02
2
3  # 各节点修改ceph.client.admin.keyring权限:
4  [root@master ceph]# chmod +r /etc/ceph/ceph.client.admin.keyring
5
6
7  # 查看状态
8  [root@master ceph]# ceph health
9  HEALTH_OK
10 [root@master ceph]# ceph -s
11     cluster 5b9eb8d2-1c12-4f6d-ae9c-85078795794b
12      health HEALTH_OK
13      monmap e1: 1 mons at {master=172.16.60.2:6789/0}
14             election epoch 3, quorum 0 master
15      osdmap e15: 3 osds: 3 up, 3 in
16             flags sortbitwise,require_jewel_osds
17       pgmap v27: 64 pgs, 1 pools, 0 bytes data, 0 objects
18             15681 MB used, 1483 GB / 1499 GB avail
19                   64 active+clean
20
21
```

2.10 部署MDS服务

我们在node01/node02上安装部署MDS服务

```
1 [root@master ceph]# ceph-deploy mds create k8s-node01 k8s-node02
```

```
# 查看状态
[root@master ceph]# ceph mds stat
e3:, 2 up:standby
[root@master ~]# ceph mon stat
e1: 1 mons at {master=172.16.60.2:6789/0}, election epoch 4, quorum 0 master

# 查看服务
[root@master ceph]# systemctl list-unit-files |grep ceph
ceph-create-keys@.service                     static
ceph-disk@.service                            static
ceph-mds@.service                             disabled
ceph-mon@.service                             enabled
ceph-osd@.service                             enabled
ceph-radosgw@.service                         disabled
ceph-mds.target                               enabled
ceph-mon.target                               enabled
ceph-osd.target                               enabled
ceph-radosgw.target                           enabled
ceph.target                                   enabled

```

至此，基本上完成了ceph存储集群的搭建。

三 创建ceph文件系统

3.1 创建文件系统

关于创建存储池确定 pg_num 取值是强制性的，因为不能自动计算。下面是几个常用的值：

- 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
- OSD 数量大于 50 时，你得理解权衡方法、以及如何自己计算 pg_num 取值
- 自己计算 pg_num 取值时可借助 pgcalc 工具
- 随着 OSD 数量的增加，正确的 pg_num 取值变得更加重要，因为它显著地影响着集群的行为、以及出错时的数据持久性（即灾难性事件导致数据丢失的概率）。

```
[root@master ceph]# ceph osd pool create cephfs_data <pg_num>
[root@master ceph]# ceph osd pool create cephfs_metadata <pg_num>

[root@master ~]# ceph osd pool ls
rbd
```

```
[root@master ~]#  ceph osd pool create kube 128
pool 'kube' created
[root@master ~]# ceph osd pool ls
rbd
kube

# 查看证书
[root@master ~]# ceph auth list
installed auth entries:

mds.node01
        key: AQB56m1dE42rOBAA0yRhsmQb3QMEaTsQ71jHdg==
        caps: [mds] allow
        caps: [mon] allow profile mds
        caps: [osd] allow rwx
mds.node02
        key: AQB66m1dWuhWKhAAtbiZN7amGcjUh6Rj/HNFkg==
        caps: [mds] allow
        caps: [mon] allow profile mds
        caps: [osd] allow rwx
osd.0
        key: AQA46W1daFx3IxAAE1esQW+t1fWJDfEQd+167w==
        caps: [mon] allow profile osd
        caps: [osd] allow *
osd.1
        key: AQBA6W1daJG9IxAAQwETgrVc3awkEZejDSaaow==
        caps: [mon] allow profile osd
        caps: [osd] allow *
osd.2
        key: AQBI6W1dot4/GxAAle3Ii3/D38RmwNC4yTCoPg==
        caps: [mon] allow profile osd
        caps: [osd] allow *
client.admin
        key: AQBu4W1d90dZKxAAH/kta03cP5znnCcWeOngzQ==
        caps: [mds] allow *
        caps: [mon] allow *
        caps: [osd] allow *
client.bootstrap-mds
        key: AQBv4W1djJ1uHhAACzBcXjVoZFgLg3lN+KEv8Q==
```

```
        caps: [mon] allow profile bootstrap-mds
client.bootstrap-mgr
        key: AQCS4W1dna9COBAAiWPu7uk3ItJxisVIwn2duA==
        caps: [mon] allow profile bootstrap-mgr
client.bootstrap-osd
        key: AQBu4W1dxappOhAA5FanGhQhAOUlizqa5uMG3A==
        caps: [mon] allow profile bootstrap-osd
client.bootstrap-rgw
        key: AQBv4W1dpwvsDhAAyp58v08XttJWzLoHWVHZow==
        caps: [mon] allow profile bootstrap-rgw

```

3.2 创建客户端密钥

```
# 创建keyring
[root@master ~]# ceph auth get-or-create client.kube mon 'allow r' osd 'allow rwx
[root@master ~]# ceph auth list

# 将密钥拷贝到node1和node2
[root@master ceph]# scp ceph.client.kube.keyring root@k8s-node01:/etc/ceph/
[root@master ceph]# scp ceph.client.kube.keyring root@k8s-node02:/etc/ceph/
```

四 卸载

```
清理机器上的ceph相关配置:
停止所有进程:  stop ceph-all
卸载所有ceph程序: ceph-deploy uninstall [{ceph-node}]
删除ceph相关的安装包: ceph-deploy purge {ceph-node} [{ceph-data}]
删除ceph相关的配置: ceph-deploy purgedata {ceph-node} [{ceph-data}]
删除key: ceph-deploy forgetkeys

卸载ceph-deploy管理: yum -y remove ceph-deploy

```

参考链接

- ceph官方文档
- ceph中文开源社区
- CentOS 7部署 Ceph分布式存储架构