

Ceph 分布式存储实践指南


版本 1.10

日期 2017-03-15

版权所有©上海云轴信息科技有限公司 2017。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标说明

和其他云轴商标均为上海云轴信息科技有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受上海云轴公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，上海云轴公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

上海云轴信息科技有限公司

地址：上海市闵行区紫竹科学园东川路 555 号 6 号楼 邮编：200241

网址：<http://www.zstack.io/>

客户服务邮箱：support@zstack.io

客户服务电话：400-962-2212

概述

本文档针对 Ceph 分布式存储实践过程进行说明。

读者对象

本文档适合以下工程师阅读：

- 运维工程师
- 测试工程师
- 存储工程师
- 预研工程师

术语定义

术语	概念
管理节点	安装ZStack系统的物理主机，提供UI管理、云系统部署功能
计算节点	也称之为物理主机，为云主机实例提供计算、内存、网络、存储的物理主机。运行KVM虚拟化的物理主机，简称“KVM主机”
云主机	ZStack 特制虚拟机，即运行在物理机上的虚拟机实例，具有独立的IP地址，可以安装部署应用服务
镜像	云主机所使用的镜像模板文件，包含了云主机的操作系统，也可以定制安装相应的软件
镜像服务器	也称之为备份存储服务器，存储云主机镜像文件的物理主机。虽然可以和管理节点或其他计算节点共享同一台物理服务器，但不建议在生产环境中这么部署
云盘	云主机的数据盘，给云主机提供额外的存储空间，一块云盘在同一时刻只能挂载到一个云主机。一个云主机最多可以挂载24块云盘
计算规格	启动云主机涉及到的CPU数量、内存大小、网络设置等规格定义
云盘规格	创建云盘容量大小的规格定义
安全组	针对云主机进行第三层网络的防火墙控制，对IP地址、网络包类型

	或网络包流向等可以设置不同的安全规则
L2NoVlanNetwork	物理主机的网络连接不采用Vlan设置
L2VlanNetwork	物理主机节点的网络连接采用Vlan设置，Vlan需要在交换机端提前进行设置
二层网络	计算节点的物理网卡设备名称，例如eth0
三层网络	云主机需要使用的网络配置，包括IP地址范围，网关，DNS等
管理网络	ZStack管理物理机和其他云资源的网络
公有网络	云主机连接和使用的网络

中英文术语对照

管理节点	Management Node
物理机	Host
云主机	VM Instance
镜像服务器（备份服务器）	Backup Storage
主存储	Primary Storage
镜像	Image
云盘	Volume
集群	Cluster
区域	Zone
二层网络	L2 Network
三层网络	L3 Network
安全组	Security Group
计算规格	Instance Offering
云盘规格	Disk Offering
扁平网络模式	Flat Network Mode
本地存储	Local Storage

修改记录

修改记录积累了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 1.10（2017-03-15）

第一次正式发布。

目录

第一章 部署实践.....	7
1.1 基础环境准备.....	7
1.2 磁盘初始化.....	10
1.2 安装 Ceph.....	13
第二章 运维管理.....	18
2.1 删除 OSD.....	18
2.2 增加 OSD.....	18
2.3 删除 MON.....	19
2.4 增加 MON.....	19

第一章 部署实践

Ceph 是开源的分布式存储解决方案，支持对象存储、块存储和文件存储访问类型。

2014 年 4 月，Redhat 收购 Inktank。Ceph 成为 Redhat 的开源项目，并提供商业支持服务。Ceph 开源项目发布遵循 GPL 和 LGPL 协议。

1.1 基础环境准备

目前，ZStack Community DVD ISO ([下载](#)) 集成 Ceph 社区版 Hammer 0.94.9。本文假设以下实施环境：

服务器	配置	IP 地址	数据磁盘
ceph-1	4 核, 内存 4GB	172.20.12.154	固态硬盘: /dev/vdb 机械硬盘: /dev/vdc, /dev/vdd
ceph-2	4 核, 内存 4GB	172.20.12.99	固态硬盘: /dev/vdb 机械硬盘: /dev/vdc, /dev/vdd
ceph-3	4 核, 内存 4GB	172.20.12.21	固态硬盘: /dev/vdb 机械硬盘: /dev/vdc, /dev/vdd

其中，每个节点配置 1 块固态硬盘 40GB 存储日志，配置 2 块机械硬盘 200G 存储数据。

通过 ZStack 定制 ISO 安装 3 个节点操作系统。安装后配置服务器的主机名和网络。配置 3 个节点的主机名：

```
[root@localhost ~]# hostnamectl set-hostname ceph-1

[root@localhost ~]# hostnamectl set-hostname ceph-2

[root@localhost ~]# hostnamectl set-hostname ceph-3
```

网络的配置方式如下：

- 配置网桥的命令：

```
zs-network-setting -b eth0 172.20.12.21 255.255.0.0 172.20.0.1
                  接口   IP 地址      掩码      网关
```

配置 3 个节点的网络信息：

```
[root@ceph-1 ~]# zs-network-setting -b eth0 172.20.12.154 255.255.0.0 172.20.0.1

[root@ceph-2 ~]# zs-network-setting -b eth0 172.20.12.99 255.255.0.0 172.20.0.1

[root@ceph-3 ~]# zs-network-setting -b eth0 172.20.12.21 255.255.0.0 172.20.0.1
```

在 ceph-1 节点设置主机解析：

```
[root@ceph-1 ~]# vim /etc/hosts
...
172.20.12.154    ceph-1
172.20.12.99    ceph-2
172.20.12.21    ceph-3
...
```

在 ceph-1 节点创建 ssh 密钥配对，并配置 ceph-1、ceph-2 和 ceph-3 无密码登陆：

```
[root@ceph-1 ~]# ssh-keygen -t rsa    # 执行后保持默认设定，直接回车

[root@ceph-1 ~]# ssh-copy-id ceph-1    # 执行后提示输入 ceph-1 的 root 密码

[root@ceph-1 ~]# ssh-copy-id ceph-2    # 执行后提示输入 ceph-2 的 root 密码

[root@ceph-1 ~]# ssh-copy-id ceph-3    # 执行后提示输入 ceph-3 的 root 密码
```

无密码配置完成后，传输/etc/hosts 文件：

```
[root@ceph-1 ~]# scp /etc/hosts ceph-2:/etc/

[root@ceph-1 ~]# scp /etc/hosts ceph-3:/etc/
```

配置 ceph-1、ceph-2 和 ceph-3 防火墙，允许互相访问：

```
[root@ceph-1 ~]# iptables -I INPUT -s 172.20.0.0/16 -j ACCEPT && service iptables save
[root@ceph-2 ~]# iptables -I INPUT -s 172.20.0.0/16 -j ACCEPT && service iptables save
[root@ceph-3 ~]# iptables -I INPUT -s 172.20.0.0/16 -j ACCEPT && service iptables save
```

配置 ceph-1、ceph-2 和 ceph-3 时间同步：

```
[root@ceph-1 ~]# yum --disablerepo=* --enablerepo=zstack-local,ceph-hammer install ntp
[root@ceph-1 ~]# systemctl restart ntpd
[root@ceph-1 ~]# systemctl enable ntpd.service
```



```
[root@ceph-2 ~]# yum --disablerepo=* --enablerepo=zstack-local,ceph-hammer install ntp
[root@ceph-2 ~]# systemctl restart ntpd
[root@ceph-2 ~]# systemctl enable ntpd.service
```

```
[root@ceph-3 ~]# yum --disablerepo=* --enablerepo=zstack-local,ceph-hammer install ntp
[root@ceph-3 ~]# systemctl restart ntpd
[root@ceph-3 ~]# systemctl enable ntpd.service
```

至此，3 个节点的主机名与网络配置完成。

1.2 磁盘初始化

根据上一节服务器环境描述，每个节点配置 1 块固态硬盘和 2 块机械硬盘，所以对固态硬盘执行分区，分 2 个日志分区。操作如下：

```
[root@ceph-1 ~]# parted /dev/vdb mklabel gpt
[root@ceph-1 ~]# parted /dev/vdb mkpart journal-1 10% 40%
[root@ceph-1 ~]# parted /dev/vdb mkpart journal-2 60% 90%

[root@ceph-2 ~]# parted /dev/vdb mklabel gpt
[root@ceph-2 ~]# parted /dev/vdb mkpart journal-1 10% 40%
[root@ceph-2 ~]# parted /dev/vdb mkpart journal-2 60% 90%

[root@ceph-3 ~]# parted /dev/vdb mklabel gpt
[root@ceph-3 ~]# parted /dev/vdb mkpart journal-1 10% 40%
[root@ceph-3 ~]# parted /dev/vdb mkpart journal-2 60% 90%
```

对 ceph-1 节点的机械硬盘操作格式化操作如下：

```
[root@ceph-1 ~]# mkfs.xfs -f -i size=512 -l size=128m,lazy-count=1 /dev/vdc
[root@ceph-1 ~]# mkfs.xfs -f -i size=512 -l size=128m,lazy-count=1 /dev/vdd
[root@ceph-1 ~]# mkdir -p /data/disk1/
[root@ceph-1 ~]# mkdir -p /data/disk2/

# 查看机械硬盘的 UUID
[root@ceph-1 ~]# ll /dev/disk/by-uuid/
total 0
lrwxrwxrwx 1 root root 10 May  5 13:10 7ea42029-e99a-4d8a-a837-d27c647ff74e -> ../../vda2
lrwxrwxrwx 1 root root 10 May  5 13:10 84d27c03-98b6-4fc4-8126-eac83015d786 -> ../../vda1
lrwxrwxrwx 1 root root  9 May  5 13:56 9bdfef51-95c2-4018-93dc-69256ab789f3 -> ../../vdd
lrwxrwxrwx 1 root root  9 May  5 13:55 da461bde-2aef-4f49-8634-89564317559e -> ../../vdc

# 通过 UUID 挂载数据目录，/dev/vdc 挂载/data/disk1，依次类推
[root@ceph-1 ~]# mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/da461bde-2aef-4f49-8634-89564317559e /data/disk1
[root@ceph-1 ~]# mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/9bdfef51-95c2-4018-93dc-69256ab789f3 /data/disk2
```

对 ceph-2 节点的机械硬盘操作格式化操作如下：

```
[root@ceph-2 ~]# mkfs.xfs -f -i size=512 -l size=128m,lazy-count=1 /dev/vdc
[root@ceph-2 ~]# mkfs.xfs -f -i size=512 -l size=128m,lazy-count=1 /dev/vdd
[root@ceph-2 ~]# mkdir -p /data/disk1/
[root@ceph-2 ~]# mkdir -p /data/disk2/

# 查看机械硬盘的 UUID
[root@ceph-2 ~]# ll /dev/disk/by-uuid/
total 0
lrwxrwxrwx 1 root root 9 May 5 14:03 62996e0d-0f7f-4e10-99fa-f87686cc0c05 -> ../../vdc
lrwxrwxrwx 1 root root 10 May 5 13:10 7ea42029-e99a-4d8a-a837-d27c647ff74e -> ../../vda2
lrwxrwxrwx 1 root root 10 May 5 13:10 84d27c03-98b6-4fc4-8126-eac83015d786 -> ../../vda1
lrwxrwxrwx 1 root root 9 May 5 14:03 ba29e5ea-81af-4ae9-859a-cdeae84ba466 -> ../../vdd

# 通过 UUID 挂载数据目录, /dev/vdc 挂载/data/disk1, 依次类推
[root@ceph-2 ~]# mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/62996e0d-0f7f-4e10-99fa-f87686cc0c05 /data/disk1
[root@ceph-2 ~]# mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/ba29e5ea-81af-4ae9-859a-cdeae84ba466 /data/disk2
```

对 ceph-3 节点的机械硬盘操作格式化操作如下：

```
[root@ceph-3 ~]# mkfs.xfs -f -i size=512 -l size=128m,lazy-count=1 /dev/vdc
[root@ceph-3 ~]# mkfs.xfs -f -i size=512 -l size=128m,lazy-count=1 /dev/vdd
[root@ceph-3 ~]# mkdir -p /data/disk1/
[root@ceph-3 ~]# mkdir -p /data/disk2/

# 查看机械硬盘的 UUID
[root@ceph-3 ~]# ll /dev/disk/by-uuid/
total 0
lrwxrwxrwx 1 root root 9 May 5 14:17 1efee887-ed2f-4c3e-8e77-717f15b19662 -> ../../vdd
lrwxrwxrwx 1 root root 10 May 5 13:10 7ea42029-e99a-4d8a-a837-d27c647ff74e -> ../../vda2
lrwxrwxrwx 1 root root 10 May 5 13:10 84d27c03-98b6-4fc4-8126-eac83015d786 -> ../../vda1
lrwxrwxrwx 1 root root 9 May 5 14:17 cb07fe55-afba-4ce6-aae1-a65c596d8899 -> ../../vdc

# 通过 UUID 挂载数据目录, /dev/vdc 挂载/data/disk1, 依次类推
[root@ceph-3 ~]# mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/cb07fe55-afba-4ce6-aae1-a65c596d8899 /data/disk1
[root@ceph-3 ~]# mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/1efee887-ed2f-4c3e-8e77-717f15b19662 /data/disk2
```

上述关于机械硬盘通过手动挂载, 服务器重启后将会失去, 建议放到/etc/rc.local：

```
[root@ceph-1 ~]# chmod +x /etc/rc.local
[root@ceph-1 ~]# vim /etc/rc.local
...
mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/da461bde-2aef-4f49-8634-89564317559e /data/disk1
mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/9bdfef51-95c2-4018-93dc-69256ab789f3 /data/disk2
...

[root@ceph-2 ~]# chmod +x /etc/rc.local
[root@ceph-2 ~]# vim /etc/rc.local
...
mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/62996e0d-0f7f-4e10-99fa-f87686cc0c05 /data/disk1
mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/ba29e5ea-81af-4ae9-859a-cdeae84ba466 /data/disk2
...

[root@ceph-3 ~]# chmod +x /etc/rc.local
[root@ceph-3 ~]# vim /etc/rc.local
...
mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/cb07fe55-afba-4ce6-aae1-a65c596d8899 /data/disk1
mount -t xfs -o noatime,nodiratime,nobarrier \
/dev/disk/by-uuid/1efee887-ed2f-4c3e-8e77-717f15b19662 /data/disk2
...
```

至此，3 个节点的磁盘初始化完成。

1.2 安装 Ceph

在 ceph-1、ceph-2 和 ceph-3 节点安装 Ceph Hammer 0.94.9 :

```
[root@ceph-1 ~]# yum --disablerepo=* --enablerepo=zstack-local,ceph-hammer install \
ceph ceph-deploy

[root@ceph-2 ~]# yum --disablerepo=* --enablerepo=zstack-local,ceph-hammer install ceph

[root@ceph-3 ~]# yum --disablerepo=* --enablerepo=zstack-local,ceph-hammer install ceph
```

在 ceph-1 节点安装 ceph-deploy，用于自动化部署。在 ceph-1 创建 Ceph 集群：

```
[root@ceph-1 ~]# mkdir ceph-config
[root@ceph-1 ~]# cd ceph-config
[root@ceph-1 ceph-config]# ceph-deploy new ceph-1 ceph-2 ceph-3

#修改 ceph.conf 文件
[root@ceph-1 ceph-config]# vim ceph.conf
...
[global]
fsid = 392106b4-e3ab-4f51-ac0e-a4e26c8abd82
mon_initial_members = ceph-1, ceph-2, ceph-3
mon_host = 172.20.12.154,172.20.12.99,172.20.12.21
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
filestore_xattr_use_omap = true

osd_pool_default_size = 3
osd_pool_default_min_size = 2
osd_pool_default_pg_num = 128
osd_pool_default_pgp_num = 128

osd_max_backfills = 1
osd_recovery_max_active = 1
osd crush update on start = 0

debug_ms = 0
debug_osd = 0

osd_recovery_max_single_start = 1
filestore_max_sync_interval = 15
filestore_min_sync_interval = 10
filestore_queue_max_ops = 65536
```

```
filestore_queue_max_bytes = 536870912
filestore_queue_committing_max_bytes = 536870912
filestore_queue_committing_max_ops = 65536

filestore_wbthrottle_xfs_bytes_start_flusher = 419430400
filestore_wbthrottle_xfs_bytes_hard_limit = 4194304000
filestore_wbthrottle_xfs_ios_start_flusher = 5000
filestore_wbthrottle_xfs_ios_hard_limit = 50000
filestore_wbthrottle_xfs_inodes_start_flusher = 5000
filestore_wbthrottle_xfs_inodes_hard_limit = 50000

journal_max_write_bytes = 1073714824
journal_max_write_entries = 5000
journal_queue_max_ops = 65536
journal_queue_max_bytes = 536870912

osd_client_message_cap = 65536
osd_client_message_size_cap = 524288000
ms_dispatch_throttle_bytes = 536870912

filestore_fd_cache_size = 4096

osd_op_threads = 10
osd_disk_threads = 2
filestore_op_threads = 6

osd_client_op_priority = 100
osd_recovery_op_priority = 5

rbd_default_format = 2
```

修改 ceph.conf 后保存，执行以下操作：

```
[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf mon create ceph-1 \
ceph-2 ceph-3

[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf config push ceph-1 \
ceph-2 ceph-3

[root@ceph-1 ceph-config]# ceph-deploy gatherkeys ceph-1 ceph-2 ceph-3

[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf osd create \
ceph-1:/data/disk1:/dev/disk/by-partlabel/journal-1 \
ceph-1:/data/disk2:/dev/disk/by-partlabel/journal-2 \
ceph-2:/data/disk1:/dev/disk/by-partlabel/journal-1 \
ceph-2:/data/disk2:/dev/disk/by-partlabel/journal-2 \
ceph-3:/data/disk1:/dev/disk/by-partlabel/journal-1 \
ceph-3:/data/disk2:/dev/disk/by-partlabel/journal-2

[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf osd activate \
ceph-1:/data/disk1:/dev/disk/by-partlabel/journal-1 \
ceph-1:/data/disk2:/dev/disk/by-partlabel/journal-2 \
ceph-2:/data/disk1:/dev/disk/by-partlabel/journal-1 \
ceph-2:/data/disk2:/dev/disk/by-partlabel/journal-2 \
ceph-3:/data/disk1:/dev/disk/by-partlabel/journal-1 \
ceph-3:/data/disk2:/dev/disk/by-partlabel/journal-2

[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf config push ceph-1 \
ceph-2 ceph-3
```

Ceph 集群初始化结束后，看查看当前运行状态：

```
[root@ceph-1 ceph-config]# ceph -s
cluster 392106b4-e3ab-4f51-ac0e-a4e26c8abd82
health HEALTH_WARN
    64 pgs stuck inactive
    64 pgs stuck unclear
    too few PGs per OSD (10 < min 30)
monmap e1: 3 mons at {ceph-1=172.20.12.154:6789/0,ceph-
2=172.20.12.99:6789/0,ceph-3=172.20.12.21:6789/0}
election epoch 4, quorum 0,1,2 ceph-3,ceph-2,ceph-1
osdmap e13: 6 osds: 6 up, 6 in
pgmap v25: 64 pgs, 1 pools, 0 bytes data, 0 objects
    195 MB used, 1199 GB / 1199 GB avail
    64 creating
```

```
[root@ceph-1 ceph-config]# ceph osd tree
```

ID	WEIGHT	TYPE	NAME	UP/DOWN	REWEIGHT	PRIMARY-AFFINITY
-1	0	root	default			
0	0	osd.0		up	1.00000	1.00000
1	0	osd.1		up	1.00000	1.00000
2	0	osd.2		up	1.00000	1.00000
3	0	osd.3		up	1.00000	1.00000
4	0	osd.4		up	1.00000	1.00000
5	0	osd.5		up	1.00000	1.00000

查看 OSD ID 号：

```
# 查看 OSD ID
[root@ceph-1 ~]# cat /data/disk[1-2]/whoami
0
1
[root@ceph-2 ~]# cat /data/disk[1-2]/whoami
2
3
[root@ceph-3 ~]# cat /data/disk[1-2]/whoami
4
5
```

根据 OSD ID 号，建立 CRUSH 结构树：

```
[root@ceph-1 ~]# ceph osd crush add-bucket ceph-1 host
[root@ceph-1 ~]# ceph osd crush add-bucket ceph-2 host
[root@ceph-1 ~]# ceph osd crush add-bucket ceph-3 host

[root@ceph-1 ~]# ceph osd crush move ceph-1 root=default
[root@ceph-1 ~]# ceph osd crush move ceph-2 root=default
[root@ceph-1 ~]# ceph osd crush move ceph-3 root=default

[root@ceph-1 ~]# ceph osd crush create-or-move osd.0 0.2 root=default host=ceph-1
[root@ceph-1 ~]# ceph osd crush create-or-move osd.1 0.2 root=default host=ceph-1

[root@ceph-1 ~]# ceph osd crush create-or-move osd.2 0.2 root=default host=ceph-2
[root@ceph-1 ~]# ceph osd crush create-or-move osd.3 0.2 root=default host=ceph-2

[root@ceph-1 ~]# ceph osd crush create-or-move osd.4 0.2 root=default host=ceph-3
[root@ceph-1 ~]# ceph osd crush create-or-move osd.5 0.2 root=default host=ceph-3
```

创建后，CRUSH 结构树如下：


```

[root@ceph-1 ~]# ceph osd tree
ID WEIGHT  TYPE NAME          UP/DOWN REWEIGHT PRIMARY-AFFINITY
-1 1.19998 root default
-2 0.39999  host ceph-1
  0 0.20000  osd.0      up  1.00000      1.00000
  1 0.20000  osd.1      up  1.00000      1.00000
-3 0.39999  host ceph-2
  2 0.20000  osd.2      up  1.00000      1.00000
  3 0.20000  osd.3      up  1.00000      1.00000
-4 0.39999  host ceph-3
  4 0.20000  osd.4      up  1.00000      1.00000
  5 0.20000  osd.5      up  1.00000      1.00000

# 查看 Ceph 集群状态
[root@ceph-1 ~]# ceph -s
  cluster 392106b4-e3ab-4f51-ac0e-a4e26c8abd82
  health HEALTH_OK
  monmap e1: 3 mons at {ceph-1=172.20.12.154:6789/0,ceph-2=172.20.12.99:6789/0,ceph-3=172.20.12.21:6789/0}
    election epoch 4, quorum 0,1,2 ceph-3,ceph-2,ceph-1
  osdmap e29: 6 osds: 6 up, 6 in
  pgmap v53: 64 pgs, 1 pools, 0 bytes data, 0 objects
    197 MB used, 1199 GB / 1199 GB avail
    64 active+clean

```

至此，Ceph 集群初始化结束。通过 ZStack 企业版初始化界面，在主存储和备份存储步骤中，添加 Ceph 集群 MON 节点。

第二章 运维管理

2.1 删除 OSD

例如，需要移除故障的数据盘 OSD.5，执行以下步骤：

```
[root@ceph-1 ~]# ceph osd out 5

# 登陆到 OSD.5 的服务器，关闭 OSD.5 服务
[root@ceph-3 ~]# service ceph stop osd.5

[root@ceph-1 ~]# ceph osd crush remove osd.5
[root@ceph-1 ~]# ceph auth del osd.5
[root@ceph-1 ~]# ceph osd rm 5

# 进入 ceph-config 目录
[root@ceph-1 ~]# cd ceph-config
[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf config push ceph-1 \
ceph-2 ceph-3
```

2.2 增加 OSD

例如，需要增加 ceph-3 节点的/data/disk2/（必须为纯净目录），执行以下步骤：

```
# 进入 ceph-config 目录
[root@ceph-1 ~]# cd ceph-config
[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf osd create \
ceph-3:/data/disk2:/dev/disk/by-partlabel/journal-2

[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf osd activate \
ceph-3:/data/disk2:/dev/disk/by-partlabel/journal-2

[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf config push ceph-1 \
ceph-2 ceph-3
```

2.3 删除 MON

例如，需要删除 ceph-3 节点的 MON 服务：

```
# 登陆 ceph-3 节点，停止 MON 服务
[root@ceph-3 ~]# service ceph stop mon.ceph-3
[root@ceph-3 ~]# ceph mon remove ceph-3

# 修改 ceph.conf 文件，删除 ceph-3 的 MON 信息
[root@ceph-1 ~]# cd ceph-config
[root@ceph-1 ceph-config]# vim ceph.conf
...
mon_initial_members = ceph-1, ceph-2
mon_host = 172.20.12.154,172.20.12.99
...

# 推送全局信息
[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf config push ceph-1 ceph-2
```

2.4 增加 MON

例如，增加 ceph-3 节点作为 MON 服务：

```
# 查看当前 MON 状态
[root@ceph-1 ~]# ceph mon dump

# 添加 MON 节点
[root@ceph-3 ~]# ceph mon add ceph-3 172.20.12.21:6789

# 增加 MON 信息
[root@ceph-1 ~]# cd ceph-config
[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf mon create ceph-3

# 修改 ceph.conf 文件
[root@ceph-1 ceph-config]# vim ceph.conf
...
mon_initial_members = ceph-1, ceph-2, ceph-3
mon_host = 172.20.12.154,172.20.12.99,172.20.12.21
...
```

```
# 推送全局信息
```

```
[root@ceph-1 ceph-config]# ceph-deploy --overwrite-conf config push ceph-1 \  
ceph-2 ceph-3
```

更多的配置与实践指导请访问 ZStack 企业版官方网站 <http://www.zstack.io/> .