誉天教育
YU TIAN EDU

Linux云计算架构师涨薪班

自定义crush map

学习目标

- CRUSH和CRUSH map简介
- CRUSH map的解译、编译和更新
- 编写自定义CRUSH map，以控制对象放置策略
- 使用命令行配置CRUSH map

Ceph集群写操作流程

- client首先访问ceph monitor获取cluster map的一个副本，知晓集群的状态和配置
- 数据被转化为一个或多个对象，每个对象都具有对象名称和存储池名称
- 以PG数为基数做hash，将对象映射到一个PG
- 根据计算出的PG，再通过CRUSH算法得到存放数据的一组OSD位置（副本个数），第一个是主，后面是从
- 客户端获得OSD ID，直接和这些OSD通信并存放数据

- 注： 以上所有操作都是在客户端完成 的，不会影响ceph集群服务端性能

一句话描述CRUSH

CRUSH的作用，就是**根据PG ID得一个OSD列表**

CRUSH和对象放置策略

- Ceph使用CRUSH算法（Controlled Replication Under Scalable Hashing 可扩展哈希下的受控复制）来计算哪些OSD存放哪些对象
- 对象分配到PG中，CRUSH决定这些PG使用哪些OSD来存储对象。理想情况下，CRUSH会将数据均匀的分布到存储中
- 当添加新OSD或者现有的OSD出现故障时，Ceph使用CRUSH在活跃的OSD上重平衡数据
- CRUSH map是CRUSH算法的中央配置机制，可通过调整CRUSH map来优化数据存放位置
- 默认情况下，CRUSH将对象放置到不同主机上的OSD中。可以配置CRUSH map和CRUSH rules，使对象放置到不同房间或者不同机柜的主机上的OSD中。也可以将SSD磁盘分配给需要高速存储的池

CRUSH map组成部分

- CRUSH hierarchy（层次结构）：一个树型结构，通常用于代表OSD所处的位置。默认情况下，有一个根bucket，它包含所有的主机bucket，而OSD则是主机bucket的树叶。这个层次结构允许我们自定义，对它重新排列或添加更多的层次，如将OSD主机分组到不同的机柜或者不同的房间
- CRUSH rule（规则）：CRUSH rule决定如何从bucket中分配OSD pg。每个池必须要有一条CRUSH rule，不同的池可map不同的CRUSH rule

CRUSH map的解译、编译和更新

- 导出CRUSH map
  - ceph osd getcrushmap -o ./crushmap.bin
- 解译CRUSH map
  - crushtool -d ./crushmap.bin -o crushmap.txt
- 修改后的CRUSH map重新编译
  - crushtool -c ./crushmap.txt -o ./crushmap-new.bin
- 更新CRUSH map
  - ceph osd setcrushmap -i crushmap-new

CRUSH map配置段

- CRUSH可调参数及其设置
- 所有物理存储设备列表
- 所有基础架构bucket以及各自含有的存储设备或其他bucket ID的列表
- 包含PG和OSD map的CRUSH rule列表

CRUSH可调参数

- 可通过选项来调整、禁用、启用CRUSH算法的功能
- CRUSH map的开头定义可调参数，可使用ceph osd crush show-tunables来查看
- 调整CRUSH map的参数可能会改变CRUSH将pg映射到OSD的方式。发生这种情况时，集群将把对象移到集群中的不同OSD，来反映重新计算后的map
- 除了修改个别可调项外，可以通过ceph osd crush tunables profile命令选择预定义的profile。一些profile需要ceph或ceph客户端的更低版本
- 预定义的profile:
  - legacy
  - Argonaut
  - bobtail
  - firefly
  - hammer
  - jewel
  - optimal：当前版本的最佳值

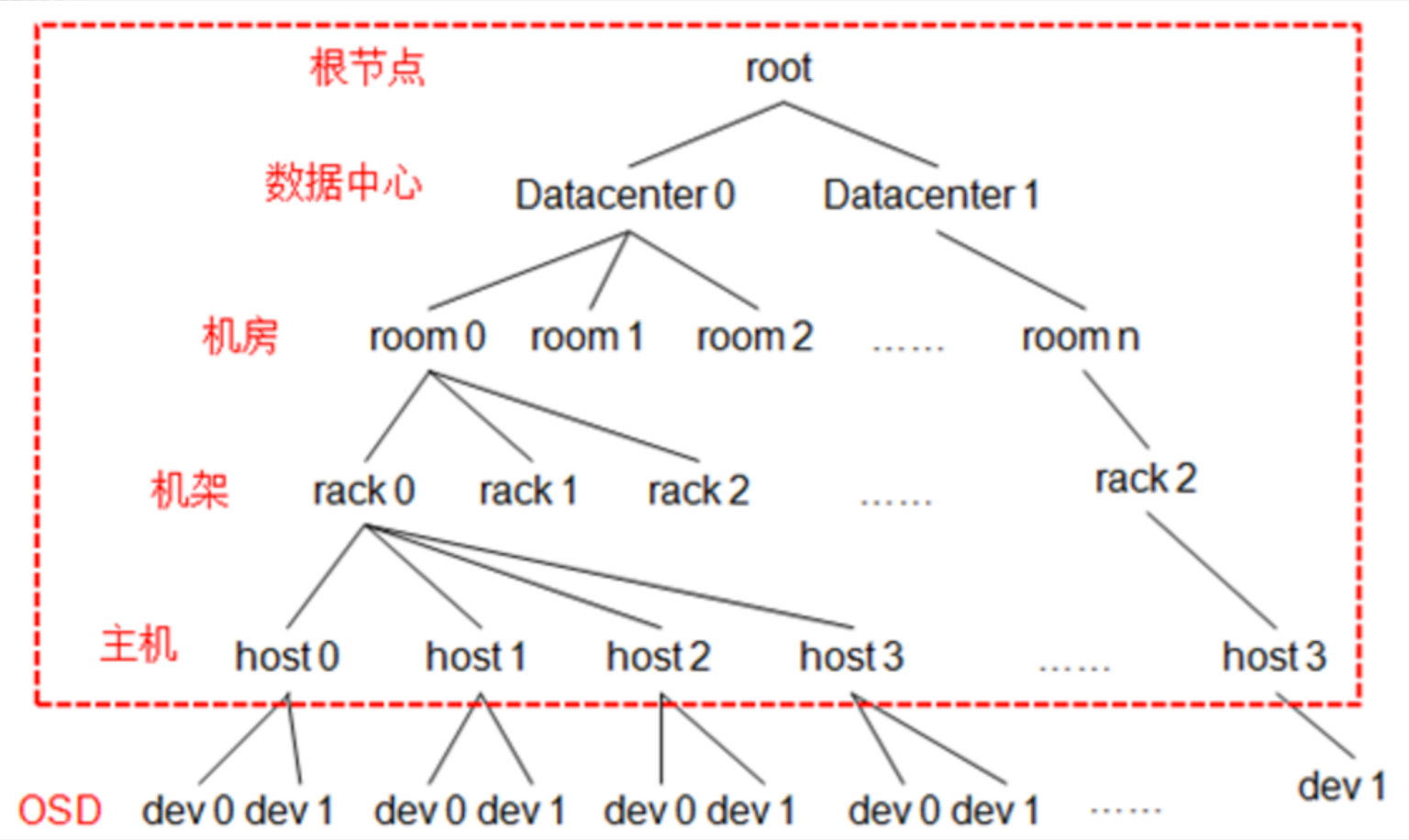
物理存储设备

- 存储设备的ID
- 存储设备的名称
- 存储设备的权重（以TB为单位的容量）
- 存储设备的类别
  - HDD
  - SSD
  - NVMe SSD
- 示例：
  - device 0 osd.0 weight 0.015 class hdd
  - device 1 osd.0 class ssd

CRUSH bucket类型

- 默认bucket类型：
  - osd
  - host
  - chassic
  - rack
  - row
  - pdu
  - pod
  - room
  - datacenter
  - region
  - root
- bucket类型支持自定义

bucket图解



CRUSH map bucket配置项

```
[bucket-type] [bucket-name] {
    id [一个负整数，以便与存储设备id区分]
    weight [权重]
    alg  [将pg map到osd时的算法，默认使用straw2]
    hash [每个bucket都有一个hash算法，目前Ceph支持rjenkins1算法，设为0即使用该算法]
    item [一个bucket包含的其他bucket或者叶子]
}

host servere {
        id -7           # do not change unnecessarily
        id -8 class hdd         # do not change unnecessarily
        # weight 0.029
        alg straw2
        hash 0  # rjenkins1
        item osd.4 weight 0.010
        item osd.6 weight 0.010
        item osd.8 weight 0.010
}
```

CRUSH规则

- CRUSH map包含数据放置规则，默认有两个规则： replicated_rule和erasure-code

- 通过ceph osd crush rule ls可列出现有规则，也可以使用ceph osd crush rule dump打印规则详细详细

CRUSH规则配置说明

```
rule <rulename> {
    id <id > [整数，规则id]
    type [replicated|erasure] [规则类型，用于复制池还是纠删码池]
    min_size <min-size> [如果池的最小副本数小于该值，则不会为当前池应用这条规则]
    max_size <max-size>[如果创建的池的最大副本大于该值，则不会为当前池应用这条规则]
    step take <bucket type> [这条规则作用的bucket，默认为default]根节点
    step [chooseleaf|choose] [firstn] <num> type <bucket-type>    叶子节点也就是故障域
        # num == 0 选择N（池的副本数）个bucket
        # num > 0且num < N   选择num个bucket
        # num < 0 选择N-num(绝对值)个bucket
    step emit
}
```

创建基于SSD的存储池

- 删除1、3、5OSD设备的类型并改为ssd
  - ceph osd crush rm-device-class osd.1　　ceph osd crush set-device-class ssd osd.1
  - ceph osd crush rm-device-class osd.3　　ceph osd crush set-device-class ssd osd.3
  - ceph osd crush rm-device-class osd.5　　ceph osd crush set-device-class ssd osd.5
- 创建crush rule规则
  - ceph osd crush rule create-replicated ssd_rule default host ssd
  - ssd_rule 是规则名 default 是root的根节点 host是普通节点 ssd是设备类型
- 创建存储池测试
  - ceph osd pool create pool_ssd ssd_rule
  - ceph pg dump pgs_brief |grep ^20 ssd池ID 查询pg是否都在ssd的osd上

从命令行添加crushmap规则

- 添加规则
  - ceph osd crush rule create-replicated <rulename> <root> <failure-domain-type> [class]
    - rulename：规则名称
    - root：CRUSH map层次结构的起始节点
    - failure-domain-type：故障域
    - class：要使用的设备类型，如ssd或者hdd，此参数可选
    - 示例：
      - ceph osd crush rule create-replicated inDC2 DC2 rack
- 查看规则
  - ceph osd crush rule ls
  - ceph osd crush rule dump <rule name>
- 应用规则
  - 创建一个新池，直接应用新的规则
    - ceph osd pool create myfirstpool 50 50 inDC2
  - 修改一个池的规则到新的规则
    - ceph osd pool set rbd crush_ruleset 1

从命令行更新CRUSH map层次结构

- 创建bucket
  - ceph osd crush add-bucket <name> <type>
  - 示例:
    - ceph osd crush add-bucket DC1 datacenter
    - ceph osd crush add-bucket rackA1 rack
    - ceph osd crush add-bucket rackB1 rack
- 将bucket整理到层次结构中
  - ceph osd move <name> type=<parent>
  - 示例:
    - ceph osd crush move rackA1 datacenter=DC1
    - ceph osd crush move rackB1 datacenter=DC1
    - ceph osd crush move DC1 root=default
- 设置osd的位置
  - ceph osd add osd.0 0.01500 root=default host=serverf

创建SSD作为主OSD的crush rule

- 创建root节点
  - ceph osd crush add-bucket cl260 root
- 创建rack节点
  - ceph osd crush add-bucket rack1 rack
  - ceph osd crush add-bucket rack1 rack
  - ceph osd crush add-bucket rack1 rack
- 创建host节点
  - ceph osd crush add-bucket hostc host
  - ceph osd crush add-bucket hostd host
  - ceph osd crush add-bucket hoste host

创建SSD作为主OSD的crush rule

- 将rack移动到root下
  - ceph osd crush move rack1 root=cl260
  - ceph osd crush move rack2 root=cl260
  - ceph osd crush move rack3 root=cl260
- 将host移动到rack下
  - ceph osd crush move hostc rack=rack1
  - ceph osd crush move hostd rack=rack2
  - ceph osd crush move hoste rack=rack3
- 将osd移动到host下（将ssd类型的osd移动到hostc下）
  - ceph osd crush move osd.0 host=hostc
  - ceph osd crush move osd.3 host=hostc
  - ceph osd crush move osd.4 host=hostc

创建SSD作为主OSD的crush rule

- 将OSD2、1、5添加给hostd
  - ceph osd crush move osd.2 host=hostd
  - ceph osd crush move osd.1 host=hostd
  - ceph osd crush move osd.5 host=hostd
- 将OSD7、6、8添加给hoste
  - ceph osd crush move osd.7 host=hoste
  - ceph osd crush move osd.6 host=hoste
  - ceph osd crush move osd.8 host=hoste

创建SSD作为主OSD的crush rule

- 导出crushmap
  - crushtool -d ./crushmap.bin -o crushmap.txt
- 添加规则

```
rule ssd_pool {
        id 2
        type replicated
        min_size 1
        max_size 10
        step take rack1
        step chooseleaf firstn 1 type host  # 第1个副本在rack1上
        step emit
        step take cl260 class hdd
        step chooseleaf firstn -1 type rack # 剩余副本在cl260根下的hdd上
        step emit
}
```

创建SSD作为主OSD的crush rule

- 重新编译crushmap
  - crushtool -c crushmap.txt -o crushmap-new.bin
- 导入crushmap
  - ceph osd setcrushmap -i crushmap-new.bin
- 创建存储池测试
  - ceph osd pool create pool_demo ssd_pool
- 查看池中PG的主OSD
  - ceph pg dump pgs_brief |grep ^21（池的ID号）

Ceph集群运行状况诊断

- 集群状态：
  - HEALTH_OK
  - HEALTH_WARN
  - HEALTH_ERR
- 常用查寻状态指令：
  - ceph health detail
  - ceph -s
  - ceph -w

管理OSD容量

- 当集群容量达到mon_osd_nearfull_ratio的值时，集群会进入HEALTH_WARN状态。这是为了在达到full_ratio之前，提醒添加OSD。默认设置为0.85，即85%
- 当集群容量达到mon_osd_full_ratio的值时，集群将停止写入，但允许读取。集群会进入到HEALTH_ERR状态。默认为0.95，即95%。这是为了防止当一个或多个OSD故障时仍留有余地能重平衡数据

- 设置方法：
  - ceph osd set-full-ratio 0.95
  - ceph osd set-nearfull-ratio 0.85
  - ceph osd dump

誉天教育
YU TIAN EDU

Thank you