Ceph MON

目录

**** 版本修订记录 ****

版本号	*修订时间*	*修订内容*
v1.0	*2018-08-09*	*初版修订*

一．简介

Ceph monitor 主要用于维护管理集群 cluster map 的主副本,在集群中的 osd 的状态发生改变的时候，实时更新 cluster map,各 monitor 之间使用改进的 Paxos 算法来提供一致性的元信息管理服务，它能为集群成员，配置和状态提供坚实的存储。客户端在和 Monitor 通信获取 cluster map 后，便能清楚的知道集群中 Ceph Monitor,Ceph OSD 进程或则 Ceph 元数据服务器的信息，从而可以和这些集群进程通信，另外，集群中的 monitor 还用来提供认证和日志服务；ceph 集群中的 map 主要包括下面五类：

- Monitor Map:主要包括了集群的 fsid 号，monitor 服务器的地址和端口，以及 map 被创建的时间，醉经修改时间和当前版本等信息,使用 ceph mon dump 命令可以查看 Monitor Map;
- OSD Map:主要包括了 osd 列表，版本信息，osd 的基本参数和集群中 Pool 的信息，使用 ceph osd dump 可以查看 OSD Map;
- PG Map:主要包括 pg 列表，pg 状态以及 pg 和 osd 的映射关系,使用 ceph pg dump 命令可以查看 PG Map;
- CRUSH Map:主要包括设备列表，bucket 类别，bucket 实例以及各类存储不同的存储规则;
- MDS Map:主要包括 map 的版本号，创建时间和修改时间以及存储文件系统元数据的 pool 信息和其服务器的状态，可使用命令 ceph fs dump 查看该 map;

二．原理

2.1 架构

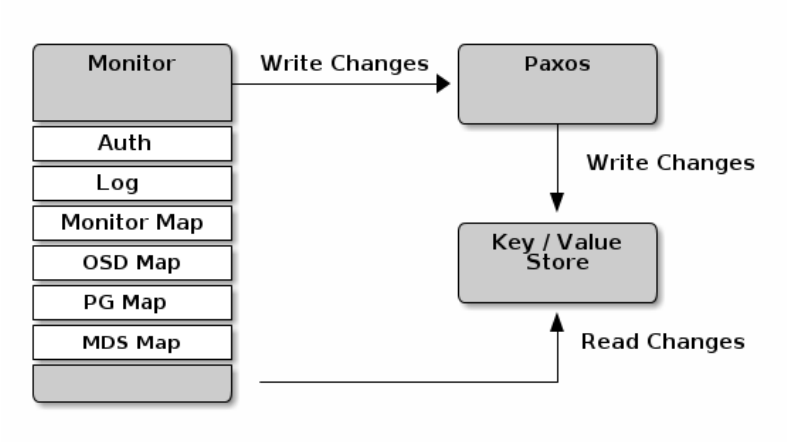

Figure 1 monitor 架构图

Monitor 服务会将所有的 Cluster Map 的变化都写到单个的 Paxos 实例中，并次持久化到后端 kv 数据库(rocksdb),其他 Monitor 在同步操作的时候可以查询到最新的 Cluster map;注意，monitor 和 monitor 之间不是通过配置文件 ceph.conf 来连接彼此，不像客服端或其他集群中的进程，它是通过 monmap 来连接的彼此的，这样设计的原因时避免了配置文件出现编写错误时导致集群失败，

并且配置文件无法集群的变化而实时更新以及保存更新之前的 monmap，这些都是由分布式一致性算法 paxos 实现的；

2.2 monitor 数据同步

在生产环境中一般存在多个 monitor,每个 monitor 都会周期性的查看是否其他的 monitor 已经生成了新的 cluster map.在 cluster map 发生变化的时候，总会存在有些 monitor 的 cluster map 的版本落后于其他的 monitor,这时版本落后的这些 monitor 会被剔除 quorum,这类 monitor 被称为 requestor,它会重新向 leader 请求最新的 cluster map,leader 会将 cluster map 同步的任务分配给 provider,该 provider 将提供给 requester 最新的 cluster map,在 monitor map 中，如图 2 所示；会存在如下的三类 monitor:

- Leader:最先生产 cluster map 并被选举成功；
- Provider:拥有最新的 cluster map,但不是最新生成的；
- Requester:该类 monitor 已被剔除了 quorum,并向 Leader 请求最新的 cluster map 后重新加入 quorum;

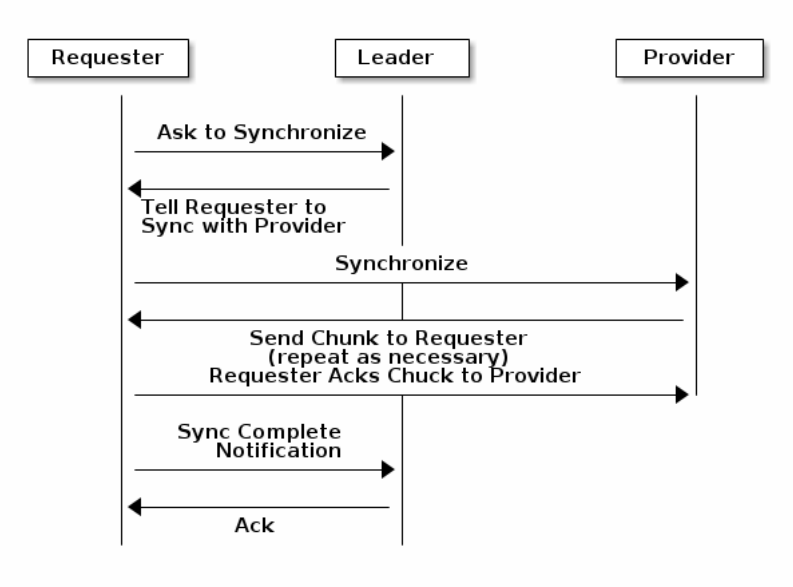

Figure 2 monitor 数据同步图

Monitor map 的数据更新严重依赖时间戳，monitor 服务器之间的时间不同步会导致很多集群的很多异常问题，比如丢弃同步数据，同步超时等，须在每台服务器上安装 ntp 服务用于时间同步；

三．操作

3.1 初始化 mon

```
ceph-authtool --create-keyring  /tmp/ceph.mon.keyring --gen-key -n mon. –cap mon ‘allow *’ #该用户是 ceph 集群的第一个用户，用于 monitor 之间的互联密钥以及创建后续的一些用户；

sudo ceph-authtool --create-keyring /etc/ceph/ceph.client.admin.keyring --gen-key -n client.amdin --set-uid=0 --cap mon ‘allow *’ --cap osd ‘sllow *’ --cap mds ‘allow *’  --cap mgr ‘allow *’  # 创建能够访问所有集群进程的用户

sudo ceph-authtool --create-keyring /var/lib/ceph/bootstrap-osd/ceph.keyring --gen-key -n client.bootstrap-osd --cap mon ‘profile bootstrap-osd’ #创建用于启动 osd 的用户

sudo ceph-authtool /tmp/ceph.mon.keyring –import-keyring /etc/ceph/ceph.client.admin.keyring
sudo ceph-authtool /tmp/ceph.mon.keyring –import-keyring /var/lib/ceph/bootstrap-osd/ceph.keyring #将 keyring 导入到 ceph.mon.keyring

monmaptool --create --add {hostname} {ip-address} --fsid {uuid} /tmp/monmap #创建一个 monmap，用于添加到 quorum 中
```

3.2 增加 mon

```
sudo mkdir /var/lib/ceph/mon/ceph-{mon-id} #创建 monitor 的数据目录
ceph auth get mon. -o /tmp/{keyring-file} #到处 mon.用户的 keyring
ceph mon getmap -o /tmp/{monmap-file} #获取 mon map 数据
sudo ceph-mon -i {mon-id} --mkfs --monmap /tmp/{monmap-file} --keyring /tmp/{keyring-file} #使用 monmap 和 keyring 初始化 monitor 数据目录
```

3.3 删除 mon

```
service ceph –a stop mon.{mon_id} # 关闭 ceph mon 进程
ceph mon remove {mon-id} # 移除集群中的 monitor
```

3.4 mon 参数

```
mon force quorum join # 使之前未在 quorum 的 monitor 强制加入 quorum 中
fsid # 集群的唯一编号
mon initial members # 初始化的 monitor 机器名，可以包含多个服务器，用于建立 quorum
mon data # monitor 数据存放的路径，建议使用默认值/var/lib/mon/$cluster-$id,另外 monitor 使用 mmap()函数写数据，它会经常将数据刷新到磁盘，所以生产环境中不要将 mon 和 osd 部署在同一台服务器上；
mon data size warn # 当 monitor 的数据达到多大时发出警告 HEALTH_WARN，默认 15G
mon data avail warn # 当 monitor 的数据所在的目录使用量少于该值时会发出警告 HEALTH_WARN,默认 30
mon data avail crit # 当 monitor 的数据所在的目录使用量少于该值时集群报错 HEALTH_ERR，默认值为 5.
```

mon warn on cache pools without hit sets # 当缓存池没有设置 hitset 类型时触发 HEALTH_WARN
mon warn on crush straw calc version zero # 当 CURSH 的 tunable straw_calc_version 为 0 时触发 HEALTH_WARN
mon warn on legacy crush tunables # 当 CRUSH tunables 的值小于 mon warn on legacy crush tunables 时触发 HEALTH_WARN
mon crush min required version # 集群需要的最小可调配置文件版本
mon warn on osd down out interval zero # 当 mon osd down out interval 的值为 0 时出发 HEALTH_WARN
mon cache target full warn ratio # pool 的值在 cache_target_full 和 target_max_object 之间时出发 HEALTH_WARN
mon health data update interval # monitor 多久和 quorum 内的 monitor 发布自己的状态，默认为 60s，该值为浮点数。
mon health to clog # 是否将集群的健康状况周期性的写入日志，默认为 true；
mon health to clog tick interval # monitor 多久检查一下集群的状态变化并在变化的时候写入日志；
mon health to clog interval # monitor 将健康状况记录到日志的周期，monitor 的健康概况无论发生变化与否都会写入日志；当该值为负数时不记录日志；

mon osd full ratio # 集群禁止读写时 osd 磁盘的最大使用率。默认值为 0.95
mon osd backfillfull ration # osd 磁盘使用率被认为不能进行回填的最大使用率，默认值 0.90
mon osd nearfull ratio # osd 磁盘被将要忙时的使用率,默认值为 0.85

mon sync trim timeout
mon sync heartbeat timeout
mon sync heartbeat interval
mon sync backoff timeout
mon sync timeout
mon sync max retries
mon sync max payload size
paxos max join drift
paxos max join drift
paxos stash full interval
paxos propose interval
paxos min
paxos min wait
paxos trim min
paxos trim max
paxos service trim min
paxos service trim max
mmon max pgmap epochs
mon max log epochs
mon mds force trim to
mon osd force trim to
mon osd cache size
mon election timeout
mon lease

mon lease renew interval factor
mon lease ack timeout factor
mon accept timeout factor
mon min osdmap epochs
mon max pgmap epochs
mon max log epochs

mon tick interval #
mon clock drift allowed
mon clock drift warn backoff
mon timecheck interval
mon timecheck skew interval

mon client hunt interval
mon client ping interval
mon client max log entries per message
mon client bytes

mon allow pool delete
osd pool default flag hashpspool
osd pool default flag nodelete
osd pool default flag nopgchange
osd pool default flag nosizechange

mon max osd
mon globalid prealloc
mon subscribe interval
mon stat smooth intervals
mon probe timeout
mon daemon bytes
mon max log entries per event
mon osd prime pg temp
mon osd prime pg temp max time
mon osd prime pg temp max time estimate
mon osd allow primary affinity
mon osd pool ec fast read
mon mds skip sanity
mon max mdsmap epochs
mon config key max entry size
mon scrub interval
mon scrub max keys
mon compact on start
mon compact on bootstrap
mon compact on trim
mon cpu threads
mon osd mapping pgs per chunk
mon osd max split count
mon session timeout #默认 300s,monitor 会中断超过这个时间的 session

3.5 其他命令

```
ceph mon dump # 获取 monitor map
```

四．参考资料

【1】Ceph Monitor 实现

【2】ceph monitor paxos 的实现(一)

【3】Paxos 算法详解

【4】mon conf ref