

|  |  |
| --- | --- |
|  | **CEPH MON** |
|  |  |
|  | Author  **与ceph同行**  8/9/18 |

Contents

[**一．简介** 4](#_Toc521942834)

[**二．原理** 4](#_Toc521942835)

[2.1 架构 4](#_Toc521942836)

[2.2 monitor数据同步 5](#_Toc521942837)

[**三．操作** 5](#_Toc521942838)

[3.1 初始化mon 5](#_Toc521942839)

[3.2 增加mon 6](#_Toc521942840)

[3.3 删除mon 6](#_Toc521942841)

[3.4 mon参数 6](#_Toc521942842)

[3.5 其他命令 9](#_Toc521942843)

[**四．参考资料** 9](#_Toc521942844)

***\* 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* |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

***\* Release Copyleft©free \****

**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

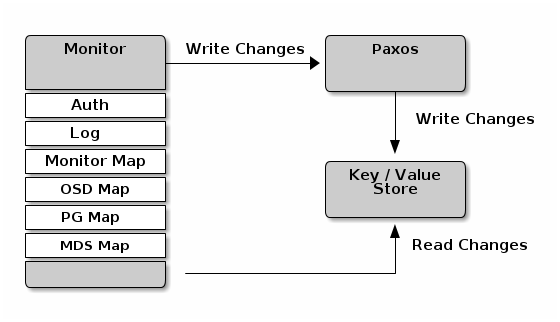


Figure 1 monitor架构图

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;

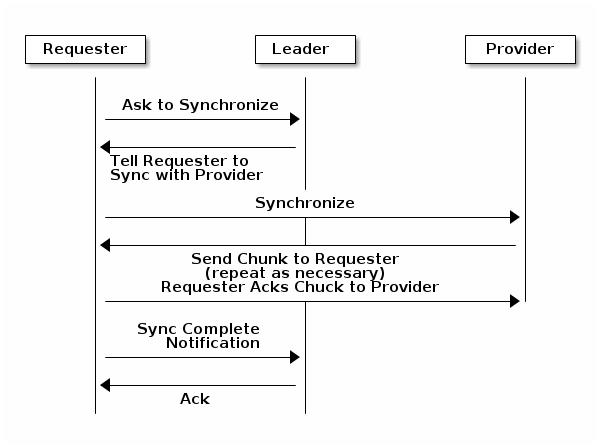


Figure 2 monitor数据同步图

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](https://www.jianshu.com/p/60b34ba5cdf2)

【2】[ceph monitor paxos的实现(一)](https://yq.aliyun.com/articles/99855?t=t1)

【3】[Paxos算法详解](http://www.cnblogs.com/endsock/p/3480093.html" \l "4029595)

【4】[mon conf ref](http://docs.ceph.com/docs/master/rados/configuration/mon-config-ref/)