Ceph monitor分析

## monitor简介

monitor作为ceph集群核心服务，负责自维护ceph集群健康状态，rados集群元数据mdsmap、monmap、osdmap、log、auth、health等，基于改进paxos算法，保证集群节点间的数据状态在同一时刻一致性。总的来说Monitor负责收集集群信息，更新集群信息，发布集群信息。



图 1 monito架构图

可以看出monitor的内部其实是一个分布式kv数据库。从下往上分别是MonitorDBStore、Paxos和PaxosService。PaxosService负责保证每次都只会有一个提案进入paxos流程。Paxos模块具体实现了multi-Paxos算法。MonitorDBStore是对底层DB的抽象封装，将DB的基本操作事务封装成统一接口，当前DB默认使用rocksdb。

## monitor启动过程

monitor服务启动过程主要包括：加载基本配置、检查和加载db、初始化网络模块、mon注册网络监听、monitor bootstrap启动。

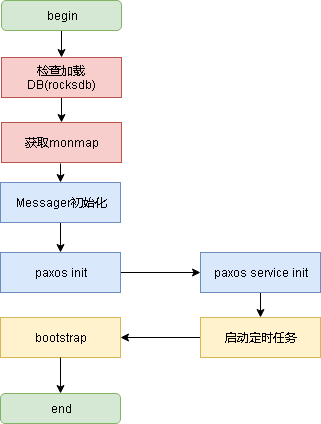


图 2 monitor启动步骤

下面介绍monitor的核心流程，选举和数据读写流程。

## monitor 核心流程

### Paxos算法介绍

Paxos算法是一种常用于分布式系统的共识一致性算法，它主要解决的问题是分布式系统中某个值(决议)如何达成一致。Paxos可以实现数据副本一致性，分布式锁，名字管理等。Paxos包括原始Paxos(basic paxos)和变种优化的multi-paxos等，其中multi-paxos更适合工程实践。

#### 1）Basic Paxos：

Basic Paxos包含的角色：

* 1. Proposer提出提案，可以是一个或多个proposer。
  2. Acceptor决策是否接受来自Proposer的提案。
  3. Learner最终提案的学习者。

Basic Paxos算法流程包含三个阶段：

1. prepare阶段

proposer向所有acceptor广播prepare的请求，请求中带有一个全局唯一且自增的proposal num（pn），acceptor接收消息后不再接收pn<=当前pn的prepare消息或propose消息。Acceptort接收之后向proposer回复promise消息。



图 3 prepare阶段a



图 4 prepare阶段b

1. accept阶段

proposer收到过半数且最大pn的提案[pn,value]，然后将提案发送给acceptor。Accept收到propose请求，大于等于本地pn，则将[pn,value]保存到本地，并且返回proposer已接收。



图 5 accept阶段a



图 6 accept阶段b

1. learn阶段

Proposer收到多数Acceptors的Accept后，决议形成，将通过的提案发送给所有Learners。



图 7 learn阶段

#### 2）Multi paxos：

原始的Paxos算法（Basic Paxos）只能对一个值形成决议，决议的形成至少需要两次网络来回，在高并发情况下可能需要更多的网络来回，极端情况下甚至可能形成活锁。如果想连续确定多个值，Basic Paxos搞不定了。因此Basic Paxos几乎只是用来做理论研究，并不直接应用在实际工程中。

实际应用中几乎都需要连续确定多个值，而且希望能有更高的效率。Multi-Paxos正是为解决此问题而提出。Multi-Paxos基于Basic Paxos做了两点改进：

1、针对每一个要确定的值，运行一次Paxos算法实例（Instance），形成决议。

2、在所有Proposers中选举一个Leader，由Leader唯一地提交Proposal给Acceptors进行表决。这样没有Proposer竞争，解决了活锁问题。在系统中仅有一个Leader进行Value提交的情况下，Prepare阶段就可以跳过，从而将两阶段变为一阶段，提高效率。

Ceph monitor采用的是Multi paxos算法。

### monitor选举过程

monitor::bootstrap()是选举入口，整个过程也由一系列状态变化而成。每个Monitor启动后，根据配置文件中的主机ip列表，发现其他monitor并且获取其他节点最新日志版本号，根据版本号大小判断是否需要从其他节点拉取db数据做同步，然后选出leader和peon，再做recovery，最后根据收到的信号shutdown。



图 8 选举状态机

首先看看probe过程：

* 本节点发送probe请求monmap里记录的其他节点，进入STATE\_PROBING状态。对端节点收到请求后，将本端first commit和last commit版本号返回。
* 接收到probe reply后比较版本号差异，节点进入STATE\_SYNCHRONIZING状态，如果本节点last commit版本号小于对端first version则需要从对端节点做全量数据同步。如果本节点的last commit和对端的last version之间只差paxos\_max\_join\_drift，则进行差异数据同步。同步数据为paxos数据。



图 9 probe过程

* 完成同步后，如果回复数超过节点数一半就开始选举，节点进入STATE\_ELECTING状态。先广播OP\_PROPOSE消息，对端接收到消息，如果发现自身rank大于发送端rank值，则回复OP\_ACK消息。这样的话rank最小的节点会收到最多的ack消息，如果收到的ack消息数与当前活动节点数相同则本端节点成为leader，然后再通知各个节点本端胜出，其他节点将成为peon。



图 10 选举过程

* 确定leader和peon后，需要做提案的recovery。做recovery的目的是为了使集群保留相同的accept pn号，并且提交上次未完成的提案。
* Recovery过程是先执行collect()收集leader节点pending\_v未完成的版本号和pending\_pn未完成的提案号以及相关数据uncommitted\_value，同时生成新pn，然后leader节点再向集群中其他peon节点发送collect消息。Peno收到collect请求，如果peno的提案号小于leader提案号则更新peno的提案号，如果peno存在uncommit value并且版本号大于leader的last\_commited，则将uncommit\_value发回leader节点。Leader收到消息，发现peon的accept\_pn较大，则继续collect()直到peno与leader的accepted\_pn相同。当所有节点的pn都一致了，而且也还存在uncommitted的提案，则完成提交begin(uncommitted\_value)。如果没有未提交提案或者提案提交完成，就更新租约extend\_lease()。



图11 recovery流程

* Monitor中实现paxos的提交过程是由begin()来完成。Leader首先更新自己的pending\_v和pending\_pn，然后发送OP\_BEGIN消息到所有的peon节点，如果peon的accepted\_pn小于leader发送的，则接受这个提案，将数据保存到本地，并且回复OP\_ACCEPT消息。Leader收到的accept的个数等于集群的quorum，则将提案提交到leader本地，并且通知各个peon去完成commit操作。同时更新租约extend\_lease()，更新租约后leader和peon之间会有个定时任务，leader会默认每3秒更新一次租约，超时10s则会重新选举，peon同样没有收到lease消息，超过10s也会重新选举。



图 12 提案提交过程

可以看到，monitor启动经历了probe、sync数据、选举、recovery过程，其中recovery涉及到了提案的提交动作，后续所有的monitor的提案提交都是通过begin()完成的。这里来看，paxos的第一阶段就是在bootstrap()里完成，后续提案提交直接可以通过leader发给peon来完成，符合multi-paxos的描述。

### monitor读写过程