curator 提供应用场景的封装

curator-reciples

master/leader选举

分布式锁（读锁、写锁）

分布式队列

…

LeaderLatch

写一个master

LeaderSelector

每一个应用都写一个临时有序节点，根据最小的节点来获得优先权

zookeeper的几个原理分析

# zookeeper集群角色

## leader

leader是zookeeper集群的核心。

1. 事务请求的唯一调度者和处理者，保证集群事务处理的顺序性
2. 集群内部各个服务器的调度者

## follower

1. 处理客户端非事务请求，以及转发事务请求给leader服务器
2. 参与事务请求提议（proposal）的投票（客户端的一个事务请求，需要半数服务器投票通过以后才能通知leader commit； leader会发起一个提案，要求follower投票）
3. 参与leader选举的投票

## observer

观察zookeeper集群中最新状态的变化并将这些状态同步到observer服务器上

增加observer不影响集群中事务处理能力，同时还能提升集群的非事务处理能力

# zookeeper的集群组成

zookeeper一般是由 2n+1台服务器组成

# leader选举

leaderElection/AuthFastLeaderElection/FastLeaderElection

QuorumPeer startLeaderElection

源码地址：<https://github.com/apache/zookeeper.git>

需要的条件： jdk 1.7以上 、ant 、idea

FastLeaderElection

serverid : 在配置server集群的时候，给定服务器的标识id（myid）

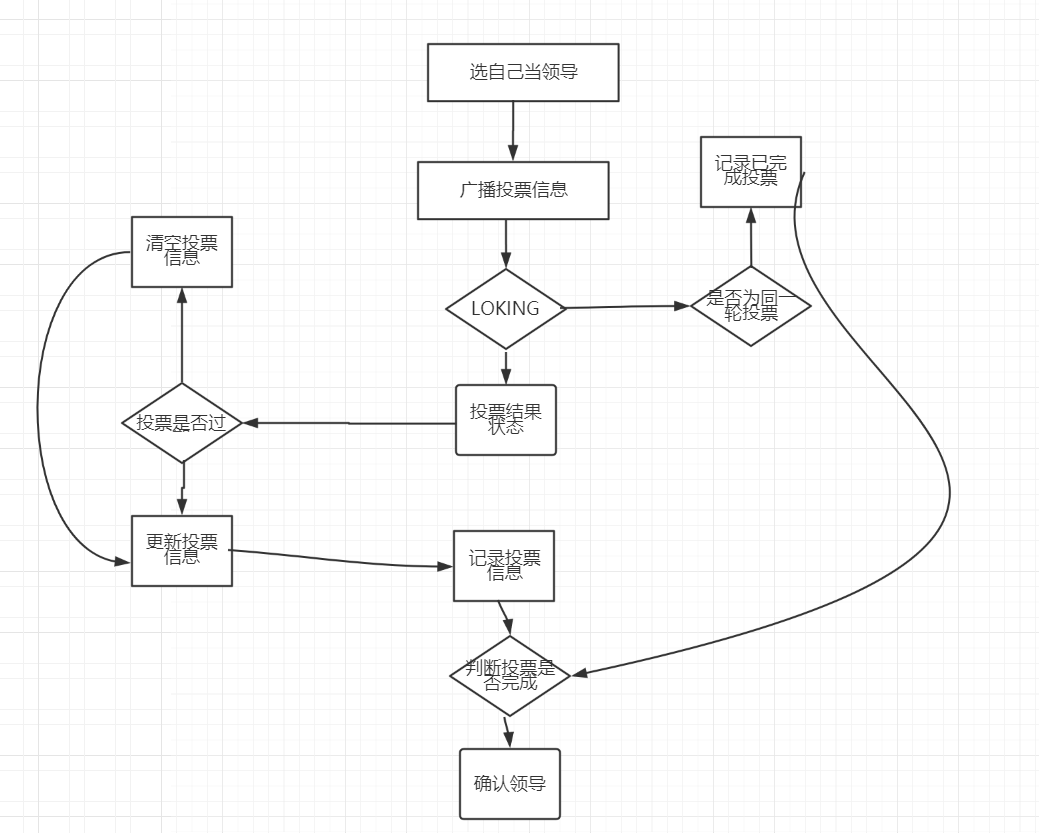
zxid : 服务器在运行时产生的数据ID， zxid的值越大，表示数据越新

Epoch: 选举的轮数

server的状态：Looking、 Following、Observering、Leading

第一次初始化启动的时候： LOOKING

1. 所有在集群中的server都会推荐自己为leader，然后把（myid、zxid、epoch）作为广播信息，广播给集群中的其他server, 然后等待其他服务器返回
2. 每个服务器都会接收来自集群中的其他服务器的投票。集群中的每个服务器在接受到投票后，开始判断投票的有效性
   1. 判断逻辑时钟(Epoch) ，如果Epoch大于自己当前的Epoch，说明自己保存的Epoch是过期。更新Epoch，同时clear其他服务器发送过来的选举数据。判断是否需要更新当前自己的选举情况
   2. 如果Epoch小于目前的Epoch，说明对方的epoch过期了，也就意味着对方服务器的选举轮数是过期的。这个时候，只需要讲自己的信息发送给对方



# ZAB协议

拜占庭问题

paxos协议主要就是如何保证在分布式环网络环境下，各个服务器如何达成一致最终保证数据的一致性问题

ZAB协议，基于paxos协议的一个改进。

zab协议为分布式协调服务zookeeper专门设计的一种支持崩溃恢复的原子广播协议

zookeeper并没有完全采用paxos算法， 而是采用zab Zookeeper atomic broadcast

## zab协议的原理

1. 在zookeeper 的主备模式下，通过zab协议来保证集群中各个副本数据的一致性
2. zookeeper使用的是单一的主进程来接收并处理所有的事务请求，并采用zab协议，

把数据的状态变更以事务请求的形式广播到其他的节点

1. zab协议在主备模型架构中，保证了同一时刻只能有一个主进程来广播服务器的状态变更
2. 所有的事务请求必须由全局唯一的服务器来协调处理，这个的服务器叫leader，其他的叫follower

leader节点主要负责把客户端的事务请求转化成一个事务提议（proposal），并分发给集群中的所有follower节点

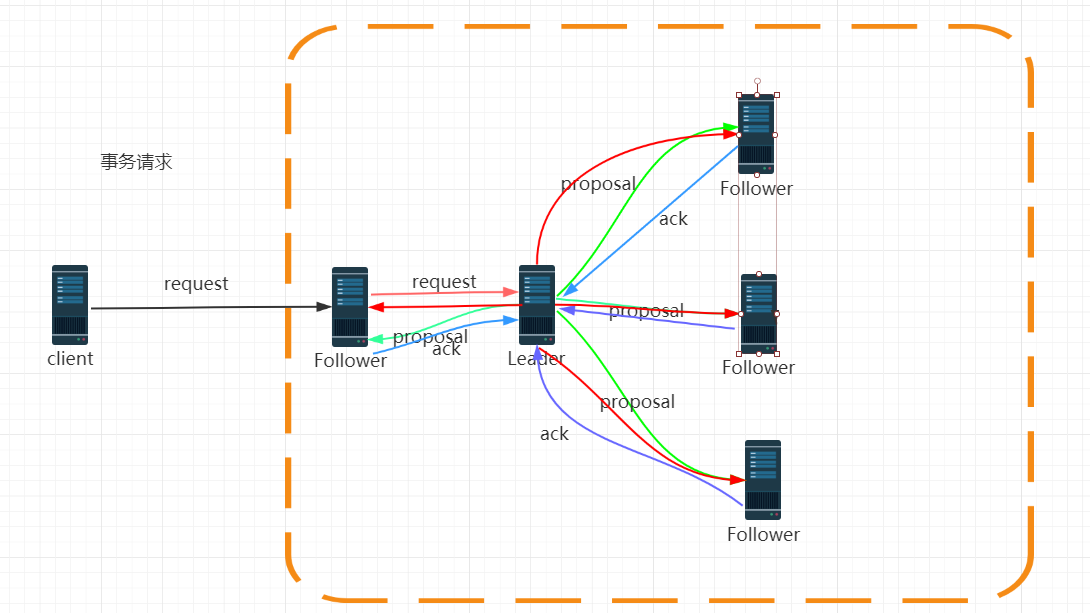
再等待所有follower节点的反馈。一旦超过半数服务器进行了正确的反馈，那么leader就会commit这条消息

崩溃恢复

原子广播

# zab协议的工作原理

1. 什么情况下zab协议会进入崩溃恢复模式
2. 当服务器启动时
3. 当leader服务器出现网络中断、崩溃或者重启的情况
4. 集群中已经不存在过半的服务器与该leader保持正常通信
5. zab协议进入崩溃恢复模式会做什么
6. 当leader出现问题，zab协议进入崩溃恢复模式，并且选举出新的leader。当新的leader选举出来以后，如果集群中已经有过半机器完成了leader服务器的状态同（数据同步），退出崩溃恢复，进入消息广播模式
7. 当新的机器加入到集群中的时候，如果已经存在leader服务器，那么新加入的服务器就会自觉进入数据恢复模式，找到leader进行数据同步



# 问题

假设一个事务在leader服务器被提交了，并且已经有过半的follower返回了ack。 在leader节点把commit消息发送给folower机器之前

leader服务器挂了怎么办

zab协议，一定需要保证已经被leader提交的事务也能够被所有follower提交

zab协议需要保证，在崩溃恢复过程中跳过哪些已经被丢弃的事务

# watcher的原理