

## 一、Zookeeper 简介

Zookeeper 是一个以文件管理为核心的分布式协调服务。

Zookeeper 中有三种角色：Leader、Follower 和 Observer。一个 Zookeeper 集群中只有一个 Leader，为客户端提供读服务和写服务，而 Follower 和 Observer 都不能提供写服务。Leader 和 Follower 通过“Leader 选举”产生，而 Observer 并不参加这个过程。尽管 Follower 并不提供写服务，但是会参与“过半写成功”策略，所以还是会影响写性能，因此 Observer 可以在不影响写性能的情况下提升集群的读性能。

在客户端访问 Zookeeper 时，会与服务器建立一个 TCP 连接，客户端可以通过这个连接进行和服务器的会话，可以向 Zookeeper 发送请求和接收相应，可以接收来自服务器的 Watch 事件通知。当因客户端导致连接断开后，只要在 SessionTimeout 规定的时间内重新连上集群的任意一台服务器即可使之前创建的会话仍然有效。

Zookeeper 的结构是一个树形结构，Leader 就相当于根节点，Follower 相当于其叶节点。Zookeeper 的节点分为两类：持久节点和临时结点。持久结点一旦被创建，除非主动对其进行移除操作，其都会保留在 Zookeeper 上；而临时结点则是由客户端的会话绑定，一旦客户端会话失效，这个客户端创建的临时节点将会被移除。

Zookeeper 允许用户向指定节点上增加 Watcher，用于监测一些关心的节点。在节点被删除，更改或是增加了子节点等情况下会向用户进行反馈。

## 二、Zookeeper 的功能

### 1. 数据的发布和订阅

对于发布和订阅的系统，一般来讲有两种模式——服务端主动向用户推送的“推模式”和用户向服务端发出请求的“拉模式”。Zookeeper 中采用的是两种相结合的模式，用户可以选择自己想要关注的节点，在节点发生变更时，服务端将推送“推”给客户端一个通知，之后有客户端自行决定是否将新的数据从服务端上“拉”回来。

### 2. 分布式协调

Zookeeper 中有着 Watcher 和异步通知机制，能够实现分布式环境下不同机器甚至是不同系统之间的通知和协调。

### 3. Leader 选举

Leader 选举是 Zookeeper 中重要的功能之一。

在没有 Leader 的情况下，可能是刚开始也可能是在之前的 Leader 挂掉后，此时便要开始 Leader 选举。

如果是在刚开始，所有服务器除 Observer 均进入 Looking 状态；如果是在之前的 Leader 挂掉后，所有 Follower 会发现无法与 Leader 取得联系，然后就进入 Looking 状态。总而言之，进行选举前所有参加选举的服务器均为 Looking 状态。

首先进行第一轮投票，每个服务器向其他机器发送自己的“选票”，其中包含了 myid 和 ZXID 两个内容。第一轮所有的服务器都会“选”自己，即把自己的 myid 和 ZXID 发送出去。

然后每个服务器就会接收到来自其他服务器的选票，接下来进行判断，判断这个选票是否有效——是否来自 Looking 状态的服务器，是否是本轮的投票。

判断为有效后，接下来与自己的选票进行对比，遵循 ZXID 优先于 myid，大优先于小这两条规则，更新自己的选票并重新投票。

当有超过半数的选票统一时，就“选”出了 Leader，其他的服务器则变成 Follower，继续正常的工作。此时选举整个过程结束。

### 4. Master 选举

Master 选举也是 Zookeeper 中重要的功能之一。

作为服务器，一般来讲会有要求全天无休工作的需求，而此时若是出现了宕机，带来的损失可想而知。因此我们需要多个“相同的”节点来保证在一个宕机时其他的可以顶上来。

不过什么时候顶也是个难题，如果在之前的节点没死掉的情况下就顶掉也会带来问题，因此备用节点可以向主节点发送 `ping` 包，若没有应答则判断主节点死掉了。然而有时可能会因为网络原因而导致没能收到应答，此时顶掉主节点会带来同样的问题。

这时，`Zookeeper` 中的临时节点就发挥作用了。临时节点在经过了 `SessionTimeout` 设定的时间之后会自动被删除掉，所以我们可以让主节点在 `Zookeeper` 的某处建立一个临时节点，相当于“宣布”了“我是主节点”。在有其他的节点也想要当主节点时，会因为已经有这个临时节点存在而失败。此时这个节点并不是直接走掉，而是将这个节点订阅，在这个节点删除时，也就是之前的主节点死掉的时候。这时这些备用节点就可以来“抢”这个主节点的位置了，当然和之前一样，只要有一个节点抢到了这个位置，它就成为了主节点，其他的节点也就不能再抢了。

通过以上的行为便可以保证主节点存在且唯一，这样就能满足全天无休的工作了。