04 | 分布式选举：国不可一日无君

2019-09-30 聂鹏程

分布式技术原理与算法解析 进入课程 >

讲述：聂鹏程
时长 13:49　大小 12.66M

你好，我是聂鹏程。今天，我来继续带你打卡分布式核心技术。

相信你对集群的概念并不陌生。简单说，集群一般是由两个或两个以上的服务器组建而成，每个服务器都是一个节点。我们经常会听到数据库集群、管理集群等概念，也知道数据库集群提供了读写功能，管理集群提供了管理、故障恢复等功能。

接下来，你开始好奇了，对于一个集群来说，多个节点到底是怎么协同，怎么管理的呢。比如，数据库集群，如何保证写入的数据在每个节点上都一致呢？

也许你会说，这还不简单，选一个“领导”来负责调度和管理其他节点就可以了啊。

这个想法一点儿也没错。这个“领导”，在分布式中叫做主节点，而选“领导”的过程在分布式领域中叫作分布式选举。

然后，你可能还会问，怎么选主呢。那接下来，我们就一起去揭开这个谜底吧。

为什么要有分布式选举?

主节点，在一个分布式集群中负责对其他节点的协调和管理，也就是说，其他节点都必须听从主节点的安排。

主节点的存在，就可以保证其他节点的有序运行，以及数据库集群中的写入数据在每个节点上的一致性。这里的一致性是指，数据在每个集群节点中都是一样的，不存在不同的情况。

当然，如果主故障了，集群就会天下大乱，就好比一个国家的皇帝驾崩了，国家大乱一样。比如，数据库集群中主节点故障后，可能导致每个节点上的数据会不一致。

这，就应了那句话“国不可一日无君”，对应到分布式系统中就是“集群不可一刻无主”。总结来说，选举的作用就是选出一个主节点，由它来协调和管理其他节点，以保证集群有序运行和节点间数据的一致性。

分布式选举的算法

那么，如何在集群中选出一个合适的主呢？这是一个技术活儿，目前常见的选主方法有基于序号选举的算法（ 比如，Bully 算法）、多数派算法（比如，Raft 算法、ZAB 算法）等。接下来，就和我一起来看看这几种算法吧。

长者为大：Bully 算法

Bully 算法是一种霸道的集群选主算法，为什么说是霸道呢？因为它的选举原则是“长者”为大，即在所有活着的节点中，选取 ID 最大的节点作为主节点。

在 Bully 算法中，节点的角色有两种：普通节点和主节点。初始化时，所有节点都是平等的，都是普通节点，并且都有成为主的权利。但是，当选主成功后，有且仅有一个节点成为主节点，其他所有节点都是普通节点。当且仅当主节点故障或与其他节点失去联系后，才会重新选主。

Bully 算法在选举过程中，需要用到以下 3 种消息：

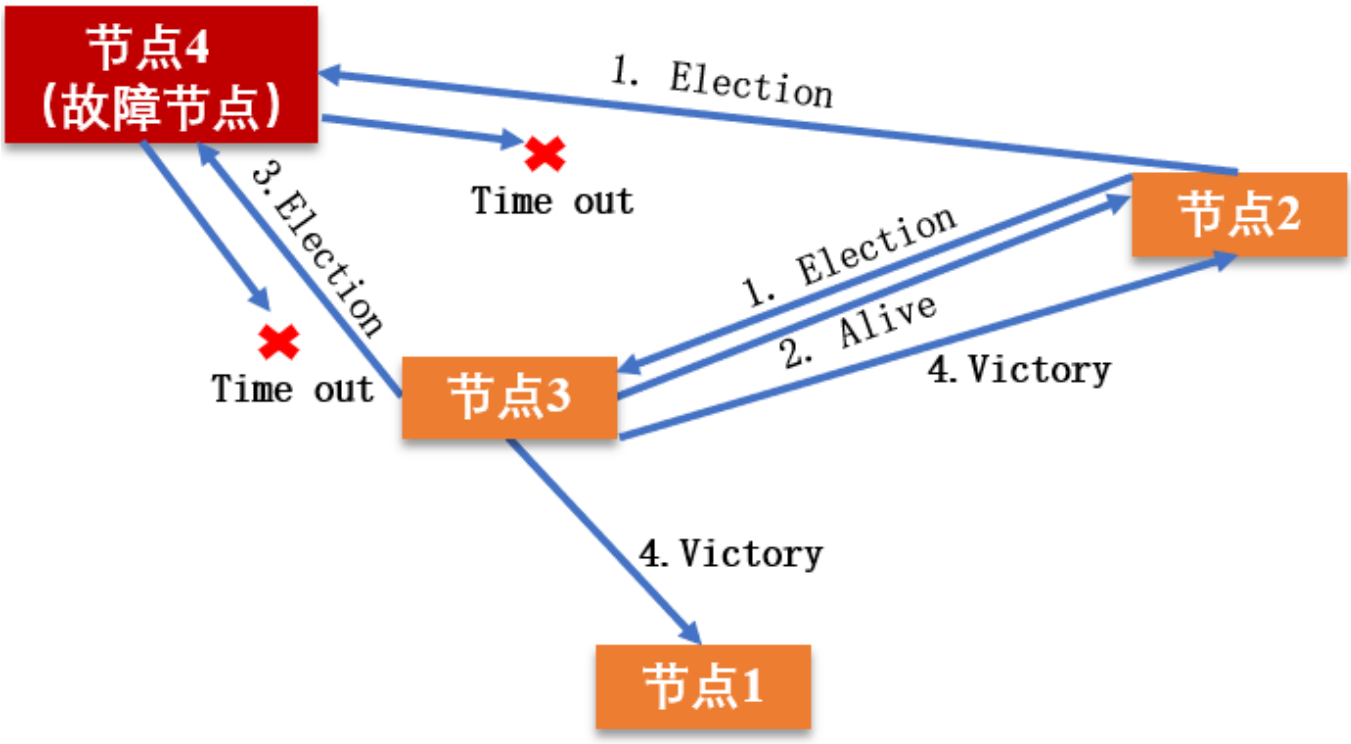
Election 消息，用于发起选举；

Alive 消息，对 Election 消息的应答；

Victory 消息，竞选成功的主节点向其他节点发送的宣誓主权的消息。

Bully 算法选举的原则是“长者为大”，意味着它的**假设条件是，集群中每个节点均知道其他节点的 ID。**在此前提下，其具体的选举过程是：

1. 集群中每个节点判断自己的 ID 是否为当前活着的节点中 ID 最大的，如果是，则直接向其他节点发送 Victory 消息，宣誓自己的主权；
2. 如果自己不是当前活着的节点中 ID 最大的，则向比自己 ID 大的所有节点发送 Election 消息，并等待其他节点的回复；
3. 若在给定的时间范围内，本节点没有收到其他节点回复的 Alive 消息，则认为自己成为主节点，并向其他节点发送 Victory 消息，宣誓自己成为主节点；若接收到来自比自己 ID 大的节点的 Alive 消息，则等待其他节点发送 Victory 消息；
4. 若本节点收到比自己 ID 小的节点发送的 Election 消息，则回复一个 Alive 消息，告知其他节点，我比你大，重新选举。



注：节点i的ID值即为i

目前已经有很多开源软件采用了 Bully 算法进行选主，比如 MongoDB 的副本集故障转移功能。MongoDB 的分布式选举中，采用节点的最后操作时间戳来表示 ID，时间戳最新的节点其 ID 最大，也就是说时间戳最新的、活着的节点是主节点。

小结一下。Bully 算法的选择特别霸道和简单，谁活着且谁的 ID 最大谁就是主节点，其他节点必须无条件服从。这种算法的优点是，选举速度快、算法复杂度低、简单易实现。

但这种算法的缺点在于，需要每个节点有全局的节点信息，因此额外信息存储较多；其次，任意一个比当前主节点 ID 大的新节点或节点故障后恢复加入集群的时候，都可能会触发重新选举，成为新的主节点，如果该节点频繁退出、加入集群，就会导致频繁切主。

民主投票：Raft 算法

Raft 算法是典型的多数派投票选举算法，其选举机制与我们日常生活中的民主投票机制类似，核心思想是“少数服从多数”。也就是说，Raft 算法中，获得投票最多的节点成为主。

采用 Raft 算法选举，集群节点的角色有 3 种：

Leader，即主节点，同一时刻只有一个 Leader，负责协调和管理其他节点；

Candidate，即候选者，每一个节点都可以成为 Candidate，节点在该角色下才可以被选为新的 Leader；

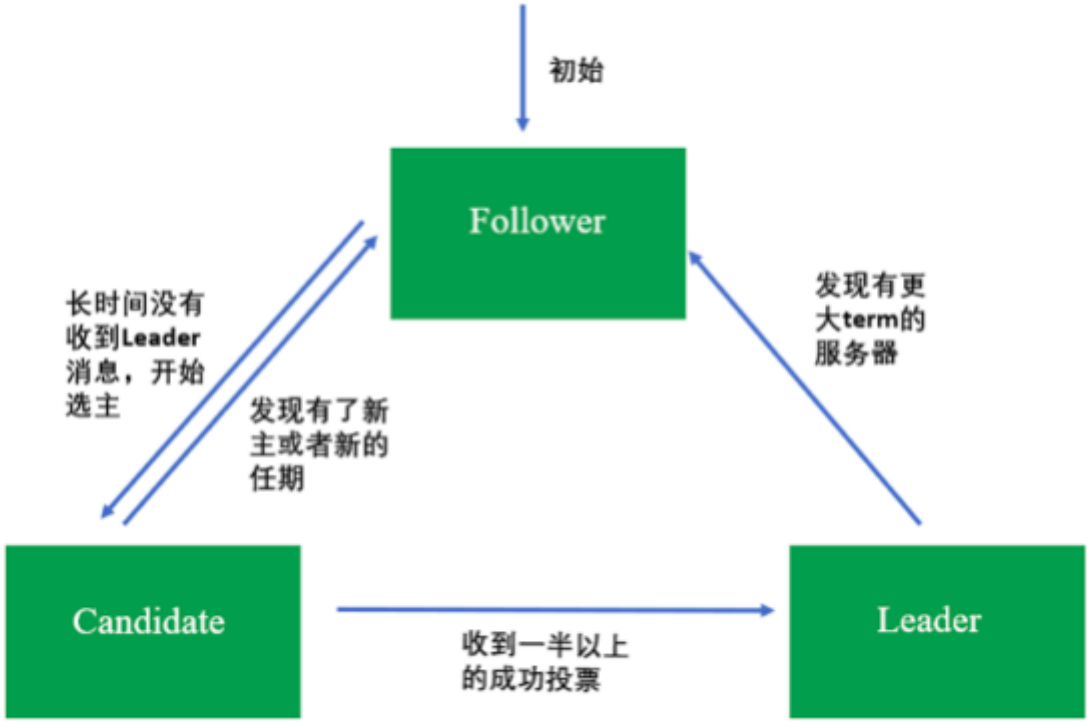
Follower，Leader 的跟随者，不可以发起选举。

Raft 选举的流程，可以分为以下几步：

1. 初始化时，所有节点均为 Follower 状态。
2. 开始选主时，所有节点的状态由 Follower 转化为 Candidate，并向其他节点发送选举请求。
3. 其他节点根据接收到的选举请求的先后顺序，回复是否同意成为主。这里需要注意的是，在每一轮选举中，一个节点只能投出一张票。
4. 若发起选举请求的节点获得超过一半的投票，则成为主节点，其状态转化为 Leader，其他节点的状态则由 Candidate 降为 Follower。Leader 节点与 Follower 节点之间会定期发送心跳包，以检测主节点是否活着。

5. 当 Leader 节点的任期到了，即发现其他服务器开始下一轮选主周期时，Leader 节点的状态由 Leader 降级为 Follower，进入新一轮选主。

节点的状态迁移如下所示（图中的 term 指的是选举周期）：



请注意，**每一轮选举，每个节点只能投一次票。**这种选举就类似人大代表选举，正常情况下每个人大代表都有一定的任期，任期到后会触发重新选举，且投票者只能将自己手里唯一的票投给其中一个候选者。对应到 Raft 算法中，选主是周期进行的，包括选主和任值两个时间段，选主阶段对应投票阶段，任值阶段对应节点成为主之后的任期。但也有例外的时候，如果主节点故障，会立马发起选举，重新选出一个主节点。

Google 开源的 Kubernetes，擅长容器管理与调度，为了保证可靠性，通常会部署 3 个节点用于数据备份。这 3 个节点中，有一个会被选为主，其他节点作为备。Kubernetes 的选主采用的是开源的 etcd 组件。而，etcd 的集群管理器 etcds，是一个高可用、强一致性的服务发现存储仓库，就是采用了 Raft 算法来实现选主和一致性的。

小结一下。Raft 算法具有选举速度快、算法复杂度低、易于实现的优点；缺点是，它要求系统内每个节点都可以相互通信，且需要获得过半的投票数才能选主成功，因此通信量大。该算法选举稳定性比 Bully 算法好，这是因为当有新节点加入或节点故障恢复后，会触发选主，但不一定会真正切主，除非新节点或故障后恢复的节点获得投票数过半，才会导致切主。

具有优先级的民主投票：ZAB 算法

ZAB（ZooKeeper Atomic Broadcast）选举算法是为 ZooKeeper 实现分布式协调功能而设计的。相较于 Raft 算法的投票机制，ZAB 算法增加了通过节点 ID 和数据 ID 作为参考进行选主，节点 ID 和数据 ID 越大，表示数据越新，优先成为主。相比较于 Raft 算法，ZAB 算法尽可能保证数据的最新性。所以，ZAB 算法可以说是对 Raft 算法的改进。

使用 ZAB 算法选举时，集群中每个节点拥有 3 种角色：

Leader，主节点；

Follower，跟随者节点；

Observer，观察者，无投票权。

选举过程中，集群中的节点拥有 4 个状态：

Looking 状态，即选举状态。当节点处于该状态时，它会认为当前集群中没有 Leader，因此自己进入选举状态。

Leading 状态，即领导者状态，表示已经选出主，且当前节点为 Leader。

Following 状态，即跟随者状态，集群中已经选出主后，其他非主节点状态更新为 Following，表示对 Leader 的追随。

Observing 状态，即观察者状态，表示当前节点为 Observer，持观望态度，没有投票权和选举权。

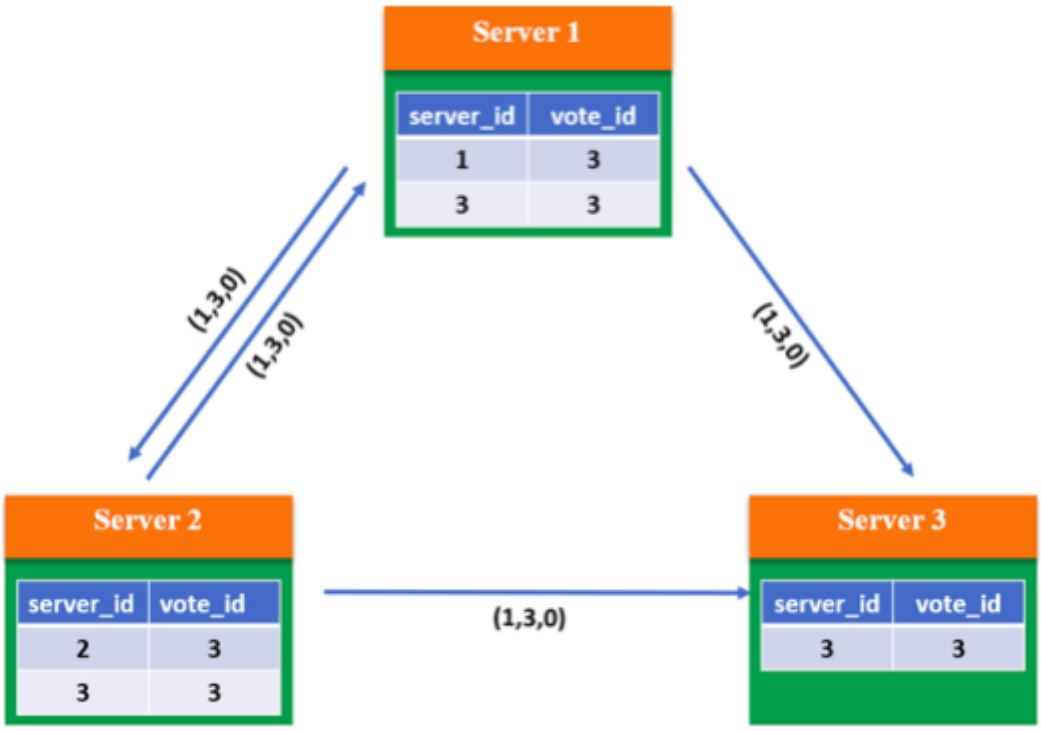
投票过程中，每个节点都有一个唯一的三元组 (server_id, server_zxID, epoch)，其中 server_id 表示本节点的唯一 ID；server_zxID 表示本节点存放的数据 ID，数据 ID 越大表示数据越新，选举权重越大；epoch 表示当前选取轮数，一般用逻辑时钟表示。

ZAB 选举算法的核心是“少数服从多数，ID 大的节点优先成为主”，因此选举过程中通过 (vote_id, vote_zxID) 来表明投票给哪个节点，其中 vote_id 表示被投票节点的 ID，vote_zxID 表示被投票节点的服务器 zxID。**ZAB 算法选主的原则是：server_zxID 最大者成为 Leader；若 server_zxID 相同，则 server_id 最大者成为 Leader。**

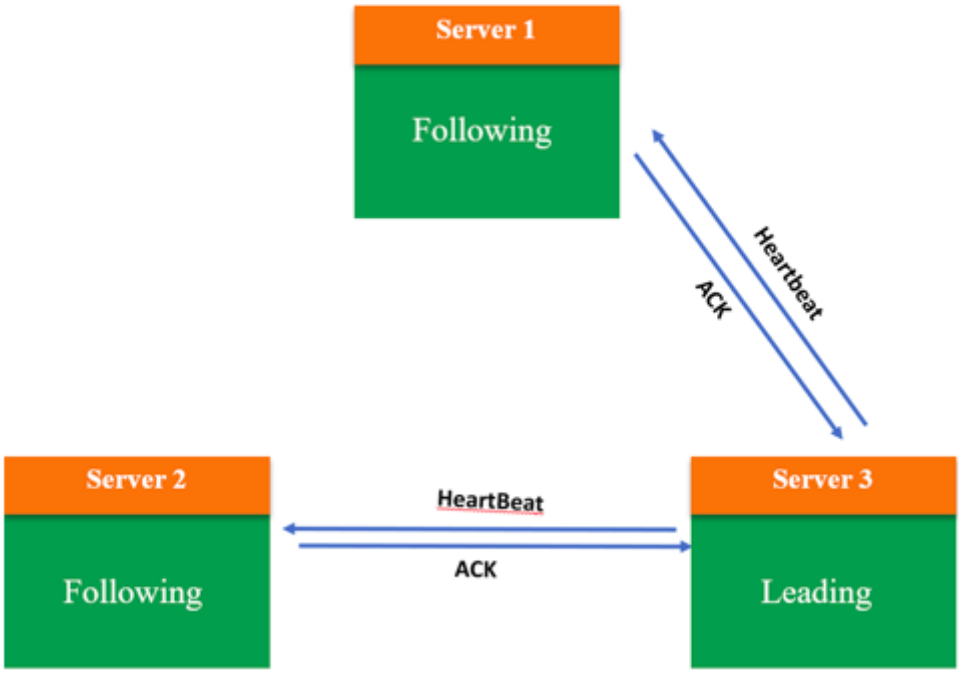
接下来，我以 3 个 Server 的集群为例，此处每个 Server 代表一个节点，与你介绍 ZAB 选主的过程。

第一步：当系统刚启动时，3 个服务器当前投票均为第一轮投票，即 epoch=1，且 zxID 均为 0。此时每个服务器都推选自己，并将选票信息 <epoch, vote_id, vote_zxID> 广播出去。

第二步：根据判断规则，由于 3 个 Server 的 epoch、zxID 都相同，因此比较 server_id，较大者即为推选对象，因此 Server 1 和 Server 2 将 vote_id 改为 3，更新自己的投票箱并重新广播自己的投票。



第三步：此时系统内所有服务器都推选了 Server 3，因此 Server 3 当选 Leader，处于 Leading 状态，向其他服务器发送心跳包并维护连接；Server1 和 Server2 处于 Following 状态。



小结一下。ZAB 算法性能高，对系统无特殊要求，采用广播方式发送信息，若节点中有 n 个节点，每个节点同时广播，则集群中信息量为 n*(n-1) 个消息，容易出现广播风暴；且除了投票，还增加了对比节点 ID 和数据 ID，这就意味着还需要知道所有节点的 ID 和数据 ID，所以选举时间相对较长。但该算法选举稳定性比较好，当有新节点加入或节点故障恢

复后，会触发选主，但不一定会真正切主，除非新节点或故障后恢复的节点数据 ID 和节点 ID 最大，且获得投票数过半，才会导致切主。

三种选举算法的对比分析

好了，我已经带你理解了分布式选举的 3 种经典算法，即 Bully 算法、Raft 算法和 ZAB 算法。那么接下来，我就从消息传递内容、选举机制和选举过程的维度，对这 3 种算法进行一个对比分析，以帮助你理解记忆。

	Bully算法	Raft算法	ZAB算法
选举消息回复类型	alive消息	同意或不同意选举的消息	投票信息 <epoch, vote_id, vote_zxID>
Leader选举机制	偏向于让ID更大的节点作为Leader	收到过半数的投票，则当选为Leader	倾向于让数据最新或者ID值最大的节点作为Leader
选举过程	只要节点发现Leader无响应时，或者ID较大的节点恢复故障时，就会发起选举	每个角色为Candidate的节点可参与竞选Leader，且每一个Follower只有一次投票权，即同意或者不同意Candidate的选举	每个节点都可以处于Looking状态参与竞选，都可以多次重新投票，根据epoch、zxID、server_id来选择最佳的节点作为Leader
选举所需时间	短	较短	较长
性能	Bully < Raft < ZAB		

知识扩展：为什么“多数派”选主算法通常采用奇数节点，而不是偶数节点呢？

多数派选主算法的核心是少数服从多数，获得投票多的节点胜出。想象一下，如果现在采用偶数节点集群，当两个节点均获得一半投票时，到底应该选谁为主呢？

答案是，在这种情况下，无法选出主，必须重新投票选举。但即使重新投票选举，两个节点拥有相同投票数的概率也会很大。因此，多数派选主算法通常采用奇数节点。

这，也是大家通常看到 ZooKeeper、 etcd、Kubernetes 等开源软件选主均采用奇数节点的一个关键原因。

总结

今天，我首先与你讲述了什么是分布式选举，以及为什么需要分布式选举。然后，我和你介绍了实现分布式选举的 3 种方法，即：Bully 算法、Raft 算法，以及 ZooKeeper 中的 ZAB 算法，并通过实例与你展示了各类方法的选举流程。

我将今天的主要内容总结为了如下所示的思维导图，来帮助你加深理解与记忆。

思考题

1. 分布式选举和一致性的关系是什么？

2. 你是否见到过一个集群中存在双主的场景呢？

我是聂鹏程，感谢你的收听，欢迎你在评论区给我留言分享你的观点，也欢迎你把这篇文章分享给更多的朋友一起阅读。我们下期再会！

上一篇　03 | 分布式互斥：有你没我，有我没你

精选留言 (7)

写留言

每天晒白牙
2019-09-30

今天这篇文章赚到了
1.分布式选举算法是为了保证数据一致性的
在集群中存在多个节点提供服务，如果每个节点都可以写数据，这样容易造成数据的不一致，所以需要选举一个leader，往leader节点中写数据，然后同步到follower节点中。这样就能更好的保证一致性...

展开

4　2

Jxin
2019-09-30

1.是实现数据一致性的一个保障高可用的算法。
2.假定老师要表达的是由于网络分区出现的脑裂现象。解决思路可以引入租凭。
3.我记得<深入分布式缓存>里zk的共识算法指的好像是paxos，和这里的zab有什么区别?

展开


1024
2019-09-30

两主的情况出现在集群因为网络原因，被划分了两部分局部可通信的区域。下面的链接详细讲解了Raft算法，及双主出现后集群是如何恢复的。
https://www.infoq.cn/article/coreos-analyse-etcd/
还有一个Raft算法动画链接
http://thesecretlivesofdata.com/raft/#election

展开


tracy
2019-09-30

redis cluster就是基于多个主节点实现的，每个主节点至少对于一个从节点，保证了集群高可用，提高了集群性能，并且还保证了集群的容错性，易于横向扩展

展开


cp★钊□□□
2019-09-30

想问下老师，选举的性能，评判的标准是什么？为什么zab的性能最好，是指哪方面的性能?


忆水寒
2019-09-30

双主机的情况是可以存在的吧，假如网络有问题。

展开


随心而至
2019-09-30

1.分布式选举和一致性，感觉是密不可分的。重新选举依靠一致性提供的数据，一致性又要依靠选举出来的主节点进行。这里我只了解过raft算法
https://www.cnblogs.com/xybaby/p/10124083.html
2.有个brain split（脑裂），比如说两个机房原来网络想通，可以正确选主，后来网络不通，每个机房都只知道自己的小山头，他们就容易各自占山为王。...
展开

1