基于区块链的共识算法Raft的实现

摘 要：Raft 是一种用来管理日志复制的一致性算法，与Paxos相比，它更易理解和工程实践化，因此广泛的应用于各种分布式系统。正是由于它高度的可理解性，可实践性以及可扩展性，目前也被应用于区块链中。业界常用的几种共识算法就是PoW，PBFT，Paxos，PoS，DPoS，其中针对于区块链中的私有链，由于网络中的所有节点是可信任的，为了效率，往往使用Raft。

本文主要围绕区块链共识算法这一点来对比上面几种常用算法的异同以及使用场景，并就Raft给出具体实现以便于理解区块链如何达成共识。其中主要实现了Raft的两个核心模块：领导选举和日志复制。

关键词：Raft，区块链，共识算法，领导选举，日志复制

**Implementation of Consensus Algorithm Raft Based on Blockchain**

**ABSTRACT:**

Raft is a consistency algorithm used to manage log replication. Compared with Paxos, it is easier to understand and practice, so it is widely used in various distributed systems. It’s because of its high degree of comprehensibility, practicality, and extensibility that it is also currently used in blockchains. Several commonly used consensus algorithms in the industry are PoW, PBFT, Paxos, PoS, and DPoS. For private blockchains, Raft is often used for efficiency because all nodes in the network are trusted.

This article mainly focuses on the blockchain consensus algorithm to compare the similarities and differences between the above common algorithms and usage scenarios, and gives a specific implementation of Raft to understand how the blockchain can achieve consensus. Among them, two core modules of Raft are implemented: leader election and log replication.

**Key words:** Raft, blockchain, consensus algorithm, leader election, log replication

1 绪论

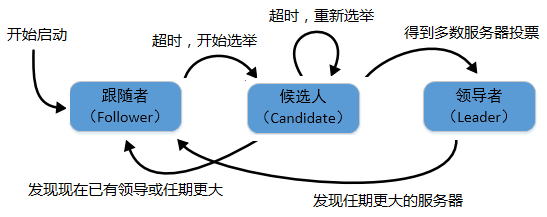
分布式系统又叫分布式计算机系统，根据《分布式系统概念与设计》一书中的定义，分布式系统是一个软件或者硬件组件分布在不同的网络机器上，彼此之间仅仅通过消息传递进行通信和协作的系统。说得通俗一点，所谓的分布式系统，实际上就是有多台计算机构成，基于网络通信传递消息构成的一个计算机系统，尽管这样的系统是有多台计算机构成的，但是对外是透明的，也就是说，当系统提供服务的时候实际上用户感觉就好像只有一台机器提供24小时无间断的服务。

从最初的单机处理到现在的各种大规模的分布式系统，分布式系统的发展越来越好，逐渐应用于我们周围的日常服务（Web搜索，社会服务，电子商务，在线游戏等等），几乎是无处不在了。比如Google，网页搜索技术的领导者，在支持用于搜索的分布式系统基础设施做出了巨大的努力，它代表了计算机史上最复杂和最大的分布式系统设施之一，它由超大数目的位于全世界多个数据中心的联网计算机组成，独有的分布式文件系统（GFS），支持超大文件快速读写，能够提供超大并行和分布式计算。随着互联网的发展和技术的演进，分布式系统必将发挥越来越重要的作用，也会有更多的技术在此基础上发展，区块链正是构建于分布式系统上的一种技术，对于区块链来说，它的核心优点就是去中心化, 不被任何个体权威控制，能够通过运用加密数据、共识算法、经济奖励和分布式数据存储等手段, 解决在分布式系统中实现点对点交易，从而为中心化机构普遍存在的低效率，高成本等问题。随着比特币近年来的快速发展, 区块链技术的应用与研究也呈现迅速增长趋势，被很多人认为是下一代云计算的雏形, 有希望像互联网一样重新塑造人类社会活动形态, 并有可能实现从当前的信息互联网向价值互联网的转变。

2共识算法

区块链首先是一个分布式系统，由传统的单机（CS模型）演变到多机的分布式系统，最为重要的一个问题就是如何保证多台机器的数据一致性，从而保证系统达成共识。而这些往往是通过某些协议（或者是共识机制，算法）来保证的。一般来说，共识算法可分为传统的分布式一致性算法以及区块链独有的共识算法，在某些特定的场景下比如私有区块链，使用传统的分布式一致性算法效率会更高，效果会更好。因此，从某种意义上来说，传统分布式一致性算法也可以算得上是共识算法的子集，本章主要会讨论以Raft，PBFT为代表的传统分布式一致性共识算法和Pow，PoS等区块链独有共识算法的异同，以及他们分别适用于区块链的什么场景。

3 Raft的设计，实现和测试

Raft源自于2013年的一篇论文《In Search of an Understandable Consensus Algorithm(Extended Version)》，它主要解决了在分布式环境下如何保证数据的一致性从而让系统达成共识。在这之前，Paxos算法在分布式环境一致性相关问题占领了统治地位，绝大多数的一致性实现都是基于Paxos或者受其影响。然而Paxos的算法最大的缺点就是难以理解并且难以运用于工程实践，而Raft算法正是在这种情况下应运而生。它不仅能满足分布式环境一致性的需求，更重要的是它比其他算法更简单且更易于理解，能广泛的应用于工程实践，同时它的安全性和效率与其他算法也不相上下。

一个 Raft 集群包括n台服务器，必须保证至少n/2+1台机器正常，比如对于一个有着5台服务器集群，最多能够允许2台机器故障，剩余的3台机器必须要能够正常运行，这样子整个系统才能保持正常服务。在任意的时间，每一个服务器一定会处于以下三种状态中的某一个状态：追随者（Follower）、候选人（Candidate）、领导人（Leader）。常规情形下，领导人只有一个，其他是追随者，响应领导人的要求。追随者们是被动的：他们不会发送任何的请求，只能够回复来自领导人和候选人的请求比如选举投票，复制日志等。如果追随者在一段时间内没有收到来自领导人的消息，它会成为一个候选人并且开始新的一次选举，收到大多数服务器投票的候选人会成为新的领导人。而对于领导人， 他会来处理所有来自客户端的请求。通常，领导人会尽最大可能性保持领导人的状态直到它自己宕机，这样子系统会重新开始一次选举选出领导人重复以上过程。上面服务器的状态图很好的说明了这个过程。

Raft算法的基本流程是初始化各个节点（Raft结构），开始领导选举，一旦领导选举成功，接受客户端请求，分发日志要求各节点复制日志，最后达成commited状态，一旦领导者宕机就需要重新选举，重复以上过程。

前面完成了算法的设计和实现，接下来主要给出了Raft算法的测试，在这里使用MIT 6.824提供的一套测试框架，因为Raft算法如果要应用于分布式系统的话，不得不面对网络通信，RPC框架，并发编程等等这些细节，这给算法的测试带来了很多的不确定性，MIT6.824提供的这套测试框架为我们模拟一套基于channel机制的RPC框架，这样不仅避免重复造轮子编写RPC框架，同时基于channel模拟的RPC框架能够尽量模拟出复杂的网络动态，比如机器的宕机重启，网络分割等等。

MIT6.824的这套测试框架所有的代码都在raft文件夹下的test\_test.go文件，所有以Test开头的函数名以及包含了t \*testing.T参数的函数都为测试函数，测试的总体流程就是模拟好环境，控制机器的连通性，最后通过领导人的状态，日志的一致性来判断算法的正确性。

4 结论

通过这次毕业设计，我学到了很多东西，最为核心的就是学习了Raft算法，然而在学习的过程中，让我对分布式系统的基础知识有了一些初步的了解，同时在实现Raft算法的过程中，学习到了Go语言的网络编程，并发编程，RPC等，通过Raft算法以点到面的让我了解了区块链的各种共识机制以及对应的应用场景。当然，限于时间和能力所限，我只完成了Raft算法的两个核心部分，领导选举和日志复制，而日志压缩和配置更新只是简单了解，并未实现，希望能在以后继续完善并且优化。

参 考 文 献

[1] [Robert Morris](http://pdos.csail.mit.edu/~rtm/), [Malte Schwarzkopf](https://people.csail.mit.edu/malte/) . MIT 6.824: Distributed Systems 2018.2

https://pdos.csail.mit.edu/6.824/

[2] [杨保华](https://book.douban.com/search/%E6%9D%A8%E4%BF%9D%E5%8D%8E), [陈昌](https://book.douban.com/search/%E9%99%88%E6%98%8C). 区块链原理、设计与应用[M]. 机械工业出版社, 2017.8

[3] 蒋勇. 白话区块链[M]. 机械工业出版社, 2017.10

[4] Nakamoto S. Bitcoin. a peer-to-peer electronic cash system

https://bitcoin.org/bitcoin.pdf, 2009.1

[5] Diego Ongaro, John Ousterhout. In Search of Understandable Consensus Algorithm 2013

[6] 袁勇，王飞跃. 区块链技术发展现状与发展 2016.4