# 如何让公链快起来<上>

从 2008 年 11 月中本聪（Satoshi Nakamoto）发表论文「Bitcoin: A Peer-to-Peer Electronic Cash System」算起，比特币即将迎来第一个十周年。这十年中，比特币与其背后的区块链技术蓬勃发展，以去中心化技术之名，大有变革整个在线数字世界的气势和雄心。

不过，雄心归雄心，正蓬勃发展的区块链技术，尤其是公链领域，有一个瓶颈却一直有待突破：以当今数字世界的规模和体量，任何一个在线系统，如果没有一个大容量、高吞吐的基础设施，就无法承载哪怕仅仅一个互联网级别的应用。

很可惜，中本聪的论文中完全没有考虑到这个问题，也许是走出这第一步实属不易，他也没想太多之后的事情，也许是这样的一个高性能的设计，在彻底去中心化的系统中难度太大。总之，近10年过去了，为了提高区块链系统的性能，前赴后继出现了大把项目，但到今天为止，并没有出现能够承载互联网级别应用的解决方案。

这是一个世界性的难题，全世界最聪明的学者、开发者都在尝试解决这个问题。我曾在微软工作多年，担任微软研究院主管研究员，很长一段时间专注于分布式系统方面的研究；离开微软之后，我又在创新工场担任负责区块链和人工智能投资方向的执行董事。多年在分布式系统方面的研究心得，以及在区块链投资领域评估多个公链项目的经验，让我深深明白，在彻底去中心化的系统中实现高性能设计，是一项难度极高、极具挑战的工作。

我看到行业内存在大量对于区块链公链性能瓶颈及解决方法的讨论，有些充满洞见，令人受益匪浅，但也有不少谬误，更有很多为了自身项目宣传而编造的似是而非的见解，颇有把讨论引入歧途的风险。在和多位该行业顶尖的学者、开发人员、投资人多次深入交流之后，他们都鼓励我把自己的看法分享出来。再三思索之后，我决定把自己对该话题的一些拙见记录下来，这样既可以让自己的一些思考能够沉淀，同时，也希望能和对该话题感兴趣的更多同仁进行一些探讨。

## 1 不要只关注性能瓶颈，而忽略了容量瓶颈

先说一下我的一个结论：**在当前以类金融为主流应用场景的情形下，区块链系统最首要的性能瓶颈是区块数据的广播延迟造成的，本质上受限于互联网的带宽和通讯延迟，这一点直接制约了吞吐量(TPS)。**只要是「Chain of Blocks」的系统，无论具体采用了什么共识算法，**无论是工作量证明（POW）、权益证明（POS）、拜占庭容错（BFT），还是委托权益证明（DPOS），在出下一个区块之前，都需要保证前一个区块在全网有一定的同步率，从而约束了每个区块不能太大，出块频率也不能太高，然后，这个问题无解。**

请注意，这里说的区块链系统特指**「Chain of Blocks」的系统**，其特征是要保证系统能最终收敛到一条单一的链表结构，并只有这条链上面的区块才是被确认的，反例是「Graph of Blocks」系统，例如所采用的 DAG 结构IOTA。

假设物理网络的带宽和延迟可以被忽略，例如基于数据中心高速链路的 EOS，系统第二个瓶颈是受限的账簿容量，本质上受限于单台全节点的内存容量，这一点直接制约了链上可以承载多少个用户（地址）以及多少个 DApp。无论具体采用了什么共识算法，只要交易验证/执行过程随时可能涉及到任何一个用户，那么单台全节点就必须随时保持全网每一个用户、每一个 DApp 相关的状态在内存里面，以供交易验证实时访问。当前所有主流的「Chain of Blocks」的系统，包括比特币区块链、以太坊、EOS 等，都有这个问题，并且同样的，这个问题也是无解的。多级缓存的数据库技术(例如RocksDB)可以稍微改善一下这个限制，使得只有活跃用户受到内存限制，而总用户基数受限于硬盘的容量。但是这并不从根本上解决问题。

「容量」这个问题的关注度远远少于吞吐量，原因很简单：因为吞吐量这个短板还没解决，所以容量问题被掩盖住了。请记住，一旦吞吐量实现了大幅提升，容量问题马上就会出现：在一个高吞吐的系统上，如果用户量上不去，很可能高性能根本跑不满。

一个典型的例子是 EOS。当 EOS 以丧失去中心化特性为代价而解决了吞吐量问题之后，容量的问题马上就凸显出来了。然后，EOS 把账簿容量瓶颈这个问题包装成了一个稀缺资源，并将其代币化，成了EOS RAM 虚拟币。当然除了内存，单台全节点 CPU 也会成为容量的瓶颈，所以也被代币化，成了 EOS CPU 虚拟币。不过，在类金融应用场景中，通常计算复杂度非常低，所以，内存会是主要瓶颈。

另外，我的另外一个观点是：**共识算法其实帮不了解决性能和容量的瓶颈，试图从标新立异的共识算法出发，提升「Chain of Blocks」系统性能的努力，基本上不会让系统性能有实质上的大幅提升。**总之，解决上面所提及的两个瓶颈问题，需要的是分布式系统设计上的巧思妙想，这和共识算法相关，也和密码学相关，但是本质的出发点不是共识算法和密码学。

## 2 性能瓶颈: 一个出块节点在做什么

首先出块节点也是全节点，接受全网的已确认区块以及未确认交易，并构造成链，不断维护账簿的最新状态，然后抓紧机会试图在链尾追加新的区块。无论采用哪种共识算法，都会历经以下几个步骤:

- 第一个步骤，根据账簿的最新状态，在未确认交易集合中选出若干验证合法的交易，然后构造一个新的区块；

- 第二个步骤，为这个新的区块，参与出块的权力的竞争或者候选，在这个阶段，大概率会因为账簿状态更新了(其他节点成功出块了)而中断，回到第一步；

- 第三个步骤，获得出块的权力之后，向全网广播这个新的区块，更新账簿状态，回到第一步。

不同的共识算法，其核心差异在于如何完成其中的第二个步骤的出块权的竞争或者候选。但是无论哪种共识算法，都有一个不可调和的性能矛盾，本质上由区块数据的广播延迟导致。这个矛盾使得如果每次出块比较大(可以包含更多的交易)，就必须有比较长的出块间隔，以保障该区块在下一次出块之前，在全网被充分传播。

如果传播不充分，在 PoW 和 PoS 系统中，将表现为较高的分叉率（出了无效的块），而在 BFT 系统中则表现为较高的失败率（区块拿不到2/3的同意票）。

**2.1 Proof-of-Work 和 Proof-of-Stake**

PoW 通过设定一个 Hash Target，要求 Hash 值必须小于一个特定的值（例如，将256位的 Hash 值当成一个大整数看待）。而 Hash 值必须根据新区块数据拼合一个 Nonce 数据计算而得。找到满足 Hash Target 对应 Nonce 的任何一个节点，便获得了出块的权力。由于只能通过随机穷举的方式找 Nonce，所以这个竞争就转换成了计算 Hash 的算力的竞争。PoS（如Peercoin）是PoW的一个变种，引入了消耗 Coin Age 来增大 Hash Target 的机制，使得出块权力的竞争可以部分地被数字货币持有的时间和数量所代替。

可以看到，PoW机制最大的好处是用一个简洁的算法，实现了完全非许可（permissionless）的出块权随机指定，竞争节点之间完全不需要协同和通讯，可以轻松支持任意数量的出块节点共同竞争，具有极佳的去中心特性。也正是由于这一点，这个算法导致了区块广播延迟和出块间隔之间的矛盾。当出块间隔较短时，一个新的区块尚未充分全网广播之前，就有另一个矿工在同样的高度出了另一个新的区块，即发生了所谓的分叉(Fork)。这种情况下，最终其中一个区块会被抛弃掉(orphaned)。发生这种情况的概率不能太高，否则会显著降低原为51%的算力攻击基准(Selfish Mining)，极端情况甚至会导致分叉始终无法到达稳定收敛。

区块广播延迟主要由区块大小和全网各个节点间的带宽决定。当前的互联网环境，大致需要 10 秒可以广播到 90% 以上的节点。所以在比特币网络中，10分钟左右的出块间隔使得区块分叉的概率极其低。2018 年整个上半年，仅出现两次分叉。而在以太坊网络中，15 秒左右的出块间隔使得区块分叉的概率始终保持在 10% 左右，即使其区块远小于比特币的区块。要注意一点，PoW的出块间隔是统计意义上的，实际情况是出块间隔时大时小，而统计期望是10分钟。这个并不是全网算力波动造成的，而是因为搜索Nonce的过程是个随机刺探过程(撞大运)，所以很多矿池都给出了自身的运气值曲线，(笑...)。

对于比特币网络来说，10 分钟的出块间隔其实在现今的互联网环境中是有很大保留的，要知道，毕竟这是在 10 年前提出的方案，这使得扩大区块大小就可以实现简单的扩容方案，但是由于区块广播延迟这一根本矛盾的存在，这种提升只在一定程度上有效。

另外，值得提一下 GHOST 协议。该协议给出了一个新的准则来判定分叉的时候，哪个叉是被接受的。其将中本聪最初提出的最长链原则， 改成了包含算力最多的子树。两个准则在分叉概率很低的时候是完全等价的，但是当概率比较高的时候(比如ETH的10%分叉率)，GHOST 协议可以规避 Selfish Mining，提高安全性。但是无论采用 GHOST 协议与否，对公链的性能无实质帮助。

**PoW 带来算力竞争，即所谓的挖矿，确实消耗了大量能源。不过这也为 PoW 系统发行的每一个币奠定了一个基础成本，使之价值有个底线。**需要指出的是，PoW 的算力和区块链系统的性能没有任何联系，任何加速 hash 算法的软件或者硬件都不会提高区块链系统单位时间的吞吐量。这就是为什么比特币区块链的全网 hash 算力提高了万亿倍，但是其吞吐量一直是 7 TPS 左右。

另外，**任何宣称节省挖矿能源的公开技术，都是不可能在实际上减少能源消耗的。**因为投入挖矿的能源总量在一个个矿场建立的时候已经确定，当有更高能效的挖矿技术或者设备出现时，算力竞争将导致所有矿工都应用新的技术，最终哄抬了全网的挖矿难度罢了。所以实际的总能源消耗，在宏观上，只和币价、电价以及数字货币的投资信心相关，和挖矿效率无关。

**2.2 拜占庭容错 （BFT）**

拜占庭容错类共识算法采用随机算法确定每一次出块的节点，根据账簿上的数字货币地址，而不是IP地址。所有参与出块候选的节点无须竞争。新的区块将被委员会(一组验证者)所有成员验证并签名(投票)，然后广播全网，继而开始下一个出块的流程。

与 PoW 不同的是，BFT出块候选是一个协作的过程，期间至少涉及 O(n^2) 的通讯复杂度，而 PoW 在出块竞争过程中无须任何通讯代价。基于 BFT 的协作过程将不会导致分叉，也不需要消耗稀缺资源(算力或者Coin Age)，但是由于这个协作的过程涉及到相当多的数据通讯，所以这个过程无法在全网候选，验证并签名的过程无法在全网展开。这就是为什么 BFT 类算法一定会涉及到一个委员会的构建过程，并且验证签名只在一个小范围里面发生，剩下的人相信他们就好了。最近出现的很多基于 BFT 的公链项目，比如 Algorand，在如何安全公平的选出这个委员会方面做了很多工作，虽然这些工作对系统性能的提升没有直接关系。

BFT 类算法的投票通常是有权重的，以规避女巫攻击 (Sybil Attack)。而这个权重多与参与者的权益相关，和PoS的精神类似，进而现在很多人将 BFT 的这类投票算法称为了 PoS 算法。而事实上，BFT 类共识算法和一开始提出的 PoS 算法(例如 Peercoin)是本质不同的机制。

上面我们提到，不同的 BFT 类算法其具体选定出块节点以及委员会成员的过程和系统的性能关系不大。和 PoW/PoS 类似，其吞吐性能同样决定于每次出块的大小，以及出块的周期。在 BFT 系统中，如果想要允许每次出块比较大，就需要出块的周期也比较大，从而大概率保证新出的块及其委员会的签名数据在委员会内部完全传播。如果这个传播不充分，将可能导致委员会成员无法达成 2/3 以上的投票，进而使得委员会内部验证并签名过程超时，最终在本出块的周期内出块失败。

理论上说，委员会的规模远小于全网，BFT 类算法中的广播延迟会比同等规模的 PoW/PoS 网络小。事实上也确实如此，但是基于 Gossip 协议的广播延迟和网络规模的对数成正比而不是线性，所以广播延迟并没有小很多。加之 BFT 类算法依赖一些额外的周期性全局同步等安全措施，使得实际效果中，BFT 类算法并没有比 PoW/PoS 系统有太多性能优势。

**2.3 小结**

PoW/PoS 系统每个出块周期需要充分传播一个区块 (例如 1MB), BFT 类系统每个出块周期也需要充分传播一个区块，加上所有委员会成员的签名 (例如 128个成员节点，至少每人84\*128B，总共1.3MB的样子)。但是 PoW/PoS 系统的广播范围是全网(例如几万个节点)，而 BFT 系统的范围限于委员会成员，这一点使得后者充分传播的时间较短一些。

不过，基于Gossip协议实现充分传播的时间，和传播的数据量呈线性关系，和传播的节点数量呈对数关系，所以 BFT 在传播时延上也没有太大的优势。结果就是，无论哪种算法，都有不可调和的区块大小和出块间隔之间的矛盾，从而无法大幅提升性能。

## 3. 容量瓶颈: 一个不出块全节点在做什么

单链的「Chain of Blocks」的系统中，大致有三种角色的节点: **出块的全节点，不出块的全节点和轻量节点。**全节点无论出块与否，都会验证并接力广播新的区块和未确认交易，这里的广播工作占据了主要的通讯量以及磁盘 I/O 的负荷，对于 TPS 只有十几的以太坊 (geth) 来说，这个通讯量约为 1.5Mbps。

为了可以实时完成对新区块和未确认交易的验证，所有用户的账簿以及所有智能合约状态都需要驻留在内存中，这个占据了主要的内存开销，当前规模的以太坊会占用将近4GB的内存。每一个全节点都会需要承担这样的一个负荷，如果要出块(PoW 的挖矿节点或者 PoS 的验证节点)还需要做额外的事情。这些负荷的代价，换来的是安全的彻底去中心化，任何一个全节点不需要预先信任任何其他节点，任何全节点也没有能力去欺骗其他全节点。

普通全节点的价值体现在两个方面：接力广播合法的数据和维护全网账簿的最新状态以供用户或者轻量节点查询。例如手机钱包这样的轻量节点不验证也不接力广播区块数据或者未确认交易，它依赖并信任预先设定好的一个或者多个全节点，通过这些全节点来获取特定用户的状态，例如账户余额，以及发起转账交易。轻量节点自身完全没有验证信息真伪的能力，更像是区块链世界里的一个终端而已。

对于单链的「Chain of Blocks」的系统，如果系统的吞吐量(TPS)提升 100 倍，需要 150Mbps 的通讯量；或者用户规模都扩大 100 倍，需要 400GB 的内存，那么基本上大部分互联网上的普通服务器都无法顺利部署一个全节点了。**全节点的参与门槛，是影响区块链系统去中心化程度重要因素。**如果全节点只能由专业矿场操作，普通人无法独立部署一个全节点的话，那么整个系统就会退化成一个多地部署的中心化云服务了，而变得容易被攻击，也容易被封禁。所以，这两个瓶颈不仅仅对于出块节点需要解决，对于普通全节点也需要解决。

## 4. 出路

前面已经说到性能瓶颈和容量瓶颈，在现在单链的「Chain of Blocks」的系统中，很难有大的提升，尤其是容量瓶颈。这就是所谓的区块链[不可能三角](https://www.8btc.com/article/7836)的由来。纵观计算机技术发展史，大容量高吞吐的设计范式，屡获大规模成功的只有一个，横向扩展(Scale-Out)。

举个例子，GPU 用了几千个性能普通的 Core 一起并行工作，实现超越 CPU 计算性能几个数量级的性能提升，而 GPU 所依赖的半导体技术并没有和 CPU 芯片有什么本质的不同。再如，现今的在线云服务系统，是用几千甚至上万台性能普通的服务器一起并行工作，来支持大容量高吞吐的在线服务。

我在这里不妨大胆设想：也许一个大容量高吞吐的区块链系统会是类似的方案，即，**让成千上万个同质的单链实例一起并行工作，切分全网的工作量，以实现整体上的大容量和高吞吐。**

这样的一个系统，可以在大幅提高 TPS 的同时，支持 10 亿以上级别的用户量，并且保持每一个参与到这个网络的中的全节点仅有一个合理的负荷，让大部分互联网上的普通服务器都可以轻松部署一个全节点，共同参与网络的维护和治理。

不过，在这样一个彻底去中心化的设定下，如何一起并行工作，如何切分工作量，又如何保证每个单链实例的安全，都是极具挑战的问题。这样的系统似乎并不容易实现，但也绝非不可能实现。我先抛出这个想法，也欢迎所有有兴趣的同仁共同思考，或批判，或贡献聪明的设想。就这个想法，我也会继续梳理，并继续通过文字分享我的一些思考。欢迎大家通过我的公众号「王嘉平」和知乎专栏「去中心化数字世界随想」就这个话题展开更多讨论。