高性能区块链系统原理及构建

陈鹏飞

中山大学

数据科学与计算机学院

中山大學
SUN YAT-SEN UNIVERSITY

区块链回顾

区块链系统中分布在全球各地的一个个节点；而这些节点可以简单理解为一台服务器服务器集群；在一个典型的P2P系统加入货币激励。

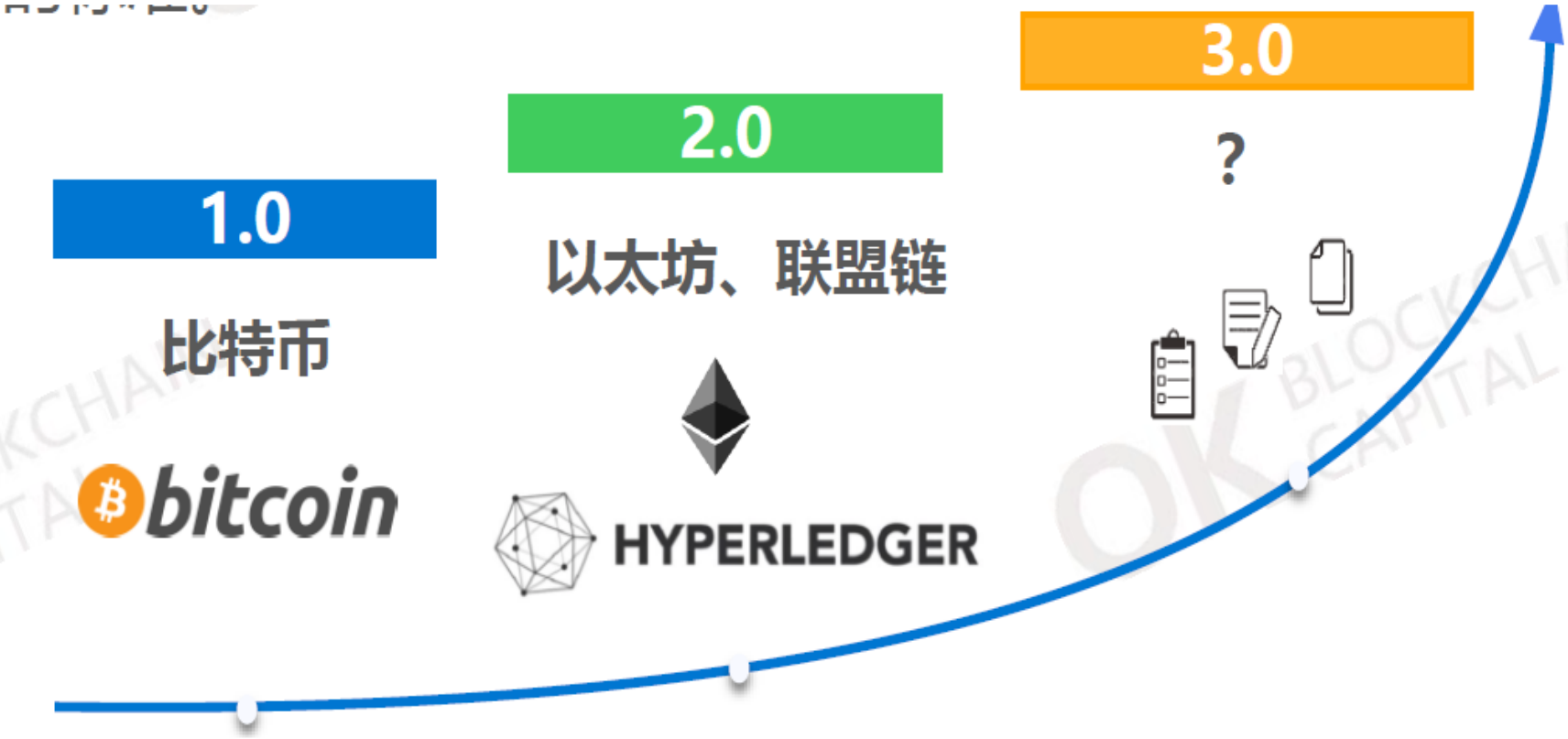
区块链回顾

区块链系统的基本角色：物、人、事

现状

投资	Santander → ripple $32m Goldman Sachs → CIRCLE $50m VISA → chain $30m	• First funding rounds for three startups raised a total of $102m • Goldman Sachs acts as a VC in the Bitcoin's Block chain space • Santander InnoVentures has created a $100m investment fund, • Visa joined a syndicate to invest $30m in Chain
内部开发	Nasdaq, IBM citibank	• NASDAQ in-house platform described on p.40 • Citibank has created citicoin, a cryptocurrency being tested internally • IBM is working on a blockchain platform for the Internet of Things
合作伙伴	earthport, WESTERN UNION WU moving money for better, fidor BANK ↔ ripple UBS ↔ clearmatics BARCLAYS ↔ safello	• Ripple has attracted many partners for its real-time payments system • UBS and Clearmatics partnered to create a securities settlement system • Barclays partnered with Safello to deliver "proof-of-concepts"

区块链系统发展



区块链3.0：侧链与跨链

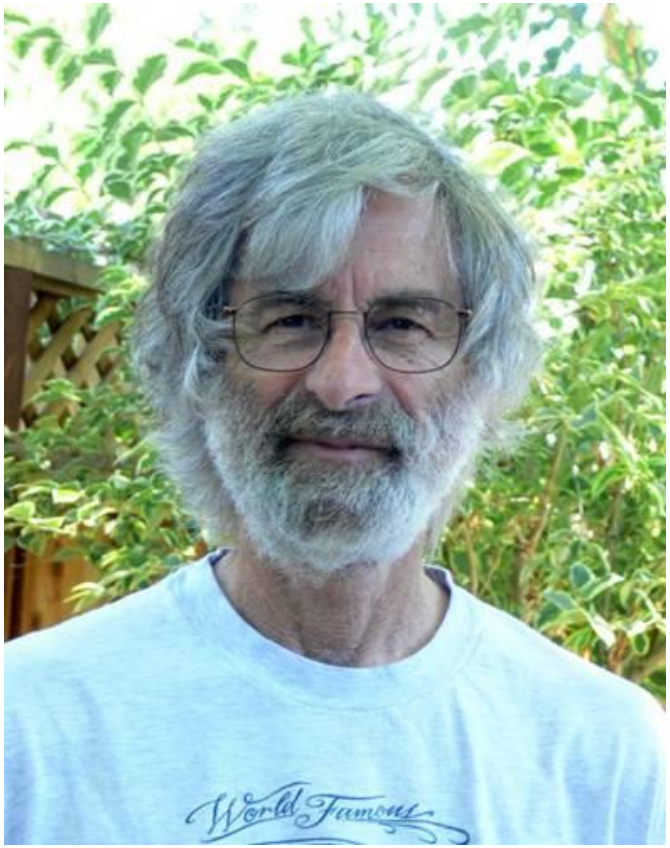
OK BlockchainCapital

区块链展望


2005年以前
目前
2025年以后
&
&
封闭和集中的物联网
以开放方式接入物联网，集中云
以开放方式接入物联网，分布云

状态机复制

Leslie B. Lamport



[PDF] Using Time Instead of Timeout for Fault-Tolerant Distributed Systems

https://www.microsoft.com/en-us/research/uploads/.../using-time-Copy.pdf ▾ 翻译此页

作者：L LAMPORT - 1984 - 被引用次数：397 - 相关文章

A general method is described for implementing a distributed system with ... use of absolute times instead of timeouts, and can be considered an extension of.

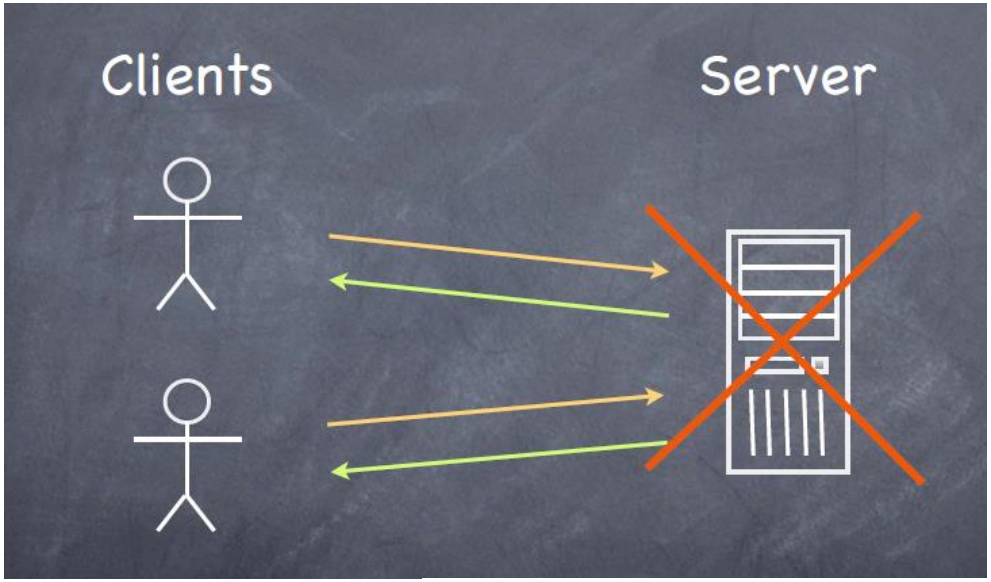
Paxos!

状态机复制

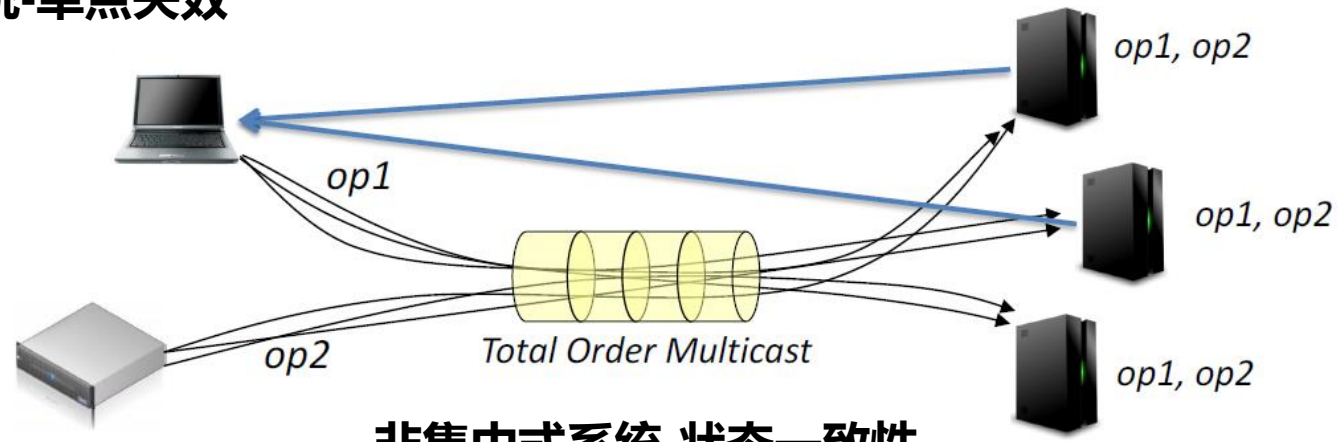
Fred Barry Schneider
康奈尔大学教授

[PDF] Implementing Fault-Tolerant Services Using the State Machine ...
https://www.cs.cornell.edu/fbs/publications/SMSurvey.pdf ▾ 翻译此页
作者：FB SCHNEIDER - 1990 - 被引用次数：2384 - 相关文章
Implementing Fault-Tolerant Services. 303 state machine and running a replica on each of the processors in a distributed sys- tem. Provided each replica being ...

状态机复制

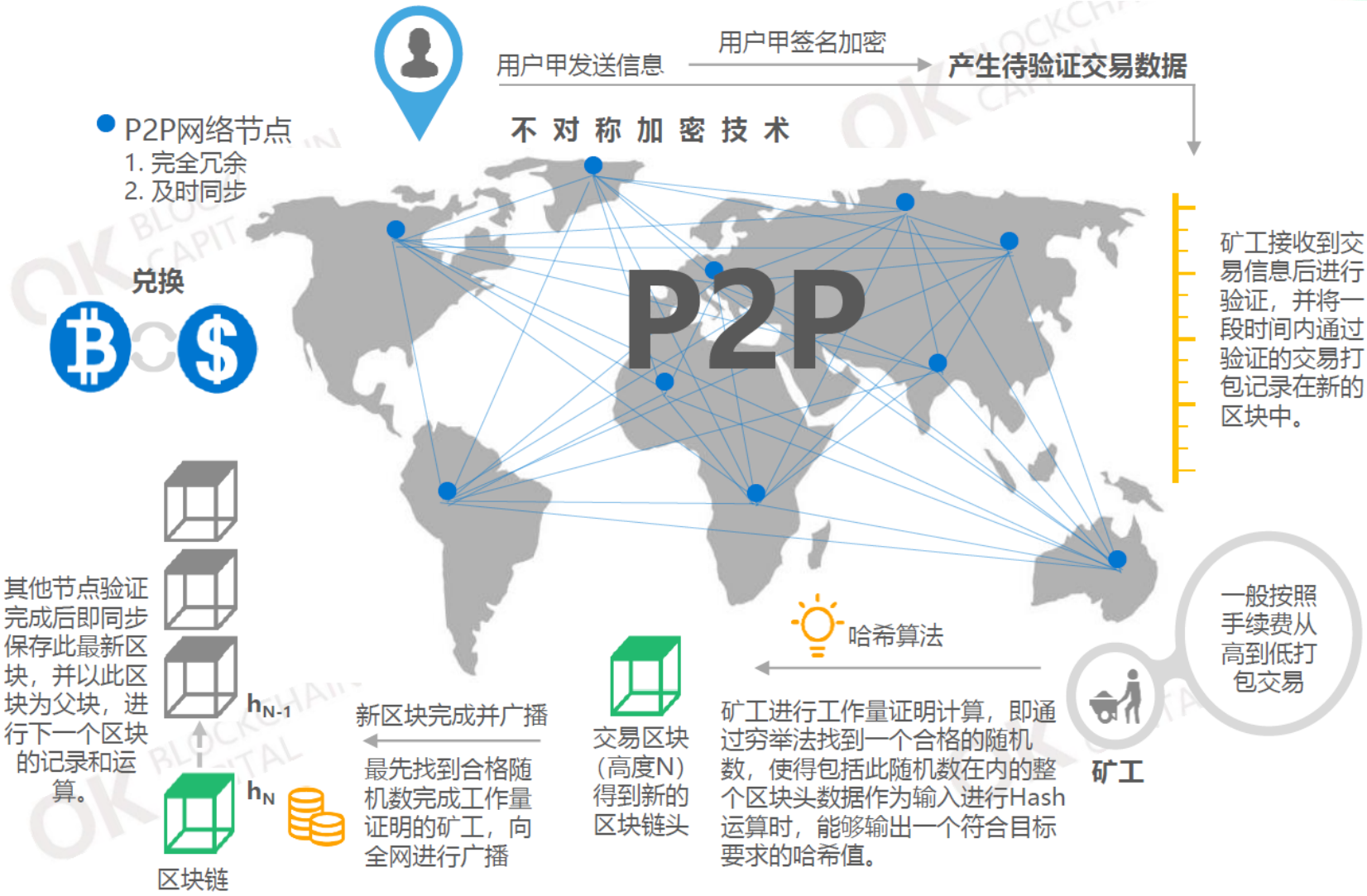


集中式系统-单点失效

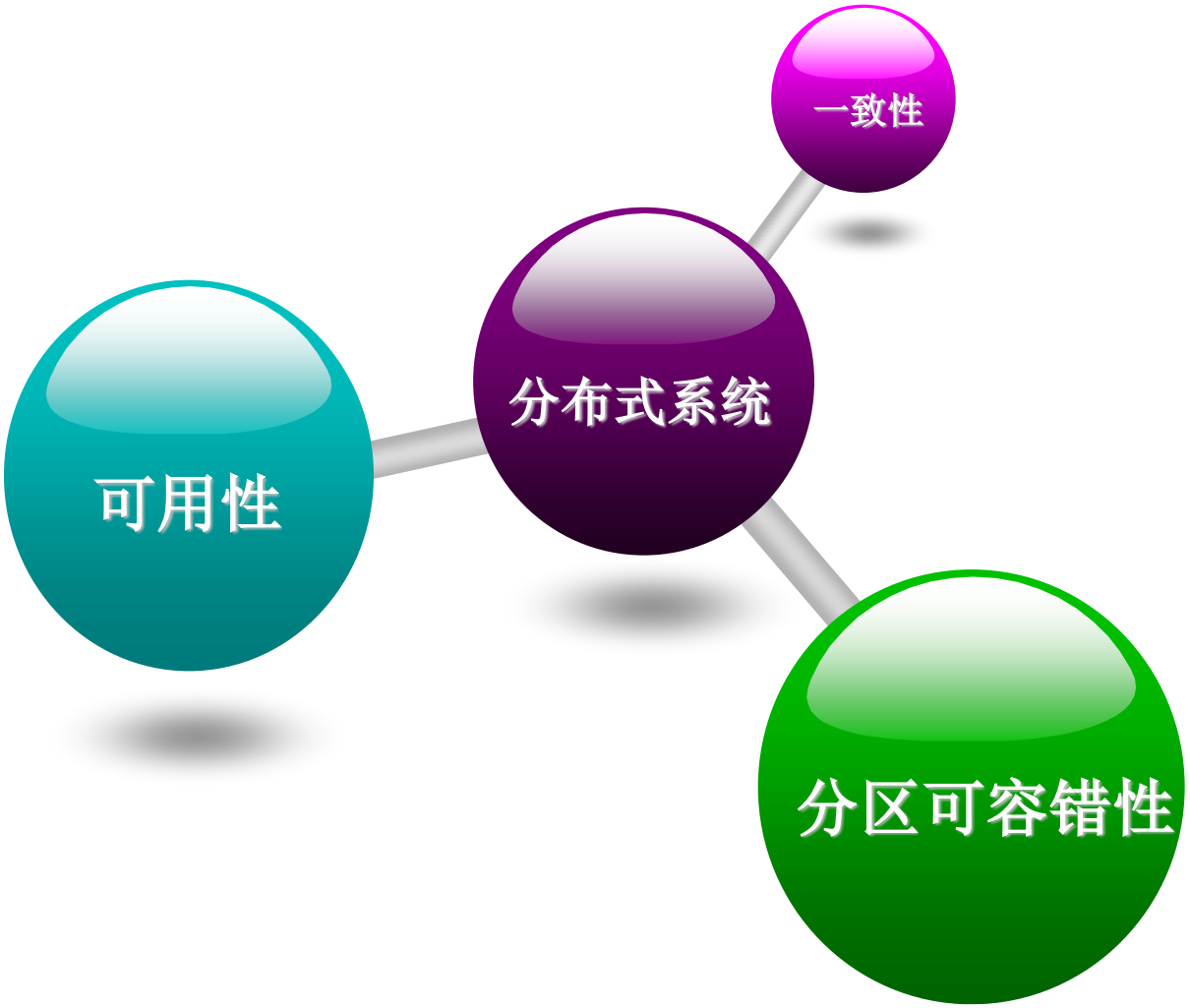


非集中式系统-状态一致性

区块链交易系统

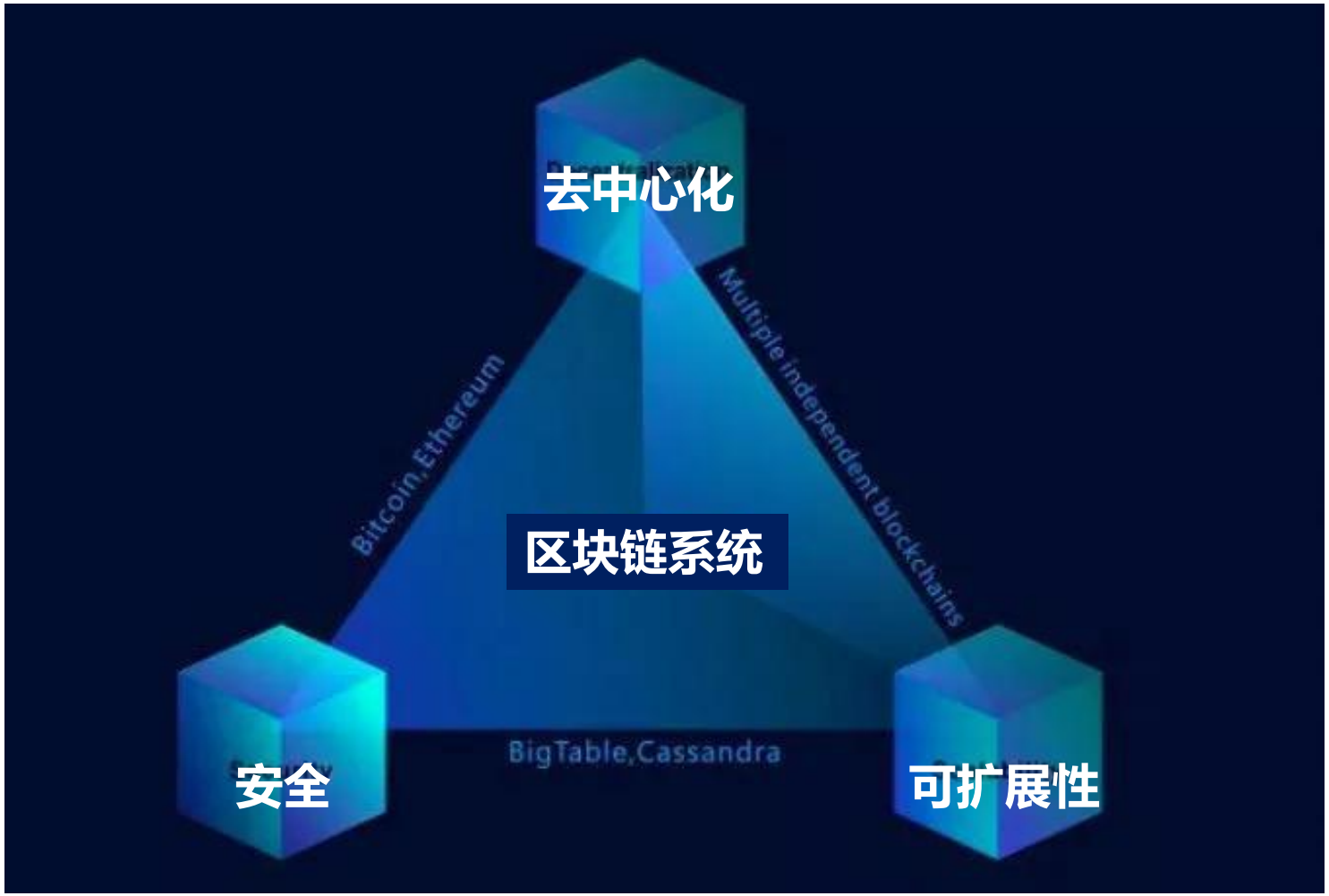


分布式系统"不可能三角"理论



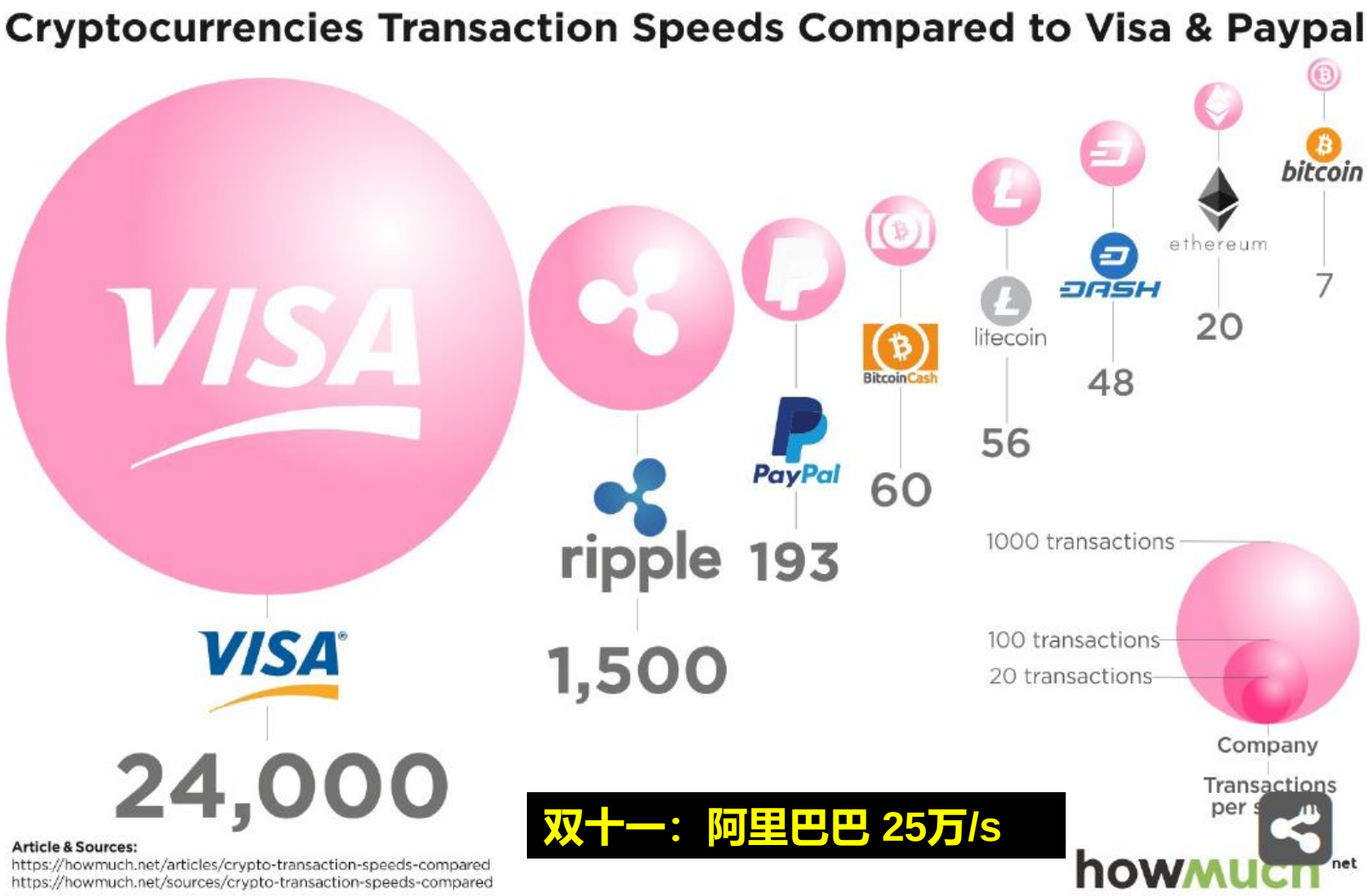
CAP原则又称CAP定理，指的是在一个分布式系统中，Consistency（一致性）、Availability（可用性）、Partition tolerance（分区容错性），三者不可得兼

区块链系统“不可能三角”理论



区块链系统同样受制于著名的理论：“不可能三角理论”，即安全性、去中心化和可扩展性三者不可兼得

电子货币的交易性能



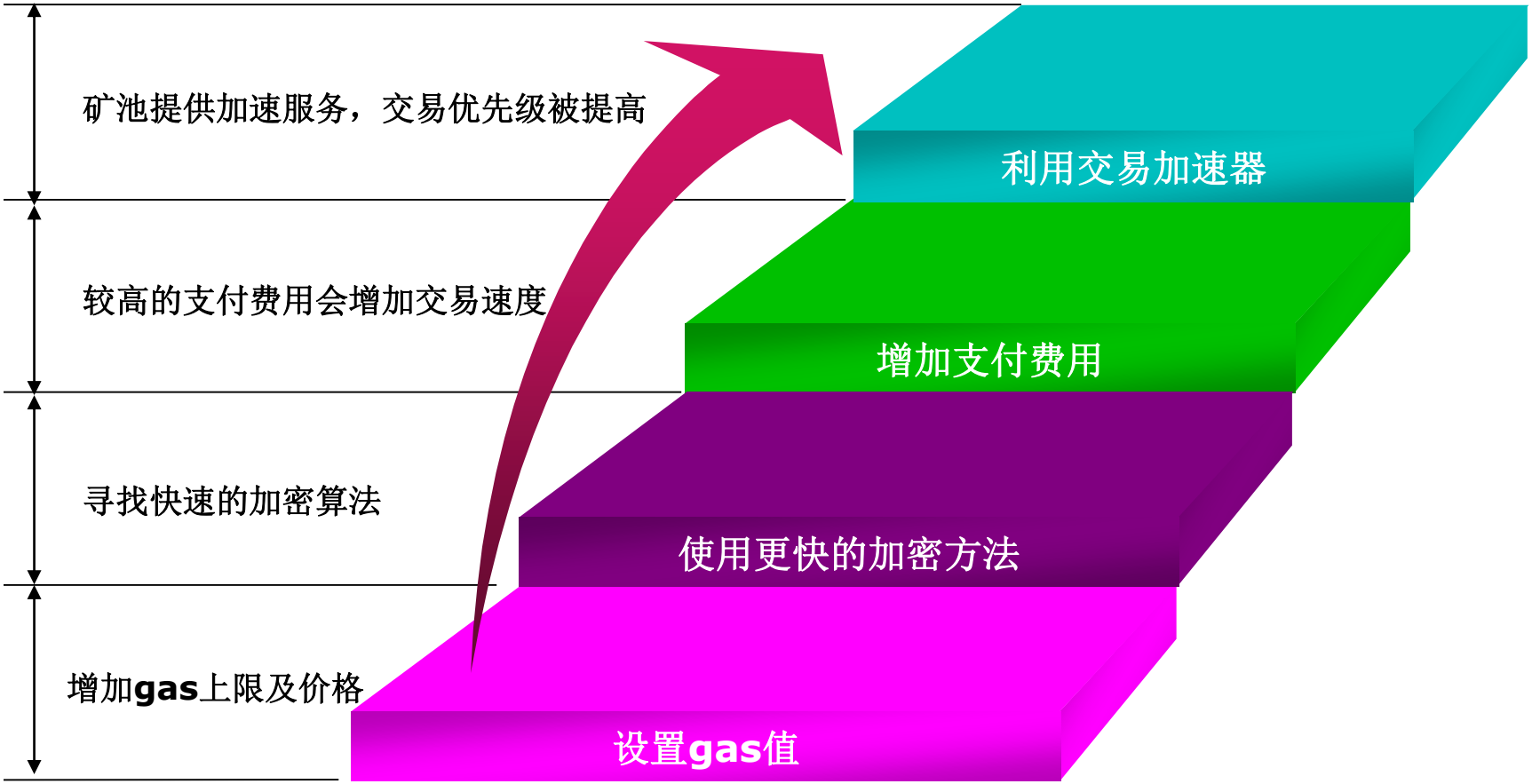
双十一：阿里巴巴 25万/s

电子货币的交易性能

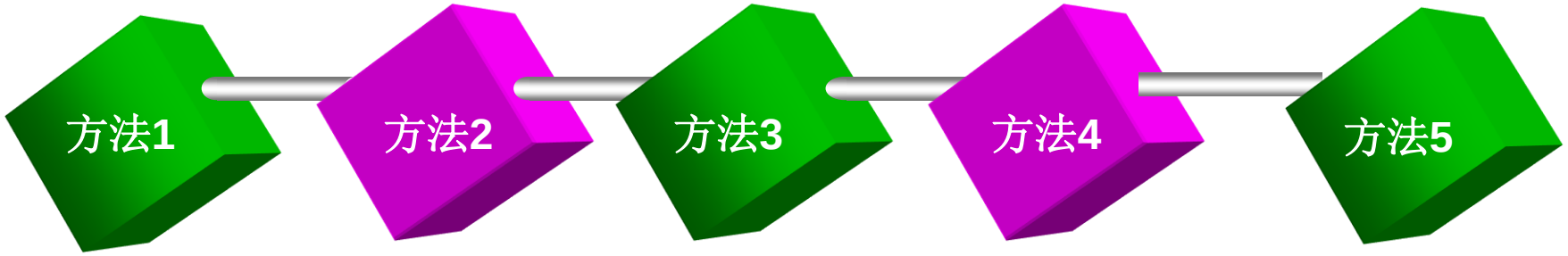
测量标准：统计交易被确认的时间

电子货币	交易性能	电子货币	交易性能
Bitcoin	单个交易时间平均78分钟；	Tron	单个交易时间平均15分钟；
Ripple	单个交易时间平均4秒；	Cardano	单个交易时间平均5分钟；
Litecoin	单个交易时间平均30分钟；	Iota	单个交易时间平均3分钟；
以太坊	单个交易时间平均6分钟；	Zcash	单个交易时间平均15分钟；
EOS	单个交易时间平均1.5秒	Monero	单个交易时间平均30分钟；
Augur	单个交易时间平均6分钟；	Steem	单个交易时间平均3秒钟；
Dash	单个交易时间平均15分钟	Ardor	单个交易时间平均1分钟；

以太坊加快交易的方法


矿池提供加速服务，交易优先级被提高
较高的支付费用会增加交易速度
寻找快速的加密算法
增加gas上限及价格
利用交易加速器
增加支付费用
使用更快的加密方法
设置gas值

高性能高扩展性区块链系统原理

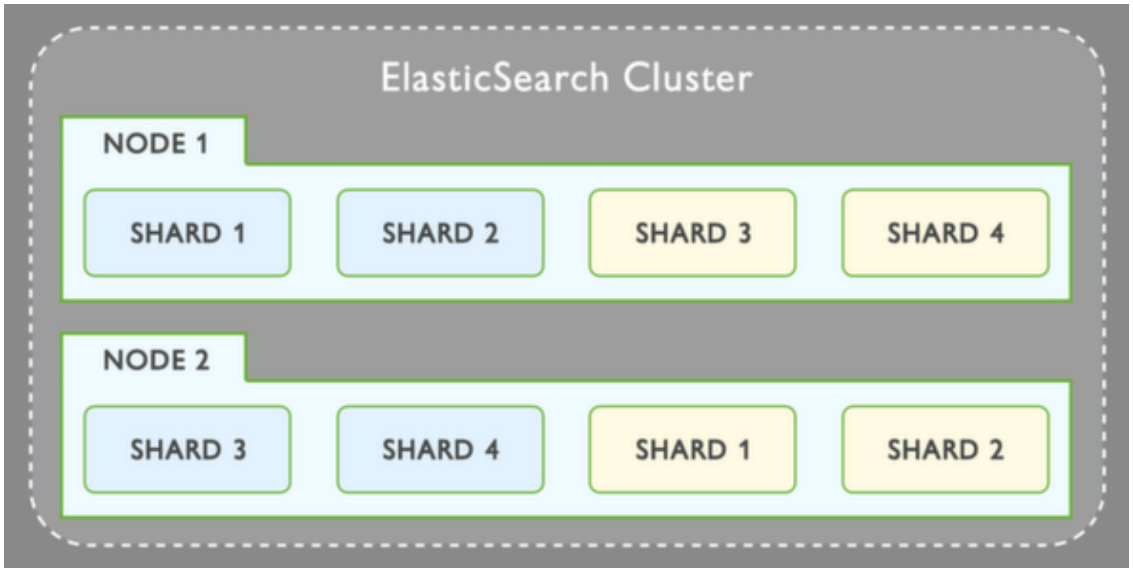


分片（Sharding）	离链计算（Off-chain computing）及侧链（Sidechain）	扩展的共识协议（Scalable consensus）	有向无环结构（DAG）	优化系统实现方法

高性能高扩展性区块链系统原理

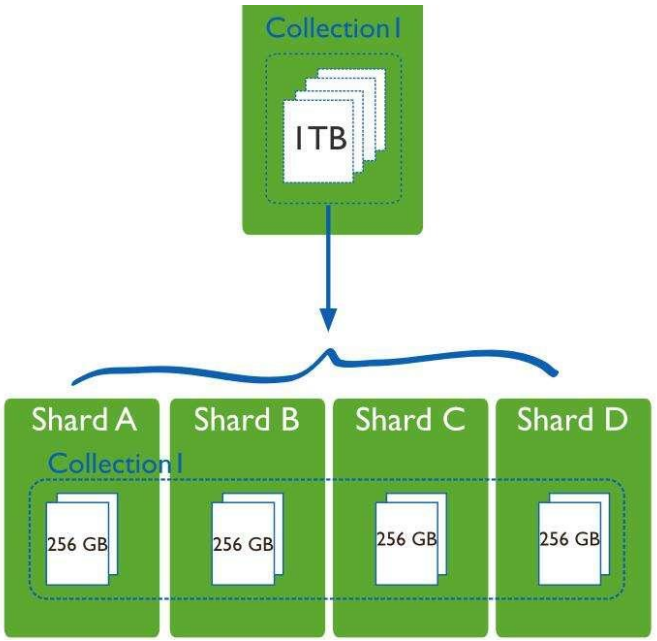
分区（Sharding）：
一种数据库技术，将大型数据分成更小、更快、更容易管理的部分

分片（Sharding）



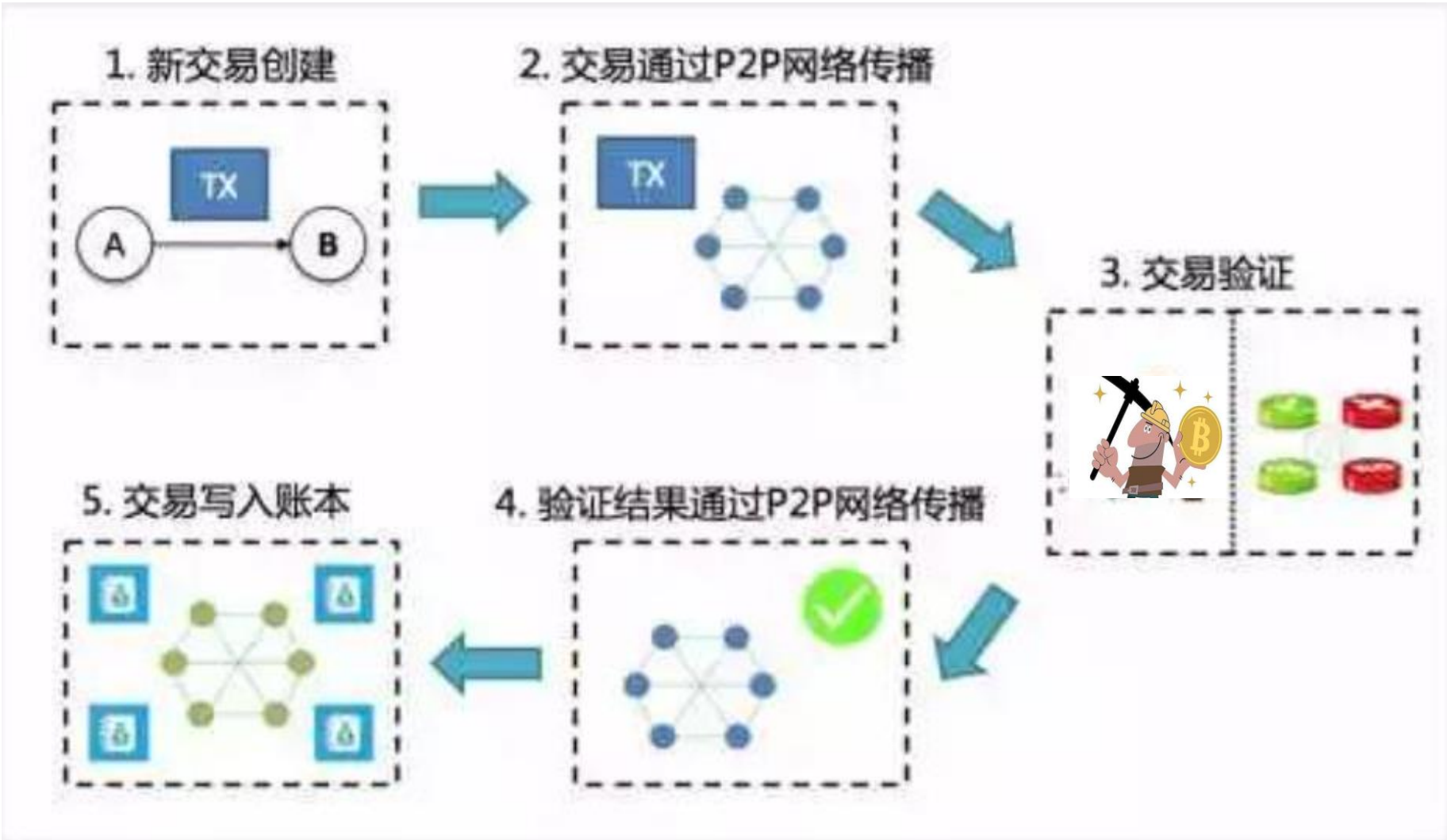
Elasticsearch sharding

- 分片（sharing）其实是一种传统的数据库技术，它将数据库分割成多个碎片并将这些碎片放置在不同的服务器上，用于提高**性能和可用性**。



MongoDB sharding

回顾


1. 新交易创建
TX
A
B
2. 交易通过P2P网络传播
TX
3. 交易验证
4. 验证结果通过P2P网络传播
5. 交易写入账本

区块分片（Sharding）

- 传统区块链单一的区块打包队列；
- 将验证、打包的过程并行化；
- 交易量提高到千笔每秒；
- 交易费用减少；

分片策略

- **网络分片**

  区块链节点的网络被分割成不同的碎片，每个碎片都能形成独立的处理过程并在不同的交易子集上达成共识。 需要保证安全性，方案：**随机性**

- **交易分片**

  根据交易的某些属性划分碎片，比如根据交易记录的Hash值 （**存在的问题？**）或者根据交易账户；

- **状态分片**

  公共节点都承担着存储交易、智能合约和各种状态的负担，这可能使其在为了获得更大的存储空间而进行巨大的花费。状态分片让每个节点只负责托管自己的分片数据，而不是存储完整的区块链状态。**存在的问题?**

分片策略

Luu L, Narayanan V, Zheng C, et al. A Secure Sharding Protocol For Open Blockchains[C]// ACM Sigsac Conference on Computer and Communications Security. ACM, 2016:17-30.

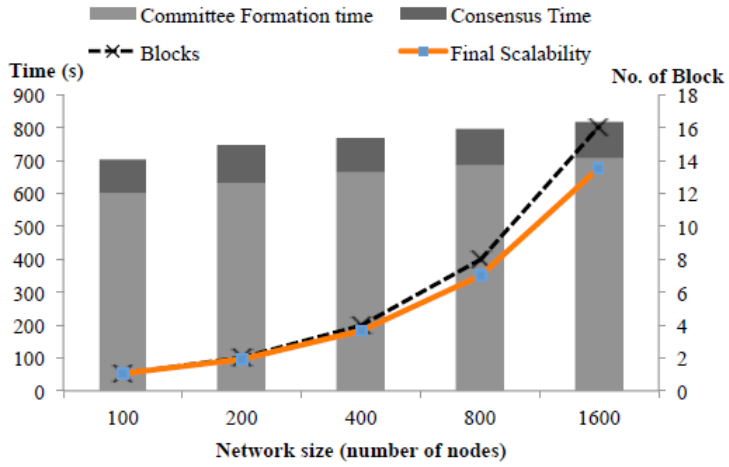
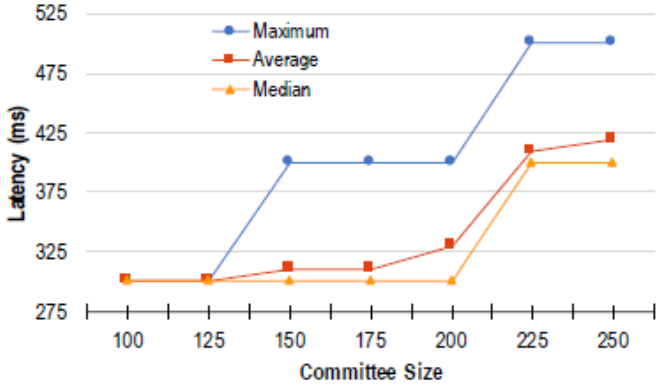
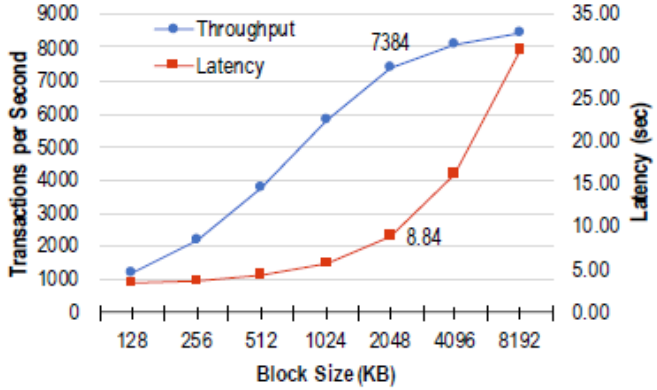


Figure 1: ELASTICO scales up the throughput nearly linearly in the computation capacity of the network.

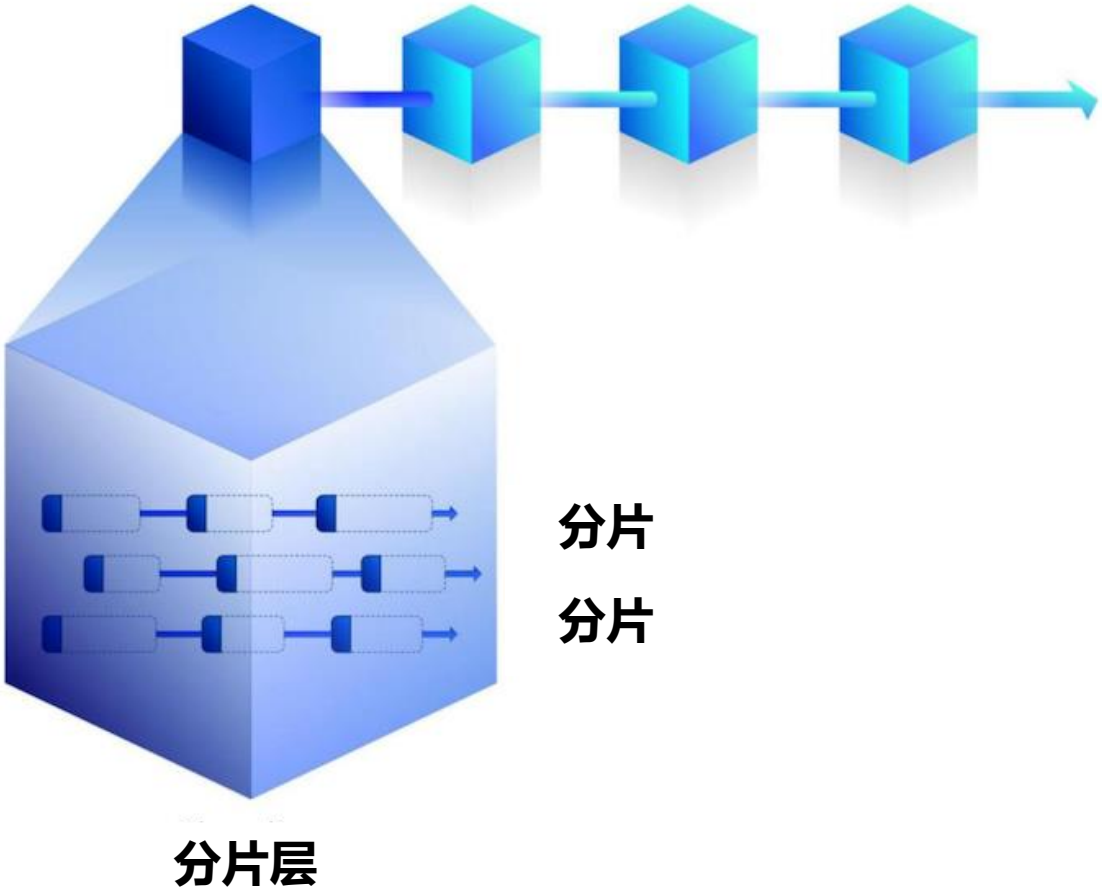
RapidChain: Scaling Blockchain via Full Sharding





夸克链（1）

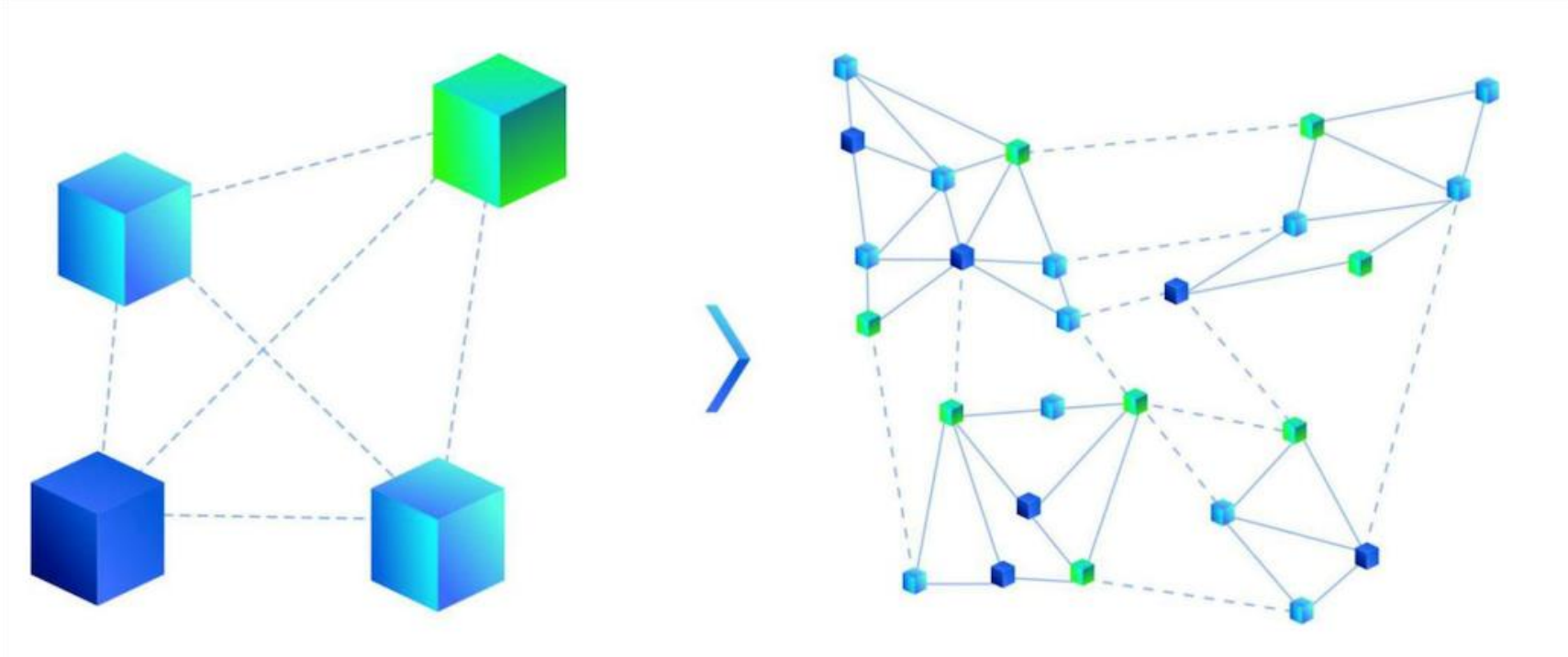
系统架构



其中每条分片处理一个事务的子集合，而根链则通过在根链中包含分区的区块头来确认分区。

夸克链（2）

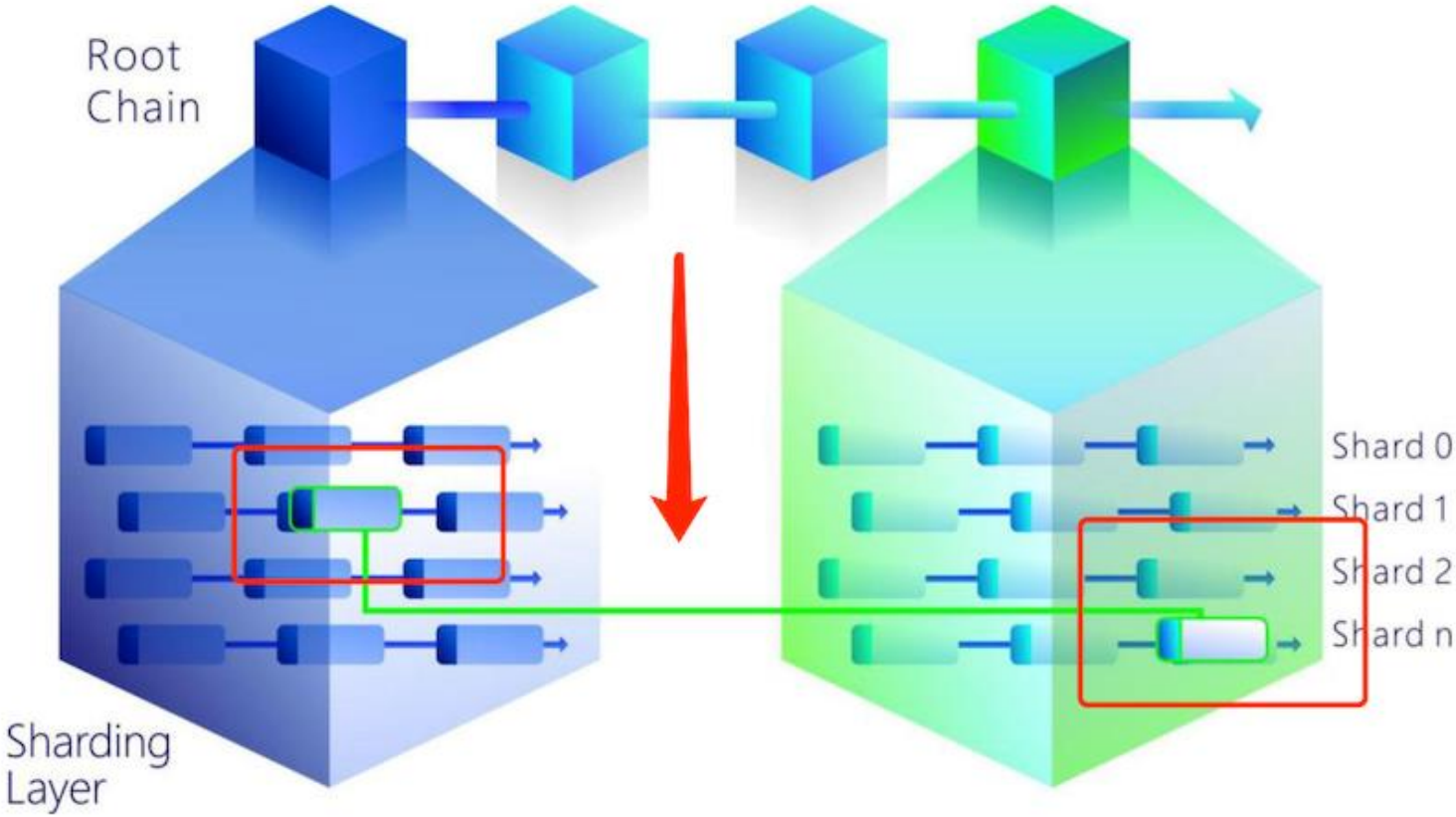
抗中心化横向扩展性



QuarkChain 网络的横向扩展性，其中四个超全节点（左）被四个节点集群（右）取代，其中每个集群中的节点彼此诚实。 （实线表示诚实的连接，虚线表示不可靠的连接）

夸克链（2）

高效、安全的分片交易



只要跨分片交易被根链确认，交易就可以确认

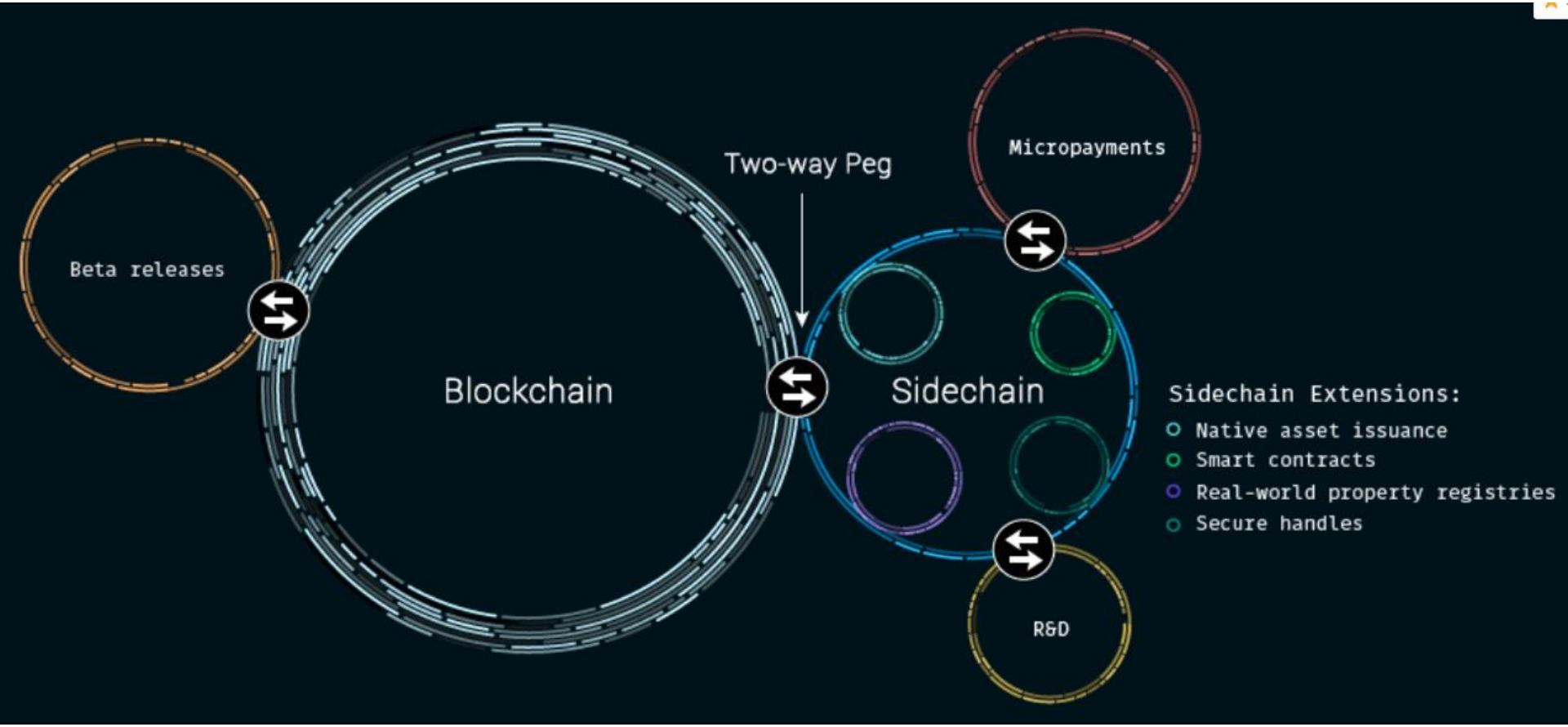
高性能高扩展性区块链系统原理

离链计算/侧链（Off-chain computing）：

在主链之下进行交易，中间交易不进入主链确认，待最后一笔交易完成后回归到主链。

侧链

- 侧链的概念是通过双向锚定实现了主链与侧链之间的价值转移，侧链的目的是为了扩展主链的功能和性能。与区跨链分片机制不同，后者是对主链划分。



区块链3.0：侧链与跨链

OK BlockchainCapital

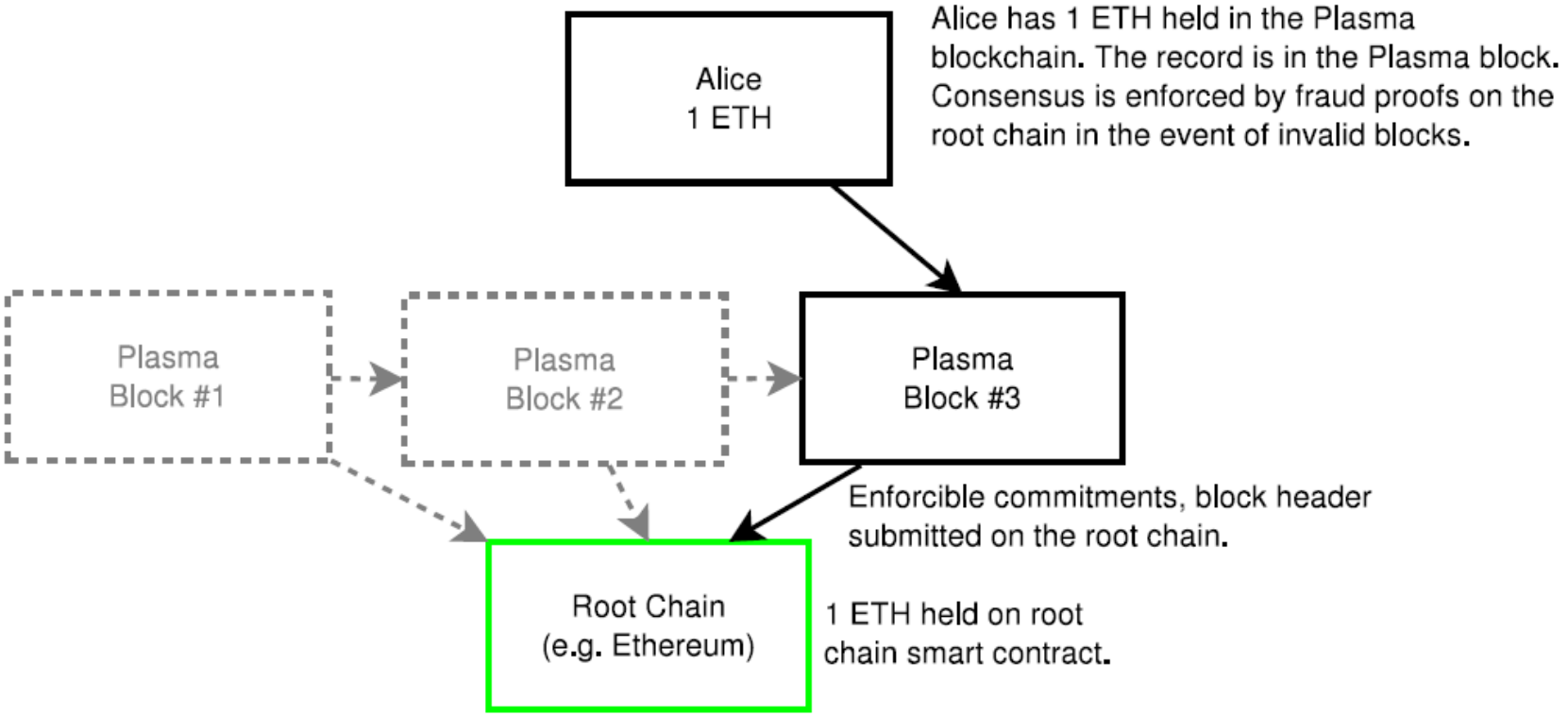
侧链要解决的关键问题

1.**跨链交易验证问题**
链与链之间如何建立一个信任机制，验证跨链之间的交易数据？解决方案有：
a. 公证人机制，b. 区块头Oracle + SPV简易验证

2. 跨链事务管理问题
跨链交易包含多个子交易，这些子交易构成了一个事务。跨链的事务管理又分为两个子问题：
2.1 如何确定子交易是否被最终确认，永不回滚？
解决方案：
a. 等待足够多确认，b. 区块纠缠，c. DPoS/xBFT
2.2 如何保证交易的原子性？所有子交易要么都成功，要么都失败。
解决方案有：哈希时间锁

3. 锁定资产管理问题
资产跨链转移时，锁定资产如何管理？
解决方案有：
a. 单一托管人b. 联盟托管人c. 智能合约托管

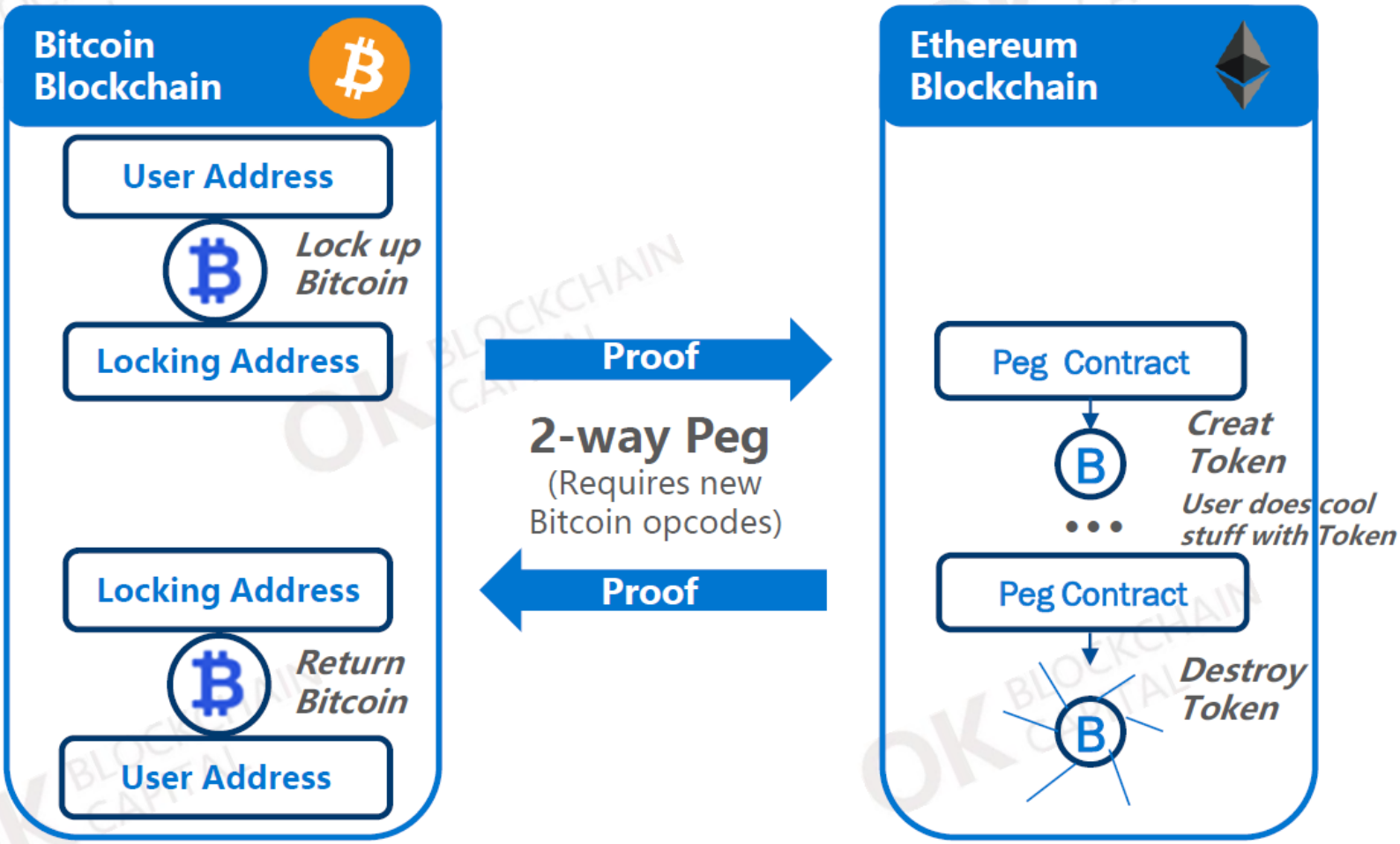
Plasma侧链



Plasma blockchains are a chain within a blockchain.

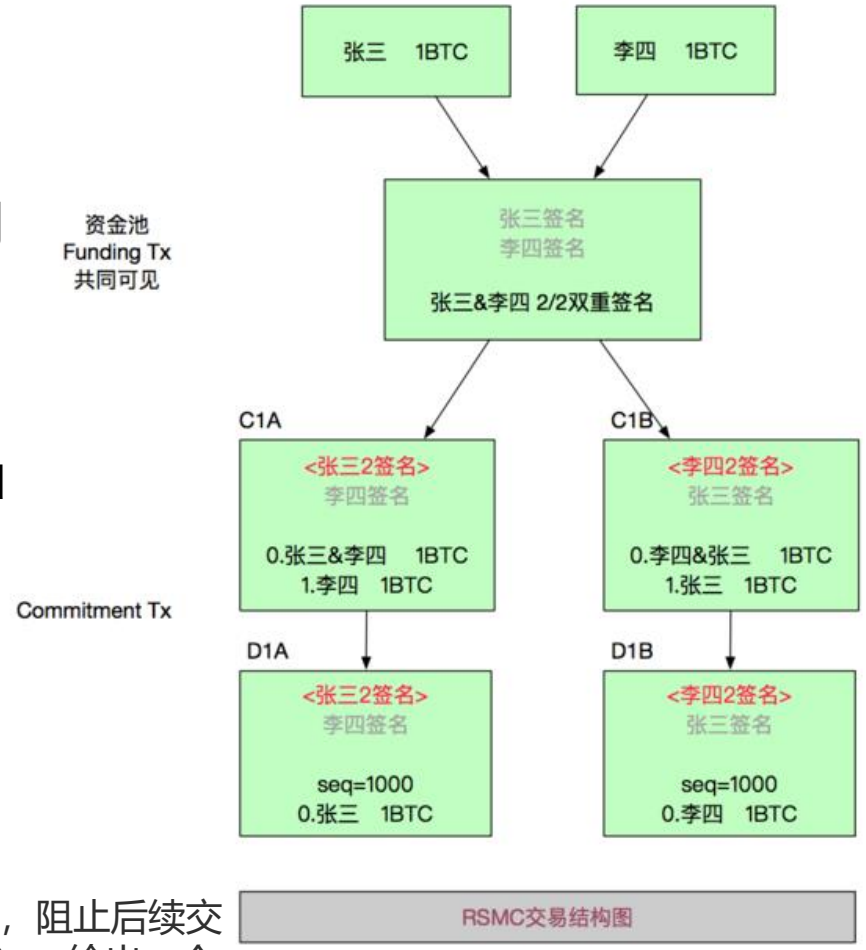
Plasma: Scalable Autonomous Smart Contracts

不同链之间的信息交换


Bitcoin Blockchain
User Address
Lock up Bitcoin
Locking Address
Locking Address
Return Bitcoin
User Address
Proof
2-way Peg
(Requires new Bitcoin opcodes)
Proof
Ethereum Blockchain
Peg Contract
Creat Token
User does cool stuff with Token
Peg Contract
Destroy Token

闪电网络

- 基于微支付通道（双向支付通道）
- 两种类型的交易合约——序列到期可撤销合约RSMC（Revocable Sequence Maturity Contract）和哈希时间锁定合约HTLC（Hashed Timelock Contract）
- RSMC解决了通道中货币单向流动问题，HTLC解决了货币跨节点传递的问题

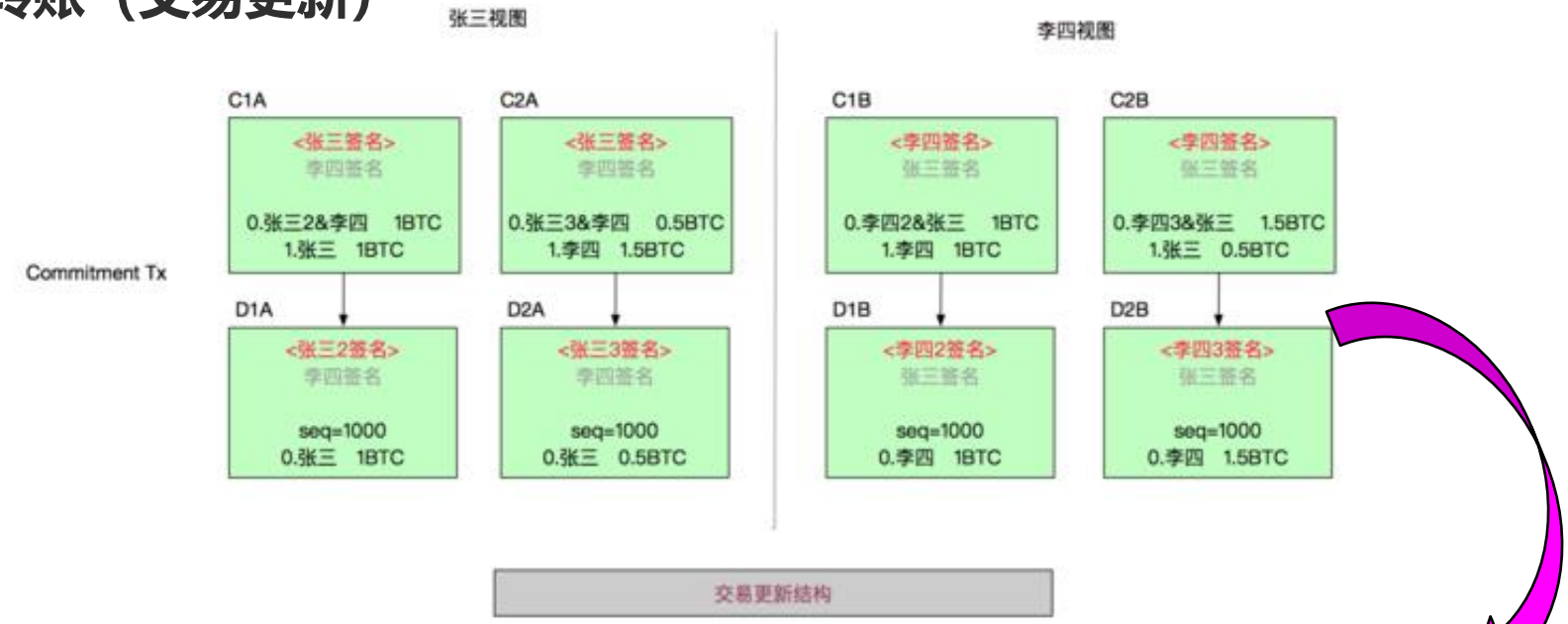


引入sequence，阻止后续交易进块（D1A），给出一个实施惩罚窗口期。

建立微（双向）支付通道

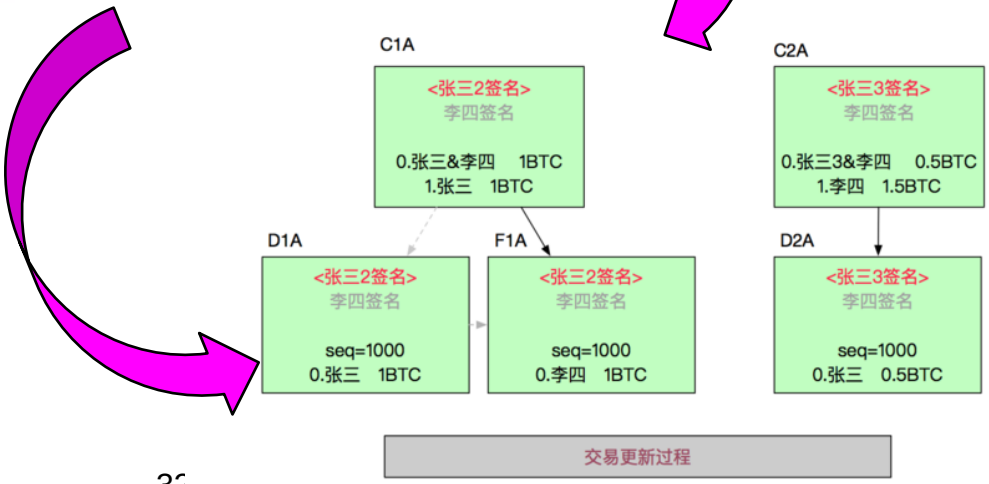
闪电网络

转账（交易更新）



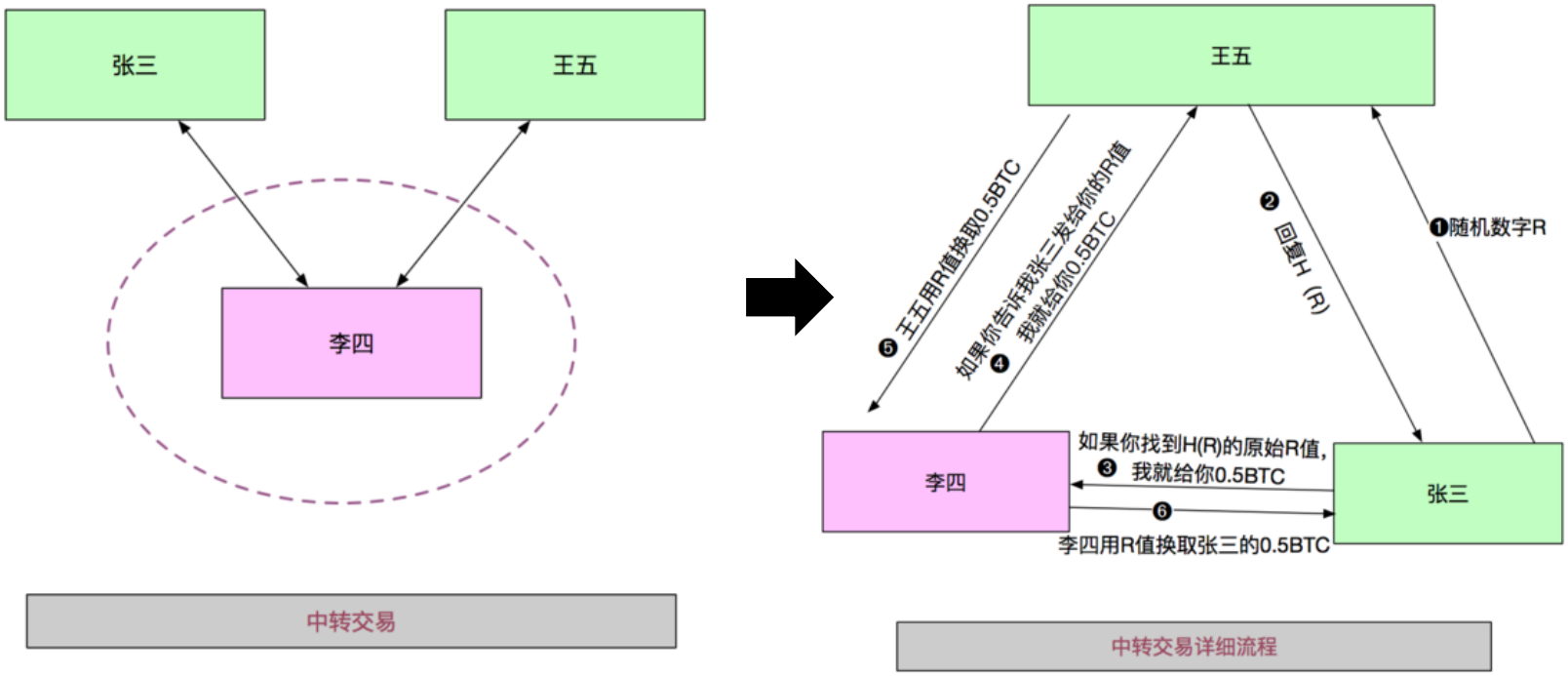
怎样才能彻底废弃掉C1A和C1B呢?

RSMC采用了一个非常巧妙的方法，在C1A的第一个输出中，采用了张三2和李四的多重签名，张三将张三2的私钥交给李四，即表示张三放弃C1A，承认C2A



闪电网络

中转交易



张三通过李四与王五进行中转交易

闪电网络

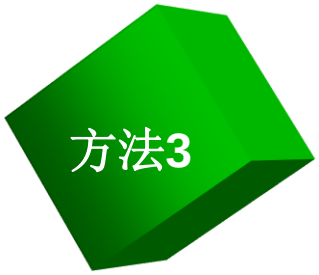
RSMC通过巧妙的设置Commitment TX的多重签名输出，以及sequence的延迟进块形成惩罚窗口期，解决了在微支付通道中的货币单向流动问题，而哈希锁定合约HTLC则解决了货币跨节点的传递。

闪电网络是一种链下交易，只有关键的环节才会发送到比特币主链，非关键环节则在链下（闪电网络上）进行计算、运行，这样大大降低了主链负荷。

构建侧链

利用以太坊构建侧链 （下回分解……）

高性能高扩展性区块链系统原理



扩展的共识协议（Scalable consensus）：
采用新的可扩展的共识协议

PoW工作量证明

计算出一个满足规则的随机数，即获得本次记账权，发出本轮需要记录的数据，全网其它节点验证后一起存储。

优点

完全去中心化，节点自由进出。

比特币

缺点

挖矿造成大量的资源浪费；共识达成的周期较长，不适合商业应用

PoS 权益证明

Pow的一种升级共识机制；根据每个节点所占代币的比例和时间；等比例的降低挖矿难度，从而加快找随机数的速度。

优点

在一定程度上缩短了共识达成的时间

未来币

缺点

需要挖矿，本质上没有解决商业应用的痛点

DPoS 权益证明

类似于董事会投票，持币者投出一定数量的节点，代理他们进行验证和记账。

优点

大幅缩小参与验证和记账节点的数量，可以达到秒级的共识验证

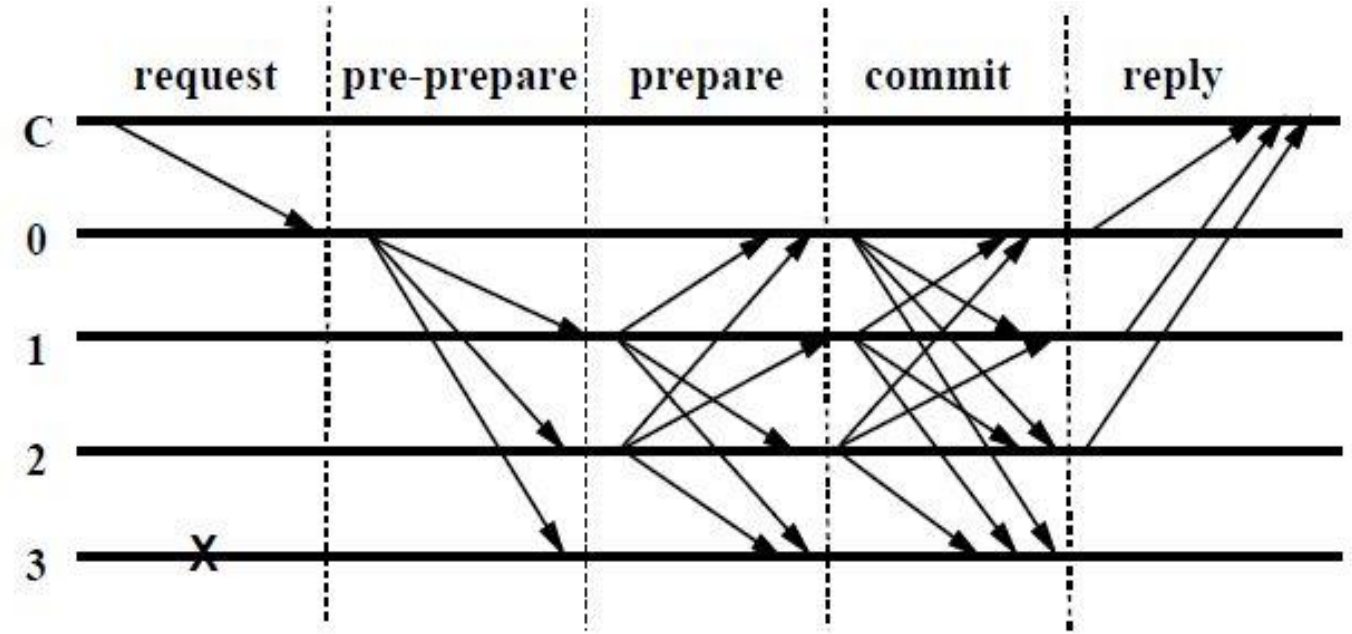
比特股

缺点

整个共识机制还是依赖于代币，很多商业应用是不需要代币存在的

PBFT拜占庭容错算法

- 当节点发现leader作恶时，通过算法选举其他的replica为leader；
- leader通过pre-prepare 消息把它选择的 value广播给其他replica节点，其他的replica节点如果接受则发送 prepare，如果失败则不发送；
- 一旦2f个节点接受prepare消息，则节点发送commit消息；
- 当2f+1个节点接受commit消息后，代表该value值被确定



PBFT拜占庭容错算法

消息传递的一致性算法，算法经过三个阶段达成一致性，这些阶段可能因为失败而重复进行

优点

- 系统运转可以脱离币的存在,安全性与稳定性由业务相关方保证;
- 共识的时延大约在2~5秒钟;
- 共识效率高;

超级账本

HYPERLEDGER

缺点

- 其封闭性（节点数目提前确定并互相联通）;
- 高性能开销(O(N^2)的消息量);
- 连接节点数量有限，无法用于公链;

Ripple共识协议

PCA对交易分两个阶段完成，第一阶段是达成交易集（即未打包进入区块的合法交易的集合）的共识，第二阶段是对新生成的区块进行提议，最终形成被共识过的区块。

优点

通过子网络内部相互信任，从而达到了更加迅速的共识速度，并保证了最终性

Ripple币

缺点

- 很容易导致分叉，安全性较弱；
- 非常中心化，很大一部分的记账节点实际上是由Ripple自己控制的

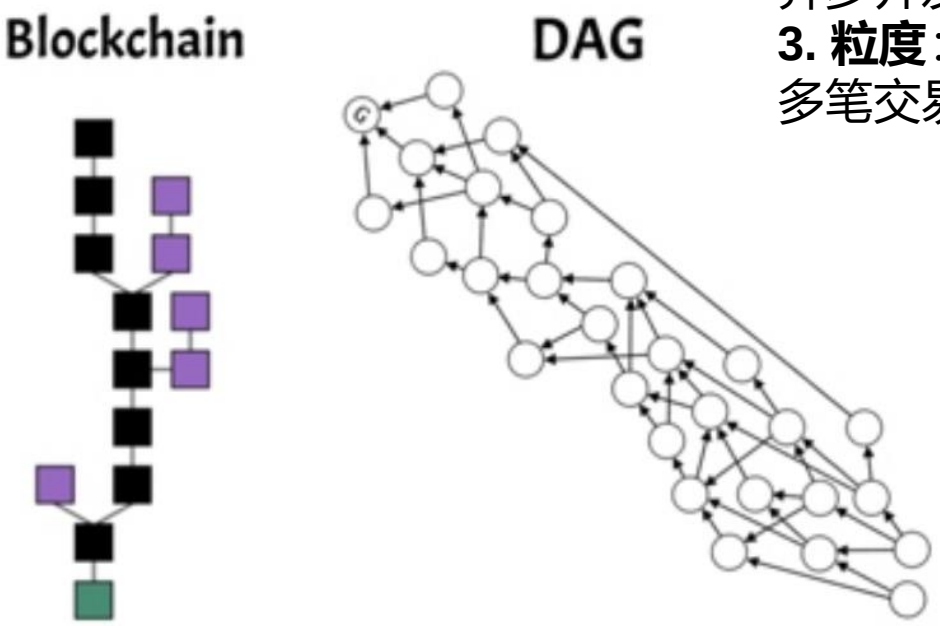
共识机制比较

共识机制	优势	劣势
POW	1. 安全稳定，节点自由度高 2. 去中心化程度高，节点系统开放	1.扩展性弱，性能低 2.没有最终性 3.造成硬件设备浪费
POS	1. 能源耗费少 2. 去中心化程度较高，节点系统开放	1.实现过程复杂 2. 存在安全漏洞
DPOS	1.能源耗费少 2. 性能高 3. 具备最终性	去中心化程度弱,节点系统相对封闭
BFT	1.性能较高 2. 具备最终性 3. 安全性好	1.去中心化程度弱，节点系统封闭 2. 容错率低

DAG

有向无环结构
（DAG）

DAG

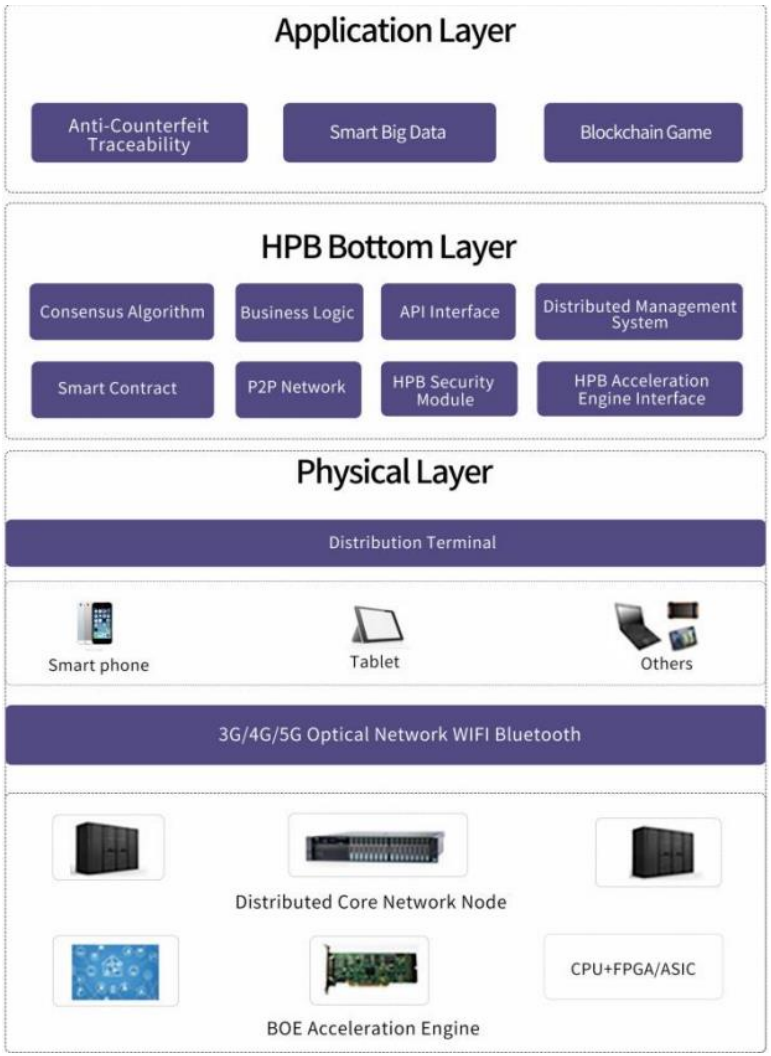


1.单元：区块链组成单元是Block（区块），DAG组成单元是TX（交易）；
2.拓扑：区块链是由Block区块组成的单链，只能按出块时间同步依次写入，好像单核单线程CPU； DAG是由交易单元组成的网络，可以异步并发写入交易，好像多核多线程CPU；
3. 粒度：区块链每个区块单元记录多个用户的多笔交易，DAG每个单元记录单个用户交易。

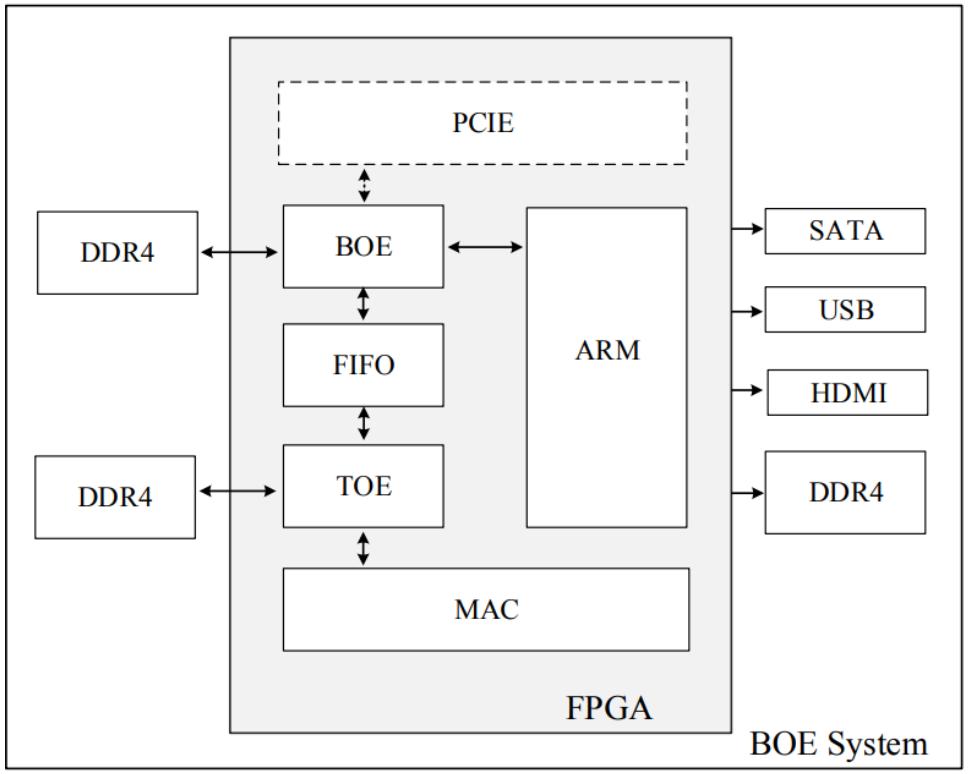
高性能高扩展性区块链系统原理

系统优化：
通过优化系统
架构、代码结构、
数据存储效率、
网络发送效率等，
提高区块链性能

HPB区块链



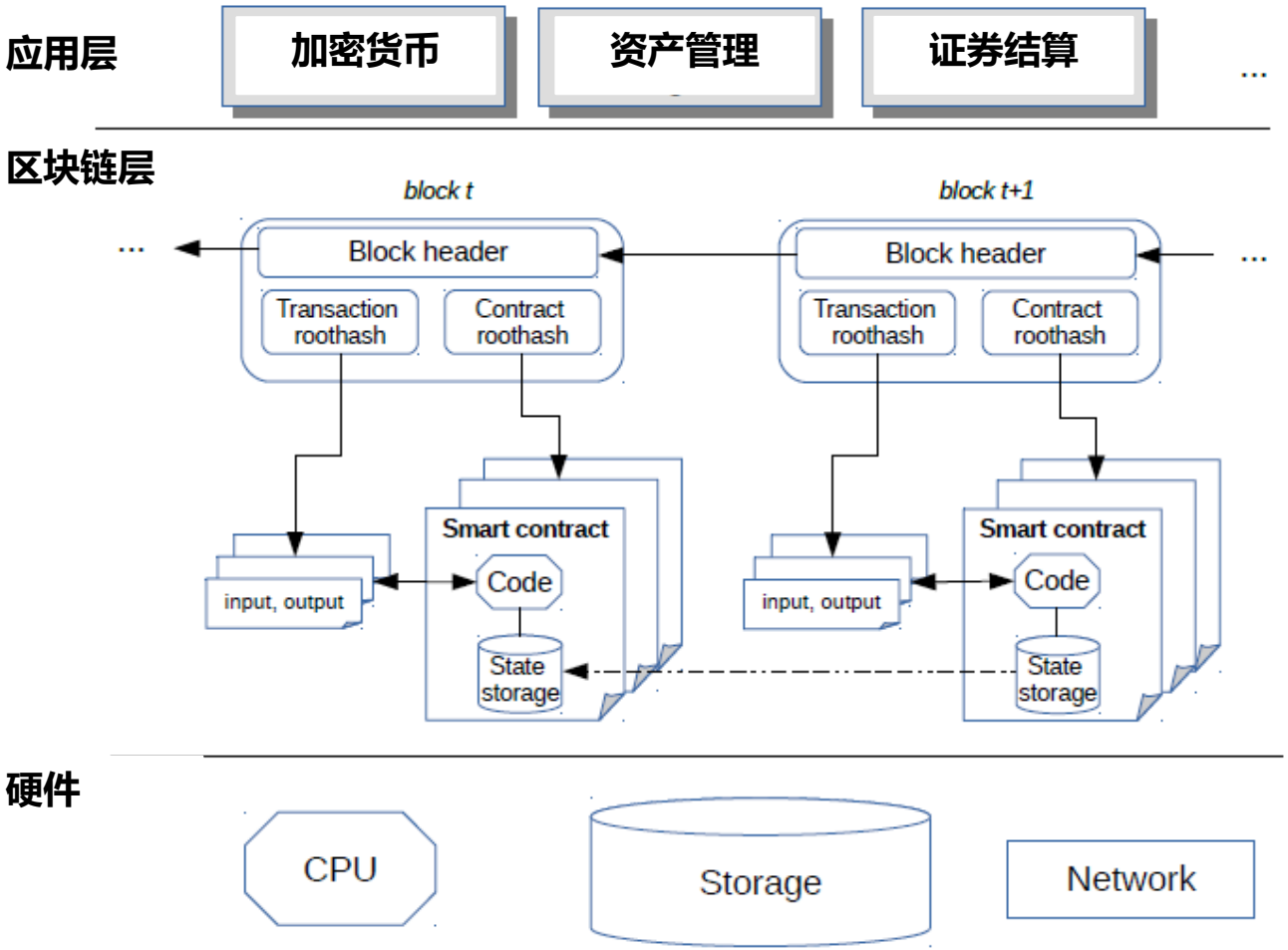
硬件加速



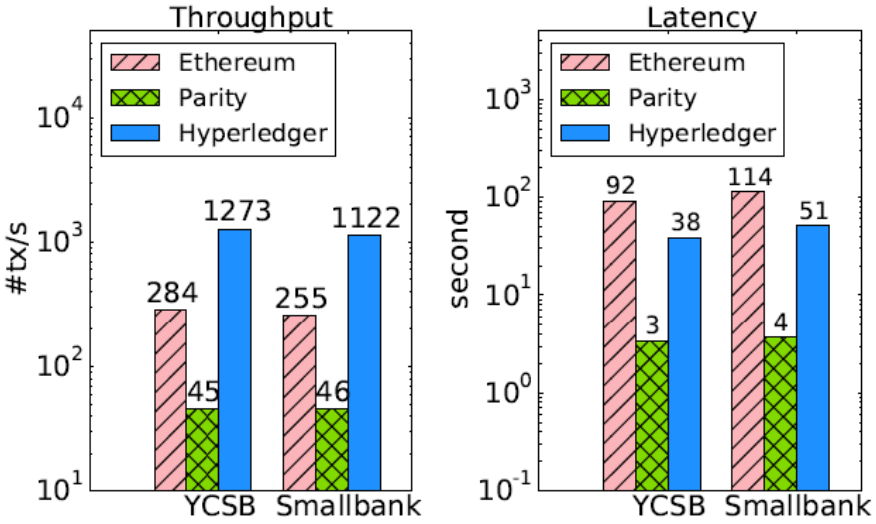
Blockchain Offload Engine (BOE)

区块链测试床

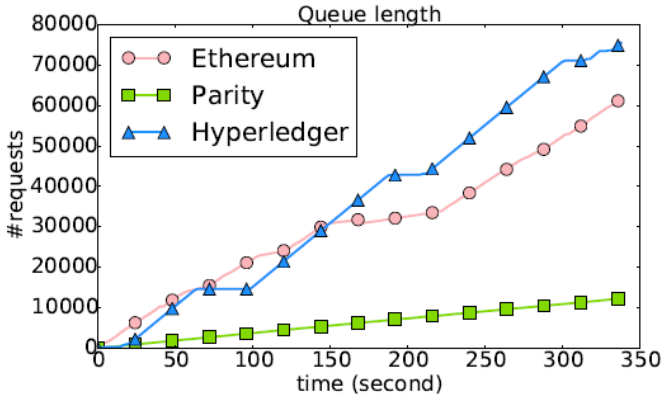
BLOCKBENCH： http://www.comp.nus.edu.sg/~dbsystem/blockbench/



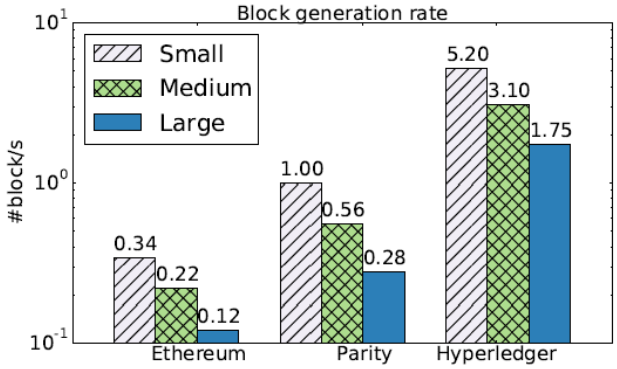
区块链测试床



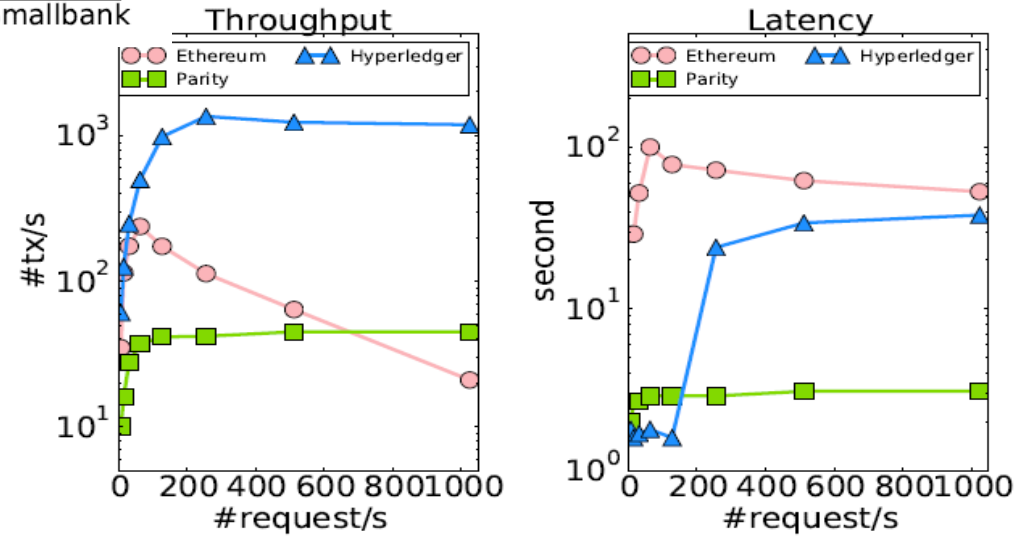
峰值的性能指标



客户端排队的长度



区块大小及产生速率



请求率不同时的性能曲线

区块链测试床

	Layer	Examples
1	View 可视化界面	UTXO (BTC), Balance (ETH), Segments (Indigo)
2	Application 应用程序	Contracts (ETH), Processes (Indigo)
3	Storage 存储	DHT, Databases, CAN
4	Execution 执行环境	Compilers, VMs, Dockers
5	Consensus 共识协议	PoW, PoS, PBFT
6	Data Model 数据模型	Blocks, Transactions, Indexing
7	P2P Propagation P2P传播	Relay Network, Fast Relay Network, Falcon, Fibre
8	Side Plane 侧链	Sidechains, Payment Channels, Hubs, Oracles, Bridges to other networks
9	Network 网络	TCP/IP, UDP

区块链系统细粒度划分

区块链系统性能指标

9	Peak Memory Usage **峰值内存利用率**	Execution	Data Model, Execution	RAM utilization (both on- and off-chain) for segment creation
10	Execution Time **执行时间**	Execution	Execution	Time to create a segment, i.e., • Off-Chain: CPU Cycles behind an agent/method execution. • On-Chain: CPU cycles for running the Validation Code (as we do in Indigo wrt the Manifest). This could be compared with the CPU cycles for a similar problem-statement solved by a Smart Contract (ETH) .
11	Availability **可用性**	Reliability **可靠性**	Network, P2P Propagation, Consensus	• The ratio of uptime to total time.

细粒度的性能指标

区块链测试标准的意义

	Comparison Type	Example
1.	Standards-Based 提供标准	How do our results compare with an industry-accepted standard?
2.	Inter-Stack Benchmark 不同平台比较	How does the performance of technology stack A compare with technology stack B, as a solution to the same high-level problem?
3.	Intra-Stack Benchmark 平台内优化	How does the performance of a given technology stack change as it configured with different options?

优化性能、完善系统

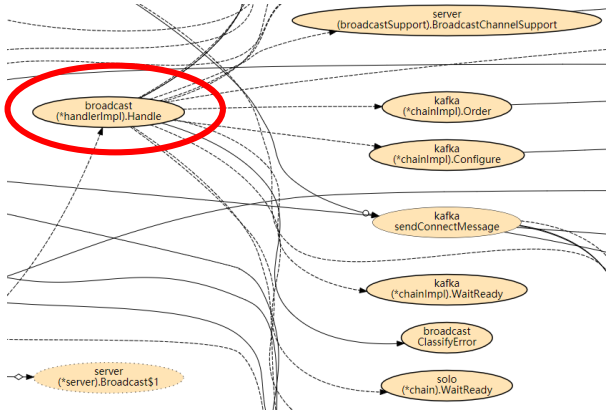
超级账本测试性能

Test	Name	Succ	Fail	Send Rate	Max Latency	Min Latency	Avg Latency	75%ile Latency	Throughput
1	simpleOperations	4000	0	200 tps	1.56 s	0.27 s	0.85 s	1.09 s	196 tps
2	simpleOperations	10000	0	499 tps	14.19 s	1.83 s	8.81 s	11.48 s	388 tps
3	simpleOperations	20000	0	950 tps	50.14 s	4.05 s	33.80 s	37.76 s	370 tps
4	query	2000	0	100 tps	0.03 s	0.00 s	0.01 s	0.01 s	100 tps
5	query	4000	0	200 tps	0.03 s	0.00 s	0.01 s	0.01 s	200 tps
6	query	10000	0	499 tps	0.06 s	0.00 s	0.01 s	0.01 s	498 tps
7	query	20000	0	1000 tps	0.46 s	0.00 s	0.08 s	0.10 s	999 tps

resource consumption

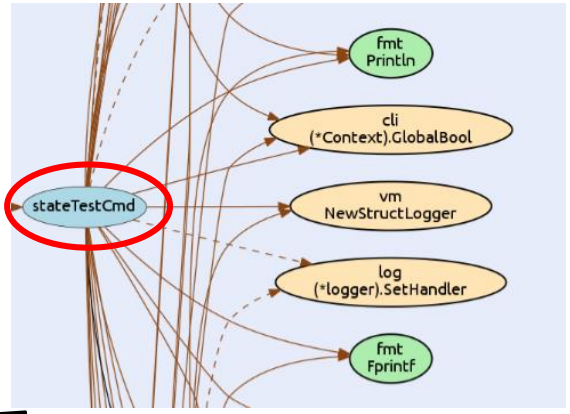
TYPE	NAME	Memory(max)	Memory(avg)	CPU(max)	CPU(avg)	Traffic In	Traffic Out
Process	node local-client.js(avg)	114.3MB	110.0MB	37.36%	28.85%	-	-
Docker	dev-peer1.org2.example.co...le-v0	6.3MB	6.1MB	8.93%	7.28%	4.8MB	2.6MB
Docker	dev-peer1.org1.example.co...le-v0	6.2MB	6.0MB	8.85%	7.40%	4.8MB	2.7MB
Docker	dev-peer0.org2.example.co...le-v0	6.0MB	5.8MB	6.04%	5.07%	3.3MB	1.9MB

区块链系统静态分析

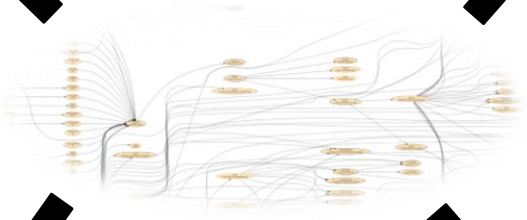


Hyperledger Ordering 组件

源代码插桩

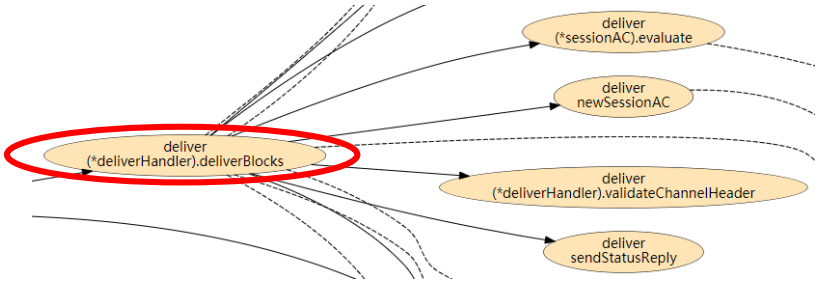


以太坊 Node组件

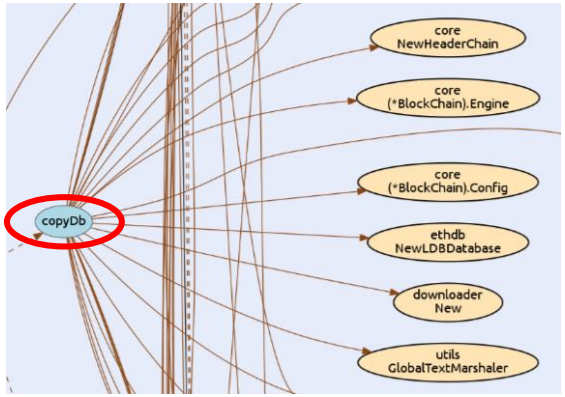


HYPERLEDGER

ethereum
HOMESTEAD RELEASE
BLOCKCHAIN APP PLATFORM

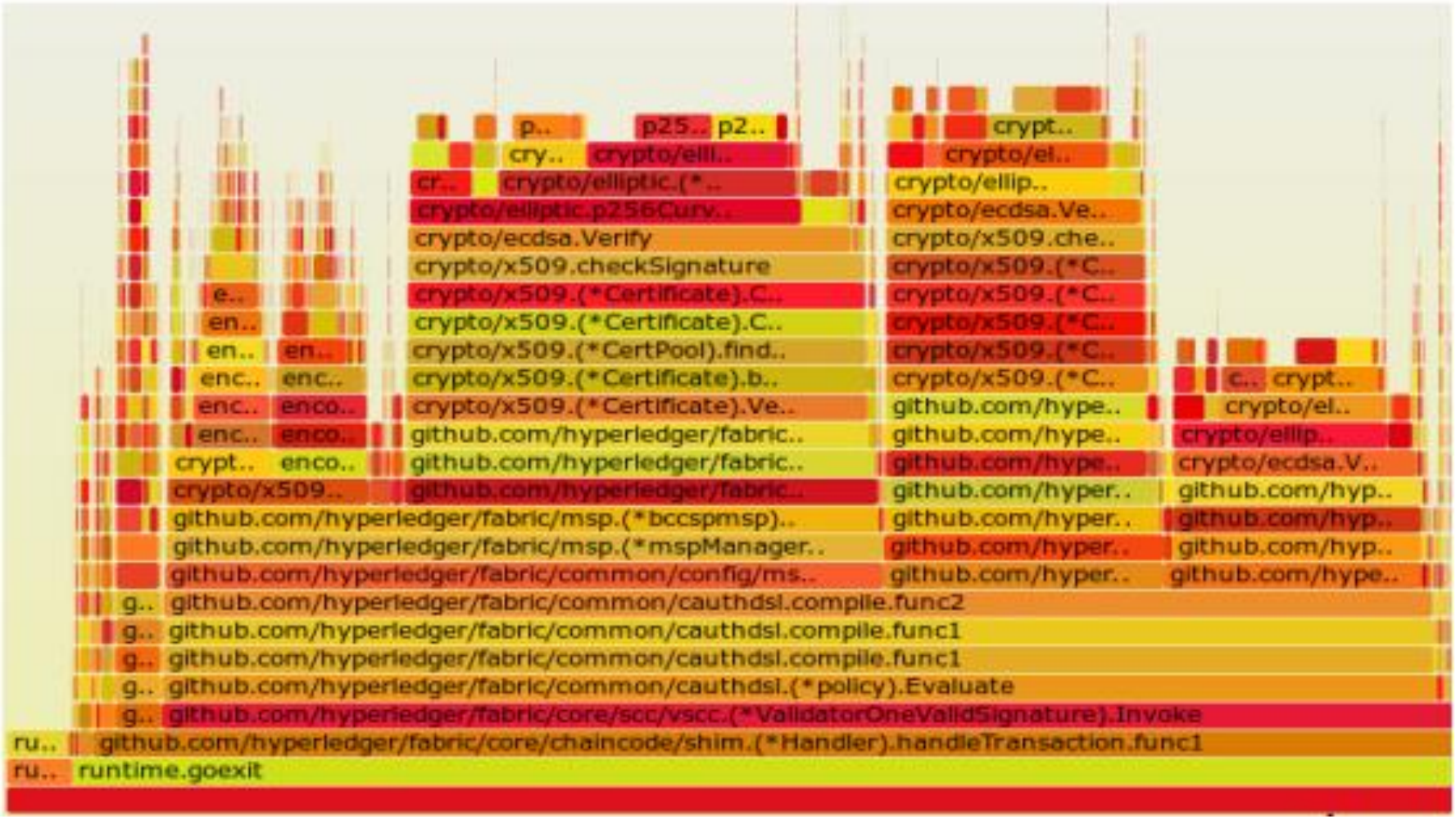


Hyperledger Peer 组件



以太坊 Geth组件

区块链系统动态分析



Hyperledger的MSP服务的函数调用次数及调用栈

区块链系统动态分析

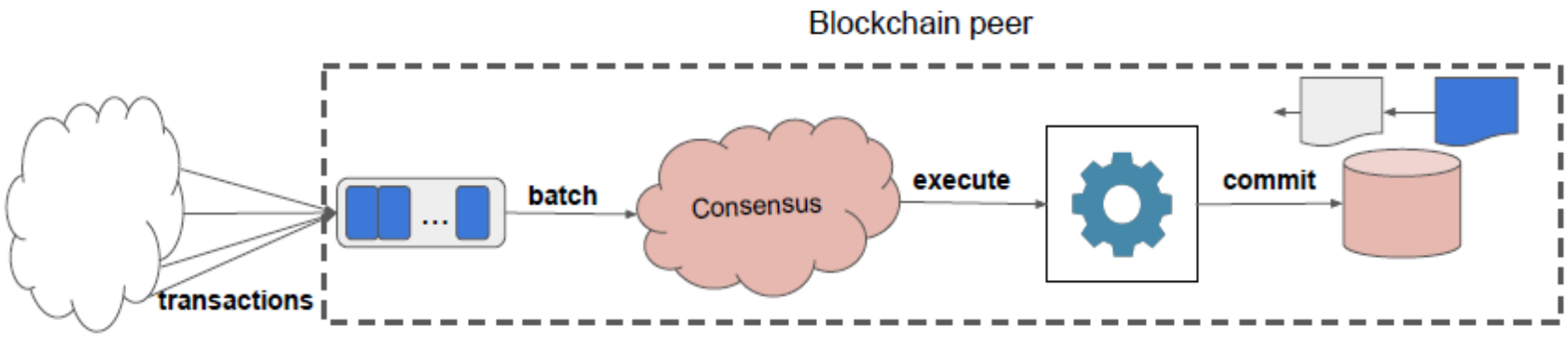
[36m2018-05-19 08:17:40.491 UTC [orderer/common/broadcast] Handle -> DEBU 48e[0m [channel: mychannel] Broadcast is processing normal message from 172.17.0.1:51660 with txid '322a0a44e60d011f44aae05c708a6e8d07b890de22646d9ea868454a8eea0cf5' of type ENDORSER_TRANSACTION

[36m2018-05-19 08:17:40.491 UTC [policies] Evaluate -> DEBU 48f[0m == Evaluating *policies.implicitMetaPolicy Policy /Channel/Writers == with txid '322a0a44e60d011f44aae05c708a6e8d07b890de22646d9ea868454a8eea0cf5' of type ENDORSER_TRANSACTION

[36m2018-05-19 08:17:40.491 UTC [policies] Evaluate -> DEBU 490[0m This is an implicit meta policy, it will trigger other policy evaluations, whose failures may be benign with txid '322a0a44e60d011f44aae05c708a6e8d07b890de22646d9ea868454a8eea0cf5' of type ENDORSER_TRANSACTION

[36m2018-05-19 08:17:40.491 UTC [policies] Evaluate -> DEBU 491[0m == Evaluating *policies.implicitMetaPolicy Policy /Channel/Application/Writers == with txid '322a0a44e60d011f44aae05c708a6e8d07b890de22646d9ea868454a8eea0cf5' of type ENDORSER_TRANSACTION

Hyperledger Fabric 运行时产生的日志



高性能区块链系统构建

BaaS （Blockchain as a Service）

应用层(Dapp)

天气预测 | 智能交通 | 医疗健康 | 消费习惯分析 | 地理环境 | 生物信息 | 视频推荐 | + 更多...

应用框架层

Tensorflow	SparkML	Theano	H2O	R 统计	自定义AI服务
Hadoop	Caffe	HTCondor	SystemML	智能合约	+ 更多...

区块链平台

以太坊 | 超级账本 (HYPERLEDGER) | Parity | Ripple (ripple)

云平台

Kubernetes 云管理平台：部署 | 调度 | 自动扩展 | + 更多...

基础设施：存储 | 计算 | 网络 | + 更多...

云运维 (AIOperator)

谢谢！