2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)

An Innovative IPFS-Based Storage Model for Blockchain

1st Qiuhong Zheng
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing,China
13121777388@163.com

1st Yi Li
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing,China
liyi@bupt.edu.cn

1st Ping Chen
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing,China
13611349651@139.com

2nd Xinghua Dong
Industrial and Commercial Bank of China
Beijing,China
87471379@qq.com

***Abstract*—Blockchain technology has received great attention in recent years. However, the data volume of blockchain grows continuously due to the features that cannot be deleted and can only be added. Currently, the total size of Bitcoin blockchain ledger has reached 200GB. Its high demand for storage space and bandwidth to synchronize data with the network prevents many nodes from joining the network. This is not only not conducive to the expansion of this decentralized network, but also becomes the bottleneck of the development of blockchain technology. This paper proposes an IPFS-based blockchain data storage model to solve this problem. In this paper, the miners deposit the transaction data into the IPFS network and pack the returned IPFS hash of transaction into the block. Utilizing the characteristics of the IPFS network and the features of the IPFS hash, the blockchain data is greatly reduced. The scheme is applied to the Bitcoin blockchain. According to the experimental results, the compression ratio can reach 0.0817. According to the analysis, it also has good performance in terms of security and synchronization speed of new node.**

Keywords—blockchain, storage optimization, distributed storage

I. INTRODUCTION

The blockchain is a immutable, cryptographically secure distributed ledger with high Byzantine fault tolerance. As the underlying technology of the Bitcoin system, it was first introduced in the Bitcoin paper [1] published by Nakamoto.Since the operation of Bitcoin network in 2009, the size of Bitcoin ledger has grown year by year, leading to storage bloating problem [2]. Currently, a complete ledger of the Bitcoin network occupies up to 200GB, and this number is constantly increasing at about 0.1GB per day, causing great pressure on data storage (for full-nodes, especially portable devices [3])and download (new node synchronization [4]). For the Bitcoin blockchain, the block body mainly contains the transfer transactions with relatively small amount of data. In the future, the scenario of the blockchain application will not be limited to this. For example, in blockchain-based IoT system [5], transaction information will involve files, video, audio and other information, leading to even larger blockchain data. The high demand for storage space of blockchain data is very unfavorable to the expansion of the network, which has become a bottleneck restricting the development of blockchain technology. A more efficient way to store and download blockchain data is urgently needed.

In the blockchain system, each node keeps a copy of the ledger. The advantage of this is to ensure the security, but the drawbacks are also obvious. Repeated storage of the same data leads to too much data redundancy in the system. Therefore, an intuitive idea is to reduce redundant data. A network coding-based distributed storage (NC-DS) framework [2] was proposed.After each block is created, it is divided into several sub-blocks, and then the network coding is adopted to encode these sub-blocks into more sub-blocks, which are then distributed into all nodes. Since the encoding and decoding process increases the complexity of the system, and there will be a large number of block requests broadcast on P2P networks if the transactions being queried by the miners are not local.Although this scheme can reduce the storage to some extent, it results in a decrease of system efficiency. At the same time, the scheme is prone to data inconsistency problems.

Another attempt is the mini blockchain project. The idea in the paper [7] is to store the balance of all non-null addresses through a database, called the "account tree", by using the structure of the balance tree to store account information on the chain. The expired transactions on the blockchain are deleted, and only the block header data is retained, thereby reducing space occupation. But this approach is too radical, deleting historical transaction data and losing the historical traceability of the blockchain.In addition, the scheme also has limitations on the types of transactions processed, which are limited to transactions involved in changing the account balance.

Another paper [7] adopted a method of recording block changes to reduce the storage of blockchain data. A summary block can be viewed as a single big transaction whose inputs are the inputs of all the transactions of its constituent blocks and whose output is all the unspent outputs of the same. But the application scenario of this solution is only for transactions involving transferable entities. Such as transactions dealing with bank accounts or land agreements, and not for storage identities. There are limitations and there is a strong dependence on the proportion of full-node in all nodes.

The paper [8] designed a similar summary block for the Bitcoin blockchain. At regular intervals, all UTXOs in the

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.000-8

current blockchain are summarized into a file and the remaining transaction data is deleted. The drawback of this is also that it cannot retain the historical traceability of blockchain. The measure adopted in this paper is to retain some full-nodes. As a result, the scheme relies heavily on the total number of nodes and does not have enough incentives to guarantee their existence.

The above schemes have some limitations and drawbacks, and the compression ratio of the data is not effective.For the optimization of blockchain data storage, this paper takes the Bitcoin blockchain as an example, and considers the transaction data to be stored in the IPFS (InterPlanetary File System) [9]. Only the IPFS hash of the transaction is packed in the block to alleviate the blockchain data storage pressure. In the existing Bitcoin network, most of the nodes maintain the same contents of the ledger.Since IPFS is content addressed. The same transaction data has the same hash in IPFS. Therefore, the IPFS hash of the transactions stored by each node in the blocks are the same, which can keep the consistency of the data in each node. At the same time, this scheme eliminates the limitation of dependence on the number of full-node in the network, has no restriction on the type of transaction, and retains all transaction data, that is, retains the traceability of the blockchain history. In addition, the synchronization speed between the newly added node and the network is accelerated. And the storage threshold of the blockchain network node is reduced. Finally, through an experiment and the statistics of the authentic Bitcoin blockchain data, the compression ratio of the storage space is calculated, and the performance of the model is analyzed from the perspectives of storage space, security and synchronization speed of the new node.

The contribution of the paper is as follows:

- We propose a new blockchain data storage model that illustrates how to use IPFS networks to reduce the storage of blockchain data.
- We implement a proof of concept showing how miners can use the model in this paper to store less blockchain data in a real mining scenario, and how new nodes can quickly synchronize with the network.
- We decode and make statistics of the authentic Bitcoin data, calculated the compression rate of the scheme, and demonstrated the excellent characteristics of the scheme through storage space, security and new node synchronization speed.

The organization of the paper is as follows. Section II describes the blockchain data storage model in detail, including system design and data processing flow. Section III calculates the data storage compression ratio of the model and analyzes the performance of the model. Section IV is a summary and future work.

II. Proposed Scheme

This section mainly describes the innovative IPFS-based storage model for blockchain and the key technologies used.

A. Miner Workflow

For miners, there are two main workflows: joining the network process and the mining process.

New Node Synchronization Process. For a node that is newly joined to the network, data synchronization with the network is required. The new node communicates with the neighboring nodes through Bitcoin P2P network to obtain the complete blockchain ledger. In addition, it needs to traverse transactions in the blockchain to build a local index and UTXO pool. The local index is a k-v database that stores the metadata of blocks and transactions, making it possible to query a block or a transaction through the block hash and transaction hash. When traversing transactions in the blockchain, the unspent output of all transactions, UTXO, is also summarized into a UTXO pool for subsequent mining process.

Mining Process. For a node that has access to the network for mining, it first need to verify the transactions it receives. The specific transaction verification process mainly involves checking the transaction format, whether the inputs of the transaction are unexpended, and whether the transaction fee and lock script meets the requirements. If a transaction is verified, the eligible transaction is deposited into the trading pool. After the previous block is verified, the miner packs the transactions generated after the block into a new block. For successfully adding the block in the blockchain, the miners need to solve a crypto-puzzle by competing and calculates the block hash[10]. The miner who first calculated the hash broadcasts the block to other nodes. At the same time, other nodes must mutually verify the correctness of the block. If it passes the validation, other miners would append the new block to the blockchain[11].

In a newly generated blocks, the vast majority of the data is the transaction data. In fact, a complete block containing all the transactions is larger than 1000 times the size of the block header [12].The main work of a miner is mining. For a mining process, it does not involve frequent access to transactions in historical blocks. Therefore, the optimization of storage can be realized by transforming transaction data in historical blocks into another storage form with smaller data volume,without reducing the mining efficiency of the system.

B. IPFS-Based Blockchain Data Storage Model

Introduction of IPFS. Interplanetary file transfer network (IPFS) is a distributed data storage protocol.IPFS is content addressed, allocating a unique hash for each stored file. It has a good deduplication mechanism, without central server limitation. Besides,data uploaded in the system can be stored permanently.For data with high frequency requests,IPFS will create duplicate data along the request path,which can be directly read locally on the next request. In summary, IPFS is a secure, high-throughput, content addressed block storage model that supports high-capacity storage and high concurrent access[13]. Bitcoin's transaction data can be up to tens of thousands of bytes in size, typically at least two hundred bytes. And an IPFS hash is only a few tens of bytes. Therefore, consider storing the transaction data in IPFS and storing the returned IPFS hash in the block, thereby achieving a large reduction in storage space.

Innovative Storage Model. Considering the requirements of blockchain data storage and the characteristics of IPFS, the IPFS-based blockchain data storage model is proposed, as shown in Fig. 1.

In the model, each miner checks the received transactions, puts the valid transactions into the trading pool, and stores them in IPFS, keeping the returned transaction IPFS hash. When calculating next block, each miner packs the IPFS hashes of the verified transactions into the new block,

calculates the merkle root and the block hash. The block structure is shown in Fig. 2. If miner A successfully calculates the block hash that meets the difficulty, the block will be broadcast to miners B, C, D, etc. After miners B, C and D receives the block, they need to verify the transactions and the block hash. In fact, most of the transactions received by miners B, C and D during the mining process are the same as those of miner A. Only a few transactions in their transaction pool differ because of the network transmission delays. Therefore, most of the transaction IPFS hashes in the new block are the same as that in the local transaction pool of miner B, C and D. If the IPFS hash of a transaction in the block sent by miner A can be found in their local transaction pools, it means that the transaction has been confirmed by B, C and D before, it will not need to be downloaded from IPFS. For the rest of the transactions, the data needs to be requested from the IPFS network through their corresponding IPFS hashes. Then the validity of transactions and the block would be confirmed. Afterwards,the new block can be appended to the blockchain ledger.

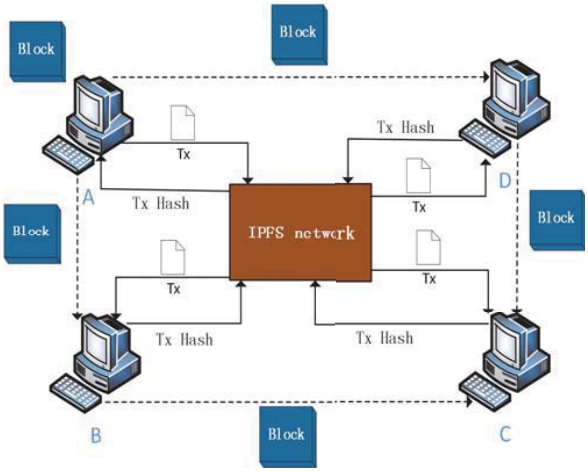

Fig. 1. IPFS-based storage model for blockchain.

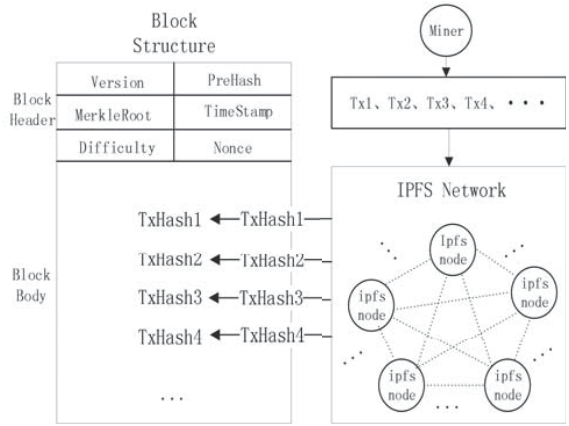

Fig. 2. Block structure.

In summary, most of the transactions in the new block can be verified locally,and a few that cannot be found in the local transaction pool need to request transaction data from the IPFS network to be verified. This will not result in a significant reduction in system efficiency Caused by the large amount of transaction data requests to IPFS network.

III. Results and Discussion

This section analyzes the model from the perspective of storage space, security and new node synchronization speed. The data compression ratio of the model is calculated experimentally.

A. Storage Space

The compression ratio of storage space is calculated as follows.The following Table 1 gives the implication of each symbol in the formula.

$$\frac{H+iHash\times N}{H+\sum_{i=1}^{N}Tx_i} \tag{1}$$

TABLE I. The implication of each symbol in the formula

Symbol	*Implication*
H	the data volume of all block header in the blockchain
iHash	the size of the IPFS hash corresponding to each transaction
N	the number of all transactions in the blockchain
Tx	the original data volume of each transaction

The block header is 80 bytes. Each block contains at least 500 transactions on average. Each transaction has at least 250 bytes [12] on average. An IPFS hash only takes up 46 bytes. It can be estimated that the optimization of the storage space of this solution is very considerable. To get authentic Bitcoin blockchain data, first install the Bitcoin Core client and download the blockchain data. Then use python to call the PRC API of Bitcoin Core client through the bitcoinrpc extension pack to obtain the data(as of July 2018). And use the python extension pack, MySQLdb, to connect to the Mysql database to save the data. Finally, statistical analysis and calculation of the obtained data.

The total number of transactions per month is shown in Fig. 3. And the comparison before and after the processing of block data in each month is shown in Fig. 4. It is intuitive to see that the solution is effective.

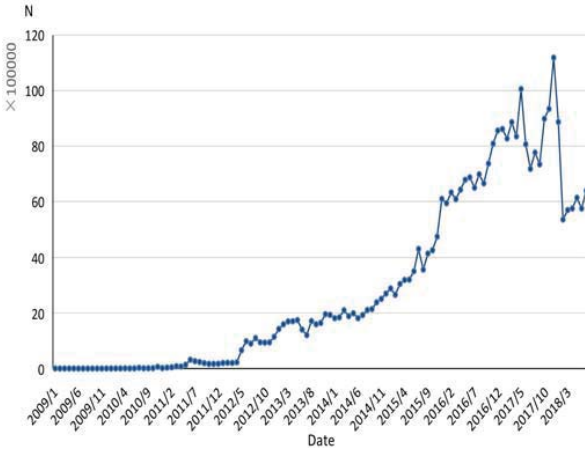

Fig. 3. The total number of transactions per month.

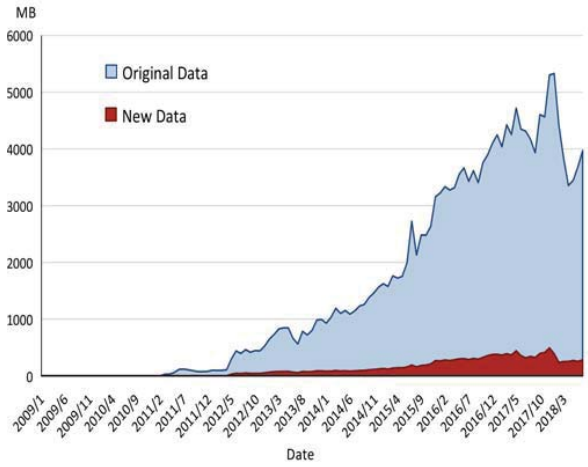

Fig. 4. Comparison of the sum of block data in each month before and after data processing

The compression ratio after data processing in each month is shown in Fig. 5.We can see that the compression rate in 2009 was relatively high, even greater than 1. This is because the number of transactions and the transaction data at that time is too small, resulting in less optimization and even larger data after data processing,which is shown in Fig. 6. After 2009, the number of transactions and block data significantly increased, compression ratio decreased, and the scheme had a good effect on storage optimization.Finally, the solution of this paper reached a total compression ratio of 0.0817.

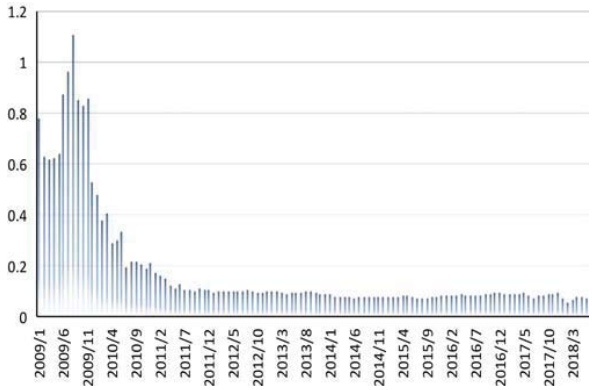

Fig. 5. Compression ratio of block data per month

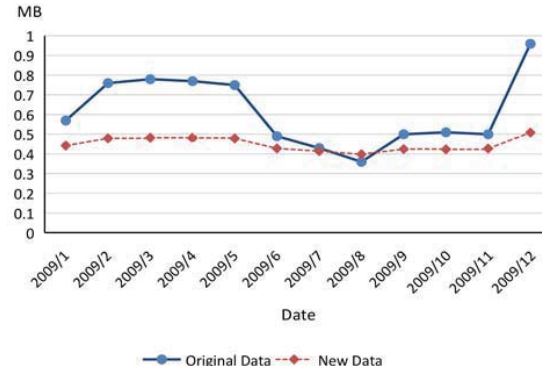

Fig. 6. Comparison before and after block data processing in each month of 2009

B. Security

The merkle root in the block header of the model is calculated based on the IPFS hash of the transaction.And the block header hash is calculated based on the merkle root. Therefore, the validity of the blockchain can be verified directly from the local blockchain data.Once a malicious node wants to change a transaction data in a block, its IPFS hash also changes.A small change will cause a huge difference in the merkle root, which will cause the block header hash to change. Therefore, the invalid block will not be recognized by the blockchain network, which ensures the security of the block data.

C. New Node Synchronization Speed

In the scheme of this paper, for nodes newly joining the network, on the one hand, it is necessary to synchronize the blockchain data from neighboring nodes. On the other hand, it is necessary to download transaction data from the IPFS network through the IPFS hash in the block.Then the validity of the blockchain is checked, and local index and UTXO pool are constructed for subsequent mining or transaction verification process.

When a new node is synchronized with the network,the blockchain data that needs to be downloaded from neighboring nodes is compressed by 0.0817. Although the new node needs to download transaction data from the IPFS network for verification, this process can be concurrent with synchronizing data with the blockchain network. In fact, due to the high speed download of IPFS, the total time for synchronization of a new nodes can be reduced. In conclusion, the model proposed in this paper can theoretically speed up the synchronization between new node and the network.

IV. Conclusion and Future Work

In this paper, an innovative IPFS-based storage model for blockchain is designed. It can be seen from the result that it has a certain degree of improvement in storage space, security and new node synchronization speed. And this solution not only works for Bitcoin type transactions, but also for other transaction types. In the scheme, because IPFS has relatively good distributed storage performance, it is chosen as the storage system. If there is a system with better performance in the future, it can also be conveniently transformed. Future work will consider how to implement the model to test performance in authentic network situations.

References

[1] Nakamoto S. Bitcoin: a peer-to-peer electronic cash system [Online], available: https://bitcoin.org/bitcoin.pdf, 2009.

[2] Dai M, Zhang S, Wang H, et al. A Low Storage Room Requirement Framework for Distributed Ledger in Blockchain[J]. IEEE Access, 2018, PP(99):1-1.

[3] K. Wang, J. Mi, C. Xu, Q. Zhu, L. Shu, and D. J. Deng, ‘ ‘Real-time load reduction in multimedia big data for mobile Internet,’ ’ ACM Trans. Multimedia Comput., Commun. Appl., vol. 12, no. 5, Oct. 2016, Art. no. 76.

[4] Bonneau J, Miller A, Clark J, et al. Research Perspectives and Challenges for Bitcoin and Cryptocurrencies[J]. 2015, to appear:104-121.

[5] Dorri A, Kanhere S S, Jurdak R, et al. Blockchain for IoT Security and Privacy: The Case Study of a Smart Home[C]// IEEE International Conference on Pervasive Computing and Communications Workshops. IEEE, 2017.

[6] J.D Bruce. The Mini-Blockchain Scheme. http://cryptonite.info/files/mbc-scheme-rev3.pdf. (March 2017).

[7] A.Palai,M.Vora,and A.Shah.Empowering Light Nodes in Blockchains with Block Summarization.In International Workshop on Blochchains and Smart Contracts.IEEE,2018.

[8] Bo Li.Analysis and Optimization of Block Chain Storage for Bitcoin Miner Node[D].Dalian Maritime University,2018.

[9] J. Benet, "IPFS – Content Addressed, Versioned, P2P File System," Protocol Labs, Inc., Tech. Rep., 2014.

[10] Tosh D K, Shetty S, Liang X, et al. Security Implications of Blockchain Cloud with Analysis of Block Withholding Attack[C]// Ieee/acm International Symposium on Cluster, Cloud and Grid Computing. IEEE, 2017.

[11] Yue Hao, Yi Li, Xinghua Dong, Li Fang , Ping Chen.: Performance Analysis of Consensus Algorithm in Private Blockchain. IEEE Intelligent Vehicles Symposium (IV) .IEEE,2018.

[12] Master bitcoin.Andreas· M· Antonopoulos (O'Reilly). Copyright 2015 Andreas· M· Antonopoulos.978-1-449-37404-4.

[13] Ziyan Wang, Xinghua Dong, Yi Li, Li Fang , Ping Chen.: IoT Security Model and Performance Evaluation: A Blockchain Approach. Proceedings of IEEE IC-NIDC 2018[C]. Beijing China: Institute of Electrical and Electronics Engineers, Inc., 2018: 053-069.