Blockchain based Architecture for Digital-right Management in Scientific Data Sharing

Xiaoying Zheng\(^1\) and Yongxin Zhu\(^1\)

1. Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China

The corresponding author: Yongxin Zhu zhuyongxin@sari.ac.cn

Abstract. The data integration and copyright protection are the essential requirements in promoting the scientific data sharing. We present a blockchain-based architecture for digital-right management in scientific data sharing. The main innovations are the multiple parallel sub-chains based block structure, and the on-chain/off-chain hybrid storage. This leads to an efficient and scalable data management platform for the large-scale scientific data and still maintains the tamper-proof and traceability features of blockchain. The qualitative analysis shows that the parallel sub-chain based architecture can improve the throughput and scalability of the original Hyperledger Fabric chain by up to \(10\times\).

1. Introduction

Data sharing is the essential part of the scientific process, which is not only able to improve the reproducibility and robustness of the science, but also can drive new science for future [1]. In data sharing practice, scientists are concerned whether they will receive credits for their work and that their data will not be abused [2]. Thus the data integration and copyright protection are the key concerns in data sharing practice. The emerging technology of blockchain provides a promising solution to solve the traditionally expensive intellectual property protection and tracking problem [3]. Blockchain is a decentralized, distributed and public digital ledger that can record transactions across many parties efficiently and in a verifiable and permanent way [4]. The blockchain’s natures of decentralization, enhanced security, tamper-proof, improved traceability, transparency allow participants to share their data in a traceable and relatively inexpensive way.

In this paper, we present a blockchain-based architecture for digital-right management in scientific data sharing. The original blockchain design seeks a paradigm of completely decentralization and experiences performance loss in terms of throughput and scalability. This makes blockchain unable to scale up in practice. We propose two innovative techniques including the multiple parallel sub-chains based block architecture and the on-chain/off-chain hybrid storage, which can improve the performance of blockchain and make it feasible in the large-scale scientific data sharing. The multiple parallel sub-chains based block architecture replaces the original single chain based block structure with a main chain and a set of multiple sub-chains. Each sub-chain is independent and can process the block commitment in parallel; data swapping among sub-chains can be implemented by submitting a block to the main chain, where the consensus only needs to be carried over a group of delegated members. By allowing multiple chains to process transactions concurrently, it can improve the system throughput greatly. Secondly, it is inefficient to store a large amount of data on the chain because the on-chain storage space is often limited. We resort to a hybrid solution where the raw data are stored on
the off-chain key-value based storage, and the metadata are stored on the blockchain. The hybrid approach can efficiently support a large volume of data storage and in the mean while still maintain the integration and traceability of data in storage.

The rest of paper is organized as follows. In Section 2, we briefly introduce the key technologies of blockchain. We present the architecture of the blockchain based digital-right management system, and describe the parallel chain structure and hybrid storage with more details in Section 3. We conclude our work in Section 4.

2. Key Technologies of Blockchain

Blockchain is a set of linked blocks distributed over multiple nodes, where each block contains the transaction data, the hash of the previous block and its own hash. The key technological components including distributed ledger, cryptography, consensus and smart contracts make blockchain a system naturally providing decentralization, enhanced security, tamper-proof, improved traceability, and transparency.

2.1. Distributed Ledger

A distributed ledger is an append-only data storage spread over multiple nodes without a central coordinator, where each node keeps a complete and identical copy of the data. The data consistency of the distributed ledger across nodes is maintained by a consensus algorithm which ensures the integrity of each data copy on different peer nodes by reaching an agreement at the commit order of transactions.

2.2. Cryptography

Various cryptography techniques such as hash, public key cryptography and zero-knowledge proofs of knowledge are widely used to ensure the integrity of the distributed ledger. A blockchain usually possesses a large-scale distributed transaction blocks over a set of nodes. Thus the integrity of data requires a reliable strategy to maintain and an efficient way to verify. The Merkle tree provides a trivial solution for integration verification. In a Merkle tree, each leaf keeps the system states; each non-leaf node keeps the hash of its child nodes. Thus, the root contains the root hash of the overall ledger and any state change will result in a new root hash. Therefore, the Merkel tree allows speedy and secure verification of the distributed ledger. On the other hand, the tamper-proof of block transaction history is protected by hash pointers. More specifically, a block with an index of $n+1$ contains the hash of the predecessor block, i.e., block $n$. Thus, any change in block $n$ will result in the invalidation of all following blocks. By combining Merkle tree and hash pointers together, the data integrity is ensured.

2.3. Consensus

Consensus is a dynamic way of reaching agreement in a group, which guarantees the unique order of transactions and therefore the data consistency. Because blockchain is often deployed in a malicious environment, consensus algorithms face the risk of the Byzantine Generals problem. There are various consensus algorithms proposed and adopted in blockchain systems which are partially Byzantine Generals tolerant, including the proof of work (PoW), the proof of stake (PoS), the delegated proof-of-stake (DPoS), and the practical Byzantine fault tolerance protocol.

2.4. Smart Contracts

Smart contracts are computer codes stored and replicated in the blockchain system that help a transaction automatically executed in a transparent, conflict-free, undeniable, faster and more secure way without the aid of a third party. Some platforms allow smart contract codes run in their native runtimes; in the meanwhile, other platforms create virtual machines for executing contract codes. For example, Hyperledger Fabric relies on Dockers to run codes.
3. System Design

In this section, we present the conceptual architecture of the blockchain based digital-right management system for scientific data sharing, and describe the key technical innovation including multiple parallel chains and on-chain/off-chain hybrid storage.

3.1. System Conceptual Architecture

Because of blockchain’s inherently distributed and peer-to-peer nature, blockchain is well known for its low throughput and poor scalability. Thus, it is impractical to store the huge volume of scientific data directly on blockchain. As many of blockchain’s functions are also provided by the traditional rational database or the key-value store, we plan to integrate blockchain and the database approaches to make the hybrid storage for scientific data. More specifically, the large volume of scientific data is stored in the database or the key-value system, and the metadata are stored in the blockchain. By combining the on-chain and off-chain storage, we can obtain the efficiency, tamper resistance and traceability of data storage. We further improve the performance bottleneck of blockchain by proposing multiple parallel chain architecture. By allowing multiple sub-chains to add blocks in parallel, the blockchain can achieve higher throughput.

In Figure 1, we present the conceptual system architecture, where the raw scientific data are stored in the database/key-value store, and the metadata are extracted and stored in blockchain. In Figure 2, the blockchain software layers are shown, where a data swapping layer is proposed. A data swapping layer helps to maintain the data integrity between different sub-chains.
3.2. Parallel Chains

The single chain based architecture has been thought to be the bottleneck of blockchain since transactions cannot be processed in parallel and have to be committed one by one under the single chain scheme. A multiple chain based blockchain architecture which allows a set of sub-chains to run in parallel can help to make the system more scalable and extendible. In the proposed architecture, there is a single main chain and a set of semi-independent sub-chains. There is a value swapping layer that connects sub-chains and the main chain, and aids the interaction between sub-chains [5]. As shown in Figure 3, in each sub-chain, the block transactions follow the traditional single chain based protocol with all peers attached to this specific sub-chain. When a sub-chain plans to exchange data with other sub-chains, i.e., to add a block to the main-chain, the consensus algorithm will be carried over the delegated members elected from all sub-chains. After an agreement is reached between the delegated members, a block is committed to the main chain. By allowing multiple sub-chains to commit transactions in parallel, the throughput of the sub-chains can be improved greatly; by selecting a subset of peers from each sub-chain as the delegated members, the size of the consensus population is reduced and the consensus can be fast.
3.3. Hybrid Storage with Off-chain Storage Enhancement
On-chain storage has a limited capacity, so we have to resort to off-chain mechanisms to store the large amount of scientific data. The scientific data are often non-structural dataset and document-oriented. Thus the key-value based distributed file system is a good off-chain storage candidate. The scientific data are stored in the key-value based file system, and the metadata are stored in the blockchain system as shown in Figure 4.

3.4. Performance analysis
We choose the Hyperledger Fabric as the blockchain platform implementation. The original Fabric architecture can scale up to no more than 16 nodes and support thousands of tps (transactions per second) [4]. In our multiple sub-chains approach, each sub-chain can be implemented as an independent Fabric single chain of blocks, and the main chain organizes all these sub-chains by running the consensus algorithm over a group of delegated members. Hence, sub-chains can process the transactions independently and concurrently. Theoretically, the overall architecture can maintain 16 sub-chains, and each sub-chain can scale up to 16 nodes. In other words, the proposed multiple chain architecture can scale up to no more than 256 nodes. As the system can consist of 10 folds of nodes compared with the original one, the throughput of the blockchain system can be increased to 10,000 tps. By the qualitative analysis, the parallel sub-chain based approach can improve the throughput and scalability of the original Fabric chain by up to 10x. To further verify the validity of our architecture design, we are working on developing the prototype of digital rights management.
system. In the prototype, the Hyperledger Fabric v1.4.4 is used as the blockchain storage for the metadata, and the IPFS (InterPlanetary File System) is used to store the off-chain data. More measured results will be reported after the prototype is completed.

4. Conclusion and Future Work
In this paper, we present a blockchain based digital rights management system for scientific data sharing. The system aims to provide the functions of digital-right protection and tracking for scientific data. The key innovations are as follows. First, the multiple parallel sub-chain based architecture is proposed to improve the transaction processing throughput. Secondly, a hybrid storage combining the on-chain and off-chain data is further applied for the system scalability, where the raw scientific data on stored in the off-chain key-value database, and the metadata are stored on the blockchain for the data integration. The overall architecture can improve the throughput and scalability of the original Hyperledger Fabric blockchain by 10x theoretically. A prototype of digital rights management system is being implemented, whose measured results will be reported to further verify the validity of our architecture design.

References
[1] Editorial, Data sharing and the future of science. 2018 Nat Commun 9, 2817
[2] Tenopir C., Dalton E. D., Allard S., Frame M., Pjesivac I., Birch B., Pollock D., and Dorsett K. 2015 PLoS ONE 10(8) pp 1-24
[3] Yang F., Shi Y., Wu Q., Li F., Zhou W., Hu Z., Xiong N. and Zhang Y. 2019 The Survey on Intellectual Property Based on Blockchain Technology, Proc. of 2019 IEEE International Conference on Industrial Cyber Physical Systems, pp 743-748
[4] Zheng X., Zhu Y. and Si X. 2019 Applied Sciences 9(22), 4731
[5] Yu Y., Liang R. and Xu J. 2018 A Scalable and Extensible Blockchain Architecture, Proc. Of 2018 IEEE International Conference on Data Mining Workshops, pp 161-163

Acknowledgments
This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (grant No. XDA19090106), and the Shanghai Committee of Science and Technology, China under Grant 19ZR1463900.