HPChain: An MPI-Based Blockchain Framework for Data Fidelity in High-Performance Computing Systems

Abdullah Al-Mamun
aalmamun@nevada.unr.edu
University of Nevada, Reno

Tonglin Li
tonglinli@lbl.gov
Lawrence Berkeley National Laborary

Mohammad Sadoghi
msadoghi@ucdavis.edu
University of California, Davis

Linhua Jiang
honorsir@yandex.com
Fudan University

Haoting Shen
hshen@unr.edu
University of Nevada, Reno

Dongfang Zhao
dzhao@unr.edu
University of Nevada, Reno

ABSTRACT

Data fidelity is of prominent importance for scientific experiments and simulations, as the data upon which scientific discovery rests must be trustworthy and retain its veracity at every point in the scientific workflow. The state-of-the-art mechanism to ensure data fidelity is through data provenance, which keeps track of the data changes and allows for auditing and reproducing scientific discoveries. However, the provenance data itself may as well exhibit unintentional human errors and malicious data manipulation. To enable a trustworthy and reliable data fidelity service, we advocate achieving the immutability and decentralization of scientific data provenance through blockchains. The challenges of leveraging blockchains in high-performance computing (HPC) are two folds. Firstly, the HPC infrastructure exhibits incompatible characteristics to the targeting platform of existing blockchain systems; Secondly, HPC's programming model MPI alone cannot meet the reliability requirements expected by blockchains. To this end, we propose HPChain, a new blockchain framework specially designed for HPC systems. HPChain employs a new consensus protocol compatible with and optimized for HPC systems. Furthermore, HPChain was implemented with MPI and integrated with an off-chain distributed provenance service to tolerate the failures caused by faulty MPI ranks. The HPChain prototype system has been deployed to 500 cores at the University of Nevada's HPC center and demonstrated strong resilience and scalability while outperforming state-of-the-art blockchains by orders of magnitude; we are working on deploying HPChain to the Cori supercomputer hosted at the Lawrence Berkeley National Laboratory.

ACM Reference Format:
Abdullah Al-Mamun, Tonglin Li, Mohammad Sadoghi, Linhua Jiang, Haoting Shen, and Dongfang Zhao. 2019. HPChain: An MPI-Based Blockchain Framework for Data Fidelity in High-Performance Computing Systems. In *Proceedings of ACM Conference (SC'19).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

SC'19, November 2019, Denver, CO, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

Data fidelity is of prominent importance for scientific experiments and simulations, as the data upon which scientific discovery rests must be trustworthy and retain its veracity at every point in the scientific workflow. Scientific data might be intentionally fabricated or falsified, might be invalidated due to system failures, or might be accidentally modified due to human errors. Regardless of the root cause, the resultant data is not trustworthy and leads to inaccurate or incorrect scientific conclusions. As a case in point, the National Cancer Institute found 0.25% of trial data are fraudulent in the year of 2015 [4]. In earth sciences, scientists emphasized the importance of maintaining data provenance in achieving the transparency of scientific discoveries [18].

The *de facto* way to audit and reproduce scientific research and data is through data provenance, which tracks the entire lifespan of the data during the experiments and simulation at various phases such as data creation, data changes, and data archival. Conventional provenance systems can be categorized into two types: centralized provenance systems and distributed provenance systems. One representative centralized provenance system is SPADE [7], where the provenance (from various data sources) is collected and managed by a centralized relational database. Domain-specific systems based on such centralized design are also available in biomedical engineering [2], computational chemistry [13], to name a few. Although having been reasonably adopted by various disciples, the centralized provenance systems are being increasingly criticized by researchers and scientists who face the exponentially-grown data in terms of velocity and volume, the so-called "Big Data." In essence, the centralized system, due to the performance bottleneck on the centralized node (not to mention its potential single point of failure), cannot meet the performance expectation of many data-intensive scientific applications and to this end, we started witnessing the boom of various distributed approaches toward **scalable provenance** [3, 22]. Indeed, those distributed provenance systems, mostly built upon distributed file systems as opposed to centralized databases, eliminated the performance bottleneck and delivered orders of magnitude higher performance than centralized approaches.

As a double-edged sword, however, distributed provenance systems pose a new concern on the provenance itself: the chance that the provenance is tempered with increases from $f\%$ to $n \cdot f\%$, where f indicates the failure rate of a single node and n the total number of nodes. Moreoever, a natural question then is, **while the provenance is supposed to audit the execution of the application, who then should audit the provenance?** Do we need

to build the provenance of provenance? So the recursion goes on and on, indefinitely. To this end, **decentralized provenance systems** were recently proposed inspired by blockchains. These systems (e.g., ProvChain [10], SmartProvenance [15]) are also called blockchain-based provenance systems that are both temper-evident and autonomous, thus guarantee trustworthy data provenance.

Multiple issues exist for applying a blockchain-based provenance system to HPC. Firstly, it is not hard to see that its space efficiency is low, its network traffic consumption is high, and the consistency is always a challenging problem. Besides, all these blockchain-based provenance systems are built in such a way that the underlying blockchain infrastructure is a black box and the provenance service works as a higher-level application by calling the programming interfaces provided by the blockchain infrastructure such as Hyperledger Fabric [8] and Ethereum [6]. In the best case, the provenance services might miss optimization and customization opportunities because the former cannot (or, prohibitively expensive and complicated to) modify the lower blockchain layer; to make it worse, the applicability of those blockchain-based provenance systems is constrained by the underlying blockchain infrastructure.

2 PROPOSED APPROACH

We propose HPChain, a new blockchain framework, specially designed for HPC systems. HPChain exhibits two key technical novelties compared with mainstream blockchain systems.

HPChain employs a new consensus protocol tailored to HPC system infrastructure. Extending a shared-nothing protocol into a shared-storage one might sound trivial, this turns out to be a challenging problem if we need to maintain the high security (i.e., the 50% quorum voting) after the extension. Specifically, the original blockchain consensus assumes that no single party could control more than 50% of the nodes[1] (each of which has its distinct storage device), however, in a shared-storage infrastructure, a single compromised storage node would be equivalent to multiple nodes from the blockchain's perspective. To this end, we design the protocol in such a way that both the compute and storage nodes are taken into account for the quorum voting (i.e., the consensus protocol), but only the former are involved in the validation procedure; the idea was partly inspired by our prior experience of building an in-memory blockchain system [1]. Furthermore, we have proven the safety of the proposed consensus protocol.

HPChain was implemented with MPI [12] and will be integrated with an off-chain distributed provenance service. As such, HPChain can be deployed to an environment without the TCP/IP stack. To overcome the resilience problem (i.e., a single MPI rank cracks down the entire blockchain job), HPChain employs a distributed, off-chain data provenance module that periodically records the provenance of the job execution. The off-chain provenance is complementary to the MPI-based on-chain data: if the HPChain fails due to a faulty rank, the off-chain provenance can continue auditing the execution of the monitored job until HPChain recovers and synchronizes with the off-chain provenance. The off-chain provenance module will be extended from our prior work [22] built upon the ZHT distributed key-value store [9].

[1] As known as "51% attack" in blockchains.

3 PRELIMINARY RESULTS

Test Bed. The HPChain prototype system was deployed to 500 cores at the University of Nevada's HPC cluster Pronghorn [14]. Pronghorn is composed of CPU, GPU, and storage subsystems interconnected by a 100 Gb/s non-blocking Intel Omni-Path fabric. Each compute node is installed with Ubuntu 16.04, Python 3.7.0, NumPy 1.15.4, mpi4py v2.0.0, and mpich2 v1.4.1.

Workloads. We took YCSB [20] as the transaction workload to evaluate the performance of HPChain; YCSB is widely accepted in measuring the performance of blockchain systems, for example in BlockBench [5] (for private blockchains) and BlockLite [19] (for public blockchains). In YCSB, each node is responsible for performing both read and write operations for each transaction. Specifically, a transaction data movement is in the form of, semantically, "data: place A to place B." In HPChain, each block contains four transactions, and we deploy more than two million transactions (2,013,590) to the prototype.

Results. We evaluated the HPChain prototype system from four perspectives: resilience, throughput, latency, and scalability. During the execution of processing the two million transactions, up to four cores (out of 500) crashed, leading to 99.4% nodes holding valid blockchain data. Note that the consensus protocol requires only 51% nodes correctly running. HPChain delivers up to 6×, 12× and 75× higher throughput than Hyperldeger, Ethereum, and Parity respectively. Throughput is defined as the number of processed transactions per second (txn/s). HPChain incurs orders of magnitude lower latency than the state-of-the-art. For each transaction, the average delay of HPChain is up to 1000×, 400×, and 5000× shorter than that of Hyperledger Fabric, Parity, and Ethereum, respectively. In terms of scalability, the throughput is only slightly decreased from 100 cores to 500 cores: 36,001 txn/s to 33,543 txn/s.

4 FUTURE WORK

One limitation of off-chain provenance is that when HPChain is offline, the off-chain data becomes vulnerable (without the validation of on-chain data). It would be a more desirable solution if we can, somehow, turn the off-chain data into another blockchain whose MPI session is independent than that of the main HPChain. In other words, there are two blockchains (thus, two MPI "worlds") concurrently running for the data fidelity job, one as the primary chain and the other as the secondary chain. The idea is not entirely new: it was called a *sidechain* [16] whose original goal was to improve the performance of the master blockchain by offloading some specific tasks to the sidechain. However, sidechain has not been widely adopted due to its huge performance overhead—it usually takes 1-2 days to switch between the main chain and its sidechain. We are designing new protocols to reduce the switching overhead.

We are also working to deploy HPChain to the Cori [17] supercomputer hosted at the Lawrence Berkeley National Laboratory. On the one hand, we plan to evaluate the effectiveness and performance of HPChain with more real-world scientific applications at extreme scales, e.g., astronomy and medical imaging we previous studied in [11], bioinformatics and seismology we previously studied in [21]). On the other hand, we will also investigate how to leverage the HPC-compatible blockchain framework HPChain in more use scenarios such as fault tolerance and distributed caching.

REFERENCES

[1] A. Al-Mamun, T. Li, M. Sadoghi, and D. Zhao. In-memory blockchain: Toward efficient and trustworthy data provenance for hpc systems. In *IEEE International Conference on Big Data (BigData)*, 2018.

[2] T. Clark, P. N. Ciccarese, and C. A. Goble. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*, 5(1):28, Jul 2014.

[3] D. Dai, Y. Chen, P. Carns, J. Jenkins, and R. Ross. Lightweight provenance service for high-performance computing. In *International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2017.

[4] Data fraud in clinical trials. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340084/, Accessed 2019.

[5] T. T. A. Dinh, J. Wang, G. Chen, R. Liu, B. C. Ooi, and K.-L. Tan. Blockbench: A framework for analyzing private blockchains. In *ACM International Conference on Management of Data (SIGMOD)*, 2017.

[6] Ethereum. https://www.ethereum.org/, Accessed 2018.

[7] A. Gehani and D. Tariq. SPADE: Support for Provenance Auditing in Distributed Environments. In *Proceedings of the 13th International Middleware Conference (Middleware)*, 2012.

[8] Hyperledger. https://www.hyperledger.org/, Accessed 2018.

[9] T. Li, X. Zhou, K. Brandstatter, D. Zhao, K. Wang, A. Rajendran, Z. Zhang, and I. Raicu. ZHT: A light-weight reliable persistent dynamic scalable zero-hop distributed hash table. In *Proceedings of IEEE International Symposium on Parallel and Distributed Processing (IPDPS)*, 2013.

[10] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla. Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, 2017.

[11] P. Mehta, S. Dorkenwald, D. Zhao, T. Kaftan, A. Cheung, M. Balazinska, A. Rokem, A. Connolly, J. Vanderplas, and Y. AlSayyad. Comparative evaluation of big-data systems on scientific image analytics workloads. In *Proceedings of the 43rd International Conference on Very Large Data Bases (VLDB)*, 2017.

[12] MPICH. http://www.mpich.org/, Accessed 2019.

[13] E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng, and T. Ferrin. Ucsf chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, Oct 2004.

[14] Pronghorn. https://www.unr.edu/research-computing/hpc, Accessed 2019.

[15] A. Ramachandran and M. Kantarcioglu. Smartprovenance: A distributed, blockchain based dataprovenance system. In *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, CODASPY '18, pages 35–42, 2018.

[16] Sidechains. https://blockstream.com/sidechains.pdf, Accessed 2019.

[17] The Cori Supercomputer. http://www.nersc.gov/users/computational-systems/cori, Accessed 2019.

[18] The Importance of Data Set Provenance for Science. https://eos.org/opinions/the-importance-of-data-set-provenance-for-science, Accessed 2019.

[19] X. Wang, A. Al-Mamun, F. Yan, and D. Zhao. Toward accurate and efficient emulation of public blockchains in the cloud. In *12nd International Conference on Cloud Computing (CLOUD)*, 2019.

[20] YCSB. https://github.com/brianfrankcooper/YCSB/wiki/Core-Workloads, Accessed 2018.

[21] D. Zhao, N. Liu, D. Kimpe, R. Ross, X.-H. Sun, and I. Raicu. Towards exploring data-intensive scientific applications at extreme scales through systems and simulations. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 27(6), 2016.

[22] D. Zhao, C. Shou, T. Malik, and I. Raicu. Distributed data provenance for large-scale data-intensive computing. In *IEEE International Conference on Cluster Computing (CLUSTER)*, 2013.