Low-latency network-scalable Byzantine Fault-tolerant Replication

Ines Messadi
TU Braunschweig, Germany
messadi@ibr.cs.tu-bs.de

ABSTRACT

Byzantine-fault tolerant (BFT) protocols allow mitigating a wide range of failures, thereby ensuring the availability and resiliency of a system. Yet, such protocols are considered costly in terms of message complexity and resource usage. The cost is to a large part caused by the multiple rounds of communication that are necessary to reach the agreement. With the availability of Remote Direct Memory Access (RDMA) in data centers, this communication overhead can be reduced. In fact, RDMA interconnects have been shown to provide very low latencies and high performance, which promises accelerated BFT systems deployable in modern data centers. However, using RDMA to speed-up BFT protocols is challenging in three ways: First, the performance is defined by the efficient use of RDMA primitives and the associated resources. Second, RDMA has been proven to open up some security issues, which can be detrimental to BFT systems. Third, scalability is still a concern for RDMA-based systems and an open issue of BFT.

KEYWORDS

RDMA, Performance, Byzantine fault tolerance

1 INTRODUCTION

Providing fault-tolerance is important for today's network-based services because their unavailability might entail serious problems. Accordingly, Byzantine fault tolerance (BFT) is a technique allowing to mitigate failures of many kinds and even malicious nodes. However, although these protocols promise resilience and reliability, they are still not used in practice because they require a lot of computation and network resources. In addition, they cannot deliver the requested low latency to be deployed in modern data center applications.

So far, many optimized BFT systems have been proposed during the last decades [1, 3, 11]. However, they all rely on a TCP/IP based communication, which is known to incur an unwanted high latency, as well as a limited throughput and scalability. In this context, Remote Direct Memory Access (RDMA) is identified as a promising solution because of its exceptional performance. In a nutshell, it is a novel technology that enables direct data movement between the memory of remote computers in a zero-copy manner, without the support of the operating system. Consequently, it helps to reduce the CPU load and to decrease the network overhead. The particular feature of RDMA is that it offers two different communication operations, the one-sided operations, allowing to directly read (RDMA read) or write (RDMA write) memory of remote machines or the two-sided communication where both applications are involved in the data transfer.

A number of recent research works has investigated the smart utilization of RDMA-based communication [5, 9, 12, 13, 14]. Indeed, despite its notable performance, it hides some design choices that could affect the expected improvement. We primarily identified three design challenges: First, every system that aims to leverage RDMA features will face the choice on how to best use the two different RDMA core primitives, as well as on how to alleviate the cost of memory registration required before any data transfer [6]. Second, the choice of RDMA operations and transport mode affect the scalability of the system. Third, as the memory keys are vulnerable, there is a threat that a malicious connected client would try to corrupt a remote memory region. Hence, it is important to design new security features for mitigating such vulnerability.

To summarize, in order to exploit the full potential of RDMA interconnects, we plan to devise the first RDMA-tailored BFT protocol through an efficient design, leading to an unprecedented performance of BFT systems.

2 RELATED WORK

First RDMA-enabled consensus protocols have been proposed [12, 14]. DARE [12], an RDMA-tailored replicated state machines protocol, substantiated the benefits of RDMA communication and proved a latency of fewer than 15 microseconds. The protocol is optimized for one-sided primitives. It replicates state machine updates through RDMA write operations. Consequently, the overhead is removed from remote parties, improving the availability and reliability of the system. APUS [14], is another recent work, which combines RDMA with agreement protocols and outperforms the traditional state machine replication protocols in term of scalability and latency. However, both protocols solve the problem only for fail-stop failures. Besides that, it is just a start and the question is how BFT systems should be adapted for RDMA primitives. To our knowledge, there is no previous work that assumed a Byzantine fault model.

3 APPROACH AND RESEARCH PLAN

Our approach targets to explore the design options of building an extended BFT framework tailored for the RDMA communication model. It is planned to use Hybster [1], a parallelizable state-machine protocol which assumes a hybrid fault model, as a basis for the planned research work. The protocol achieves over 1 million operations per second but needs multiple Ethernet cards to scale. In this context, RDMA offers a higher bandwidth. We expect RDMA to extend the performance and capabilities of Hybster in terms of latency and scalability. Although the protocol is using a trusted execution environment based on Intel's Software Guard Extensions

EuroDW'18, April 23, 2018, Porto, Portugal
2018.

(SGX), our approach is independent and intended for any traditional BFT protocol. Further, RDMA contradicts the concept of the memory protection offered by the SGX enclaves, as DMA accesses are not allowed.

According to the described challenges, the research plan is split into three milestones:

Devise of a new RDMA-tailored BFT protocol Initially, we start by analyzing the existing SMR and BFT replication protocols for their usefulness in context with RDMA. In order to fully take advantage of RDMA without harming Hybster's performance, we plan to devise an ideal RDMA setting for BFT systems. A number of recent works has largely focused on the usage of the one-sided and the two-sided RDMA primitives [4, 10, 13]. In fact, although the attractiveness of the one-sided operations, the RDMA read imposes many communication rounds, and a coordination of the issuing sides, while the RDMA write would saturate the servers with a scaling number of replicas [13]. On the other hand, the connected RDMA transport implies that these operations limit the scalability of the application. DARE is taking advantage of the one-sided write operations to achieve a good performance but does not maintain the lower consensus latency with an increasing number of connections. Our approach is to combine the RDMA one-sided communication and the two-sided for an ideal setting to reduce the replication cost and aim for a scalable, secure, and fast RDMA-based BFT. We then investigate our first evaluations of RDMA performance using different workloads. Besides the new design, new optimizations should be added or the existing BFT optimization techniques need to be adapted. Hybster buffer management and memory registration will be handled according to the related work [6].

Analyze the use of RDMA for security issues and implement counter-measures This milestone will be running in parallel with the first. As mentioned so far the memory keys used for RDMA are not secure and can be guessed by a malicious replica. So that, at some point, the latter replica can try to write corrupted data into the memory without having access permission. Moreover, many other forms of attacks are possible such as spoofing and denial of service [8]. As security is a crucial aspect in consensus models, we plan to consider RDMA vulnerabilities and implement countermeasures. The idea is to design new security features for mitigating such attacks and to design new mechanisms for the resilient use of RDMA in context of BFT.

Implement example applications: coordination service and blockchain ordering service The prominent emerging of Blockchain technology makes it an important application for the newly designed BFT framework. The core idea of blockchains is a distributed ledger, composed of blocks of records, where the blocks are related to each other through a cryptographic hash. In such systems, nodes may crash or behave maliciously. Thus, an RDMA-based consensus protocol can be used to have an accurate blockchain.

A second application is to build a coordination service having a similar interface as the Yahoo ZooKeeper [7], and Chubby [2], which supports strong consistency and availability. We expect the new implementation to benefit from the improved performance and be Byzantine fault tolerant on top.

4 SUMMARY

The aim of our approach is to build a framework of an RDMA-tailored consensus algorithm. So far, there is a lack of research that aims to integrate RDMA into consensus protocols. Our deduction is that its integration will decrease the message exchange latency and therefore the time to reach an agreement, as well as reducing the risk of bottlenecks in the nodes. Such improvement will ease the deployment of BFT protocol into real-world programs and modern data centers.

REFERENCES

[1] Johannes Behl, Tobias Distler, and Rüdiger Kapitza. "Hybrids on Steroids: SGX-Based High Performance BFT". In: *Proceedings of the Twelfth European Conference on Computer Systems*. EuroSys '17. Belgrade, Serbia: ACM, 2017, pp. 222–237. ISBN: 978-1-4503-4938-3. DOI: 10.1145/3064176.3064213.

[2] Mike Burrows. "The Chubby lock service for loosely-coupled distributed systems". In: *Proceedings of the 7th symposium on Operating systems design and implementation*. USENIX Association. 2006, pp. 335–350.

[3] Miguel Castro, Barbara Liskov, et al. "Practical Byzantine fault tolerance". In: *OSDI*. Vol. 99. 1999, pp. 173–186.

[4] Aleksandar Dragojevic, Dushyanth Narayanan, and Miguel Castro. "RDMA Reads: To Use or Not to Use?" In: *IEEE Data Eng. Bull.* 40.1 (2017), pp. 3–14.

[5] Aleksandar Dragojević et al. "FaRM: Fast remote memory". In: *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*. 2014, pp. 401–414.

[6] Philip Werner Frey and Gustavo Alonso. "Minimizing the hidden cost of RDMA". In: *Distributed Computing Systems, 2009. ICDCS'09. 29th IEEE International Conference on*. IEEE. 2009, pp. 553–560.

[7] Patrick Hunt et al. "ZooKeeper: Wait-free Coordination for Internet-scale Systems." In: *USENIX annual technical conference*. Vol. 8. 9. Boston, MA, USA. 2010.

[8] E. Deleganes J. Pinkerton. *Direct Data Placement Protocol (DDP) / Remote Direct Memory Access Protocol (RDMAP) Security*. 2007.

[9] Anuj Kalia, Michael Kaminsky, and David G Andersen. "FaSST: Fast, Scalable and Simple Distributed Transactions with Two-Sided (RDMA) Datagram RPCs." In: *OSDI*. Vol. 16. 2016, pp. 185–201.

[10] Anuj Kalia, Michael Kaminsky, and David G Andersen. "Using RDMA efficiently for key-value services". In: *ACM SIGCOMM Computer Communication Review*. Vol. 44. 4. ACM. 2014, pp. 295–306.

[11] Ramakrishna Kotla et al. "Zyzzyva: speculative byzantine fault tolerance". In: *ACM SIGOPS Operating Systems Review*. Vol. 41. 6. ACM. 2007, pp. 45–58.

[12] Marius Poke and Torsten Hoefler. "DARE: High-performance state machine replication on RDMA networks". In: *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*. ACM. 2015, pp. 107–118.

[13] Maomeng Su et al. "RFP: When RPC is faster than server-bypass with RDMA". In: *Proceedings of the Twelfth European Conference on Computer Systems*. ACM. 2017, pp. 1–15.

[14] Cheng Wang et al. "APUS: Fast and scalable Paxos on RDMA". In: *Proceedings of the 2017 Symposium on Cloud Computing*. ACM. 2017, pp. 94–107.