Ph.D. Dissertation Summary

Global-Scale Data Management

Faisal Nawab
http://nawab.me

Processing large quantities of data is becoming more ubiquitous and is the driving force behind the sustained growth and impact of Internet Services and Big Data analytics. The way data-intensive applications are deployed has been radically transformed by the cloud computing paradigm realized through massive-scale datacenters. However, datacenter-scale failures have occurred numerous times in the past and continue occurring many times annually due to various events such as power outages and natural disasters. These failures impact all services in a datacenter for extended periods of time and cause disruption to a large number of users, leading to losses in revenue and utility of Internet and Big Data applications. Also, deployment at a single datacenter gives rise to very high latency for user requests that are geographically distant.

Recently, Big Data and large-scale systems have addressed datacenter-scale failures and high service response times by distributing applications and data globally across multiple datacenters, thus providing datacenter-scale fault-tolerance and data proximity to users. Methods based on asynchronous replication are widely used in practice to keep the replicas up-to-date with each other. Asynchronous replication, however, is vulnerable to data losses and inconsistency when failures occur, which threaten the integrity of the applications. This also complicates application design and recovery for developers and system administrators. This is now broadly recognized as a serious drawback that is limiting the adoption of multi-datacenter deployments. This has led to a growing interest from both industry and academia in *global-scale* systems with stringent *consistency* guarantees.

Research Summary. Through my dissertation research, *I investigate the fundamental challenges and best practices for designing consistent **Global-Scale Data Management (GSDM)** systems*. Consistent GSDM systems allow developers and administrators to deploy applications globally without sacrificing the benefits of easy-to-use and easy-to-manage traditional data management systems while maintaining the integrity of the applications. Thus, it facilitates the adoption of the fault-tolerance and performance benefits of GSDM for a wide-range of Big Data applications and Internet Services. Consistency, however, requires *coordination* between replicas to ensure that requests do not overwrite or contradict each other. Global-scale coordination is expensive due to the large wide-area latency between datacenters. The wide-area latency ranges from hundreds of milliseconds up to several seconds. This is 2–4 orders of magnitude larger than the typical communication latency between replicas within a single datacenter. Such a high latency, studies show, drives a significant percentage of users to abandon the application, *e.g.*, more than half the users leave applications with a multi-second delay. High latency, in addition to being a disadvantage in itself, also makes other performance characteristics, such as scalability and throughput, more challenging.

In my work, I have focused on addressing the grand challenges of adopting multi-datacenter deployments for Big Data and large-scale systems. Application developers and administrators are torn between easy-to-use consistent GSDM systems that perform poorly and high-performance asynchronous replication that is difficult to develop and use correctly. My work aims at solving this dilemma by *designing a new generation of consistent GSDM systems that are scalable and achieve high performance with significantly improved coordination latency compared to traditional consistent GSDM systems.* My approach to research work begins with studying aspects of the global-scale environment and GSDM systems to extract the main challenges and problems of GSDM. I then use newfound insights and understanding from such studies to propose fundamental design principles that improve the performance of consistent GSDM systems. In the following, I summarize my dissertation research [2, 9].

Global-Scale Data Management

Understanding Global-Scale Coordination. To design efficient consistent GSDM systems, it is essential to understand the fundamental challenges and characteristics of global-scale coordination. To gain such an understanding, I started my Ph.D. work with studying existing consistent GSDM systems to extract the design characteristics that led to poor performance. With this understanding, I propose evolutions of these protocols that perform better in the global-scale environments. Our work on **Paxos-CP** [10] is a product of such a study on a consensus protocol called Paxos. Paxos is an essential component in many large-scale consistent systems such as Google Megastore. However, it was designed for distributed systems without considering the intricate challenges of global-scale Big Data applications. Paxos orders requests sequentially, thus limiting its throughput in environments with large communication latency. Paxos-CP identifies the sequential guarantee of Paxos as too strict for data management (transactional) workloads that only need *serializable* consistency. With this observation, techniques to reach consensus that allow more concurrency while maintaining serializability are developed. Our other work on **Replicated Commit** [1] studies the limitations of a traditional approach of fault-tolerance for partitioned data that is adopted widely by GSDM systems such as Google Spanner. In this traditional approach, each partition is made fault-tolerant by replicating a log of processing steps to other datacenters. Multiple steps are replicated for each request, causing coordination latency to be amplified. Replicated Commit proposes a method to replicate the whole request as a single piece—rather than replicating each step—even for systems that partition data like Spanner. Thus, only a single wide-area round of communication is needed, reducing the request latency significantly. Paxos-CP and Replicated Commit are two of the first studies that shed light on the effect of the wide-area latency limit on coordination latency and proposed designs specifically to improve coordination performance in a global-scale environment.

Breaking the Latency Barrier of the Request-Response Paradigm. The studies conducted for Paxos-CP and Replicated Commit exposed a fundamental coordination latency limit. This limit is due to the polling nature of traditional protocols that I call the Request-Response paradigm. In the Request-Response paradigm, the coordination for a request starts *after* the request is made, where the replica that received it polls other replicas to inquire about their state and detect conflicts. The request is served only after receiving a *response* from other replicas. This makes a Round-Trip Time (RTT) of communication inevitable—an expensive cost in GSDM. This leads to the question: *Is it possible to avoid the Request-Response paradigm of coordination?* My work on **Message Futures** [3] demonstrates this possibility by an observation that coordination of future requests can start before they arrive. As requests arrive, they are assigned to a predetermined future coordination point. I call this approach *Futures Coordination*. Coordination points are judiciously calculated to ensure conflicts are detected. A coordination point still needs an RTT for coordination. However, because a request is assigned to a coordination point that already started, the request's observable latency is less than RTT. Message Futures is the first protocol that shows the possibility of faster-than-RTT coordination for all the replicas of a distributed system. Also, it introduces Futures Coordination, a new approach to coordination that overcomes the limitations of the Request-Response paradigm.

Theoretical Lower-Bound on Coordination Latency. Breaking the RTT latency barrier via the Futures Coordination paradigm invalidates the previously held convention that coordination cannot be performed faster than the RTT latency. Thus, it opens the question: *What is the lower-bound on coordination latency?* This is a fundamental question in understanding the extent of the effect of the wide-area latency limit on coordination latency. Such a lower-bound, if proven, will provide system designers and researchers with a theoretical foundation on what is achievable by current and future systems.

To tackle the question of lower-bound coordination latency, I begin by formalizing and modeling the concept of coordination, which is the process of detecting a conflict between two requests. This model of coordination is then used to answer the question: *What are the cases that make detecting a conflict between two requests impossible*? To answer this question, I observe that for any potential conflict between two requests, a and b, one of them must know about the other before making the decision (commit or abort). If they both commit without knowing about the other, then they do not detect the conflict, possibly leading to a consistency violation. My work shows that these cases are inevitable if the coordination latency of a *plus* the coordination latency of b is less than the RTT between the datacenters hosting them. Thus, for any consistent global-scale system, the sum of the coordination latency of any two transactions must be greater than or equal to the RTT between their host datacenters [8]. For example, it is possible that the latency of both a and b is equal to half the RTT between their host datacenters. The coordination model inspired a protocol based on Futures Coordination, called **Helios** [8] that theoretically achieves the lower-bound, thus proving that the lower-bound is tight.

The lower-bound result shows that the coordination latency can be faster than what is previously achieved by traditional protocols and even faster than what is achieved by Message Futures. The model of coordination, in addition to being essential for deriving the lower-bound, advances our understanding of the cost of global-scale coordination. It also brings a newfound understanding of the latency characteristics of traditional and Futures Coordination protocols.

Global-Scale Data Communication. Large-scale Big Data applications process massive amounts of data that cannot be supported by traditional communication protocols. I address this problem by investigating communication designs that scale to the needs of large-scale applications and be able to support coordination-oriented communication, such as the communication needed for Message Futures, Helios, and other consistent GSDM protocols. **Chariots** [7] is the product of this investigation. To scale Chariots to large-scale workloads, the task of communication is made a priority. Chariots manages a group of machines dedicated for multi-datacenter communication that provides global-scale communication as a service to applications. The problem of communications scalability is tackled by observing that the traditional total ordering guarantees of communication protocols are too strict for coordination-related communication. Rather, it turns out that *causal-order* guarantees are sufficient for coordination-related communication. The design of Chariots proposes novel methods of managing distributed causally-ordered communication that enable it to scale to the demands of large-scale applications.

Machine Learning with Globally-Generated Data. Machine learning is essential for Big Data analytics. This motivated my work on **COP** [6] that specifically targets efficiently supporting global-scale machine learning workloads. In typical global-scale machine learning, data is collected at different locations around the world and then processed at a centralized location. COP targets improving the learning performance for this *Collect then Learn* pattern. COP's main purpose is to pre-process data as it is collected so that when it is received at the centralized location it can be processed faster. COP's approach ensures a partial order of the execution that will preserve the used machine learning algorithm's theoretical properties. The pre-processing allows COP to enforce the partial order with light-weight operations that outperform traditional methods.

A data infrastructure that spans both cloud and edge nodes[1]**.** The utilization of edge nodes is inevitable for the success and growth of emerging low latency applications, such as Augmented and Virtual Reality (AR/VR) and vehicular networks [5]. Such applications have stringent latency requirements that the current cloud model cannot satisfy. This is due to the large communication latency between users and their closest datacenter. We propose Dynamic Paxos (DPaxos) [4] to extend GSDM to span edge nodes closer to users. DPaxos is a variant of paxos—a widely-used protocol in distributed data management systems. One of the main challenges of utilizing edge nodes for data management is that coordination protocols, such as paxos, requires communicating with a significant subset (*i.e.*, majority quorum) of all nodes to reach agreement. With edge nodes—that are large in number and distributed across large distances—these protocols become infeasible. DPaxos solves this problem by adopting recent advances in the paxos protocol—proposed in Flexible Paxos—that allows a more flexible design of quorums. DPaxos extends this design and proposes a dynamic quorum allocation strategy that starts out small—spanning only a handful of close-by edge nodes—and only expands in the presence of conflicts. Unlike other paxos variants, DPaxos is the only paxos variant that can have non-intersecting Leader Election and Replication quorums. This, together with the dynamic nature of DPaxos, enable extending GSDM systems efficiently to span edge nodes closer to users.

References

[1] H. Mahmoud, F. Nawab, A. Pucher, D. Agrawal, and A. El Abbadi. Low-latency multi-datacenter databases using replicated commit. *Proc. VLDB Endow.*, 6(9):661–672, July 2013.

[2] F. Nawab. *Global-Scale Data Management with Strong Consistency Guarantees*. PhD thesis, University of California, Santa Barbara, 2018.

[3] F. Nawab, D. Agrawal, and A. El Abbadi. Message futures: Fast commitment of transactions in multi-datacenter environments. In *CIDR*, 2013.

[4] F. Nawab, D. Agrawal, and A. El Abbadi. Dpaxos: Managing data closer to users for low-latency and mobile applications. In *SIGMOD*, pages 1221–1236. ACM, 2018.

[5] F. Nawab, D. Agrawal, and A. El Abbadi. Nomadic datacenters at the network edge: Data management challenges for the cloud with mobile infrastructure. In *EDBT*, pages 497–500, 2018.

[6] F. Nawab, D. Agrawal, A. El Abbadi, and S. Chawla. COP: Planning conflicts for faster parallel transactional machine learning. In *EDBT*, 2017.

[7] F. Nawab, V. Arora, D. Agrawal, and A. El Abbadi. Chariots: A scalable shared log for data management in multi-datacenter cloud environments. In *EDBT*, pages 13–24, 2015.

[8] F. Nawab, V. Arora, D. Agrawal, and A. El Abbadi. Minimizing commit latency of transactions in geo-replicated data stores. In *SIGMOD*, pages 1279–1294, 2015.

[9] F. Nawab, V. Arora, V. Zakhary, D. Agrawal, and A. El Abbadi. A system infrastructure for strongly consistent transactions on globally-replicated data. *IEEE Data Eng. Bull.*, 40(4):3–14, 2017.

[10] S. Patterson, A. J. Elmore, F. Nawab, D. Agrawal, and A. El Abbadi. Serializability, not serial: Concurrency control and availability in multi-datacenter datastores. *Proc. VLDB Endow.*, 5(11):1459–1470, July 2012.

[1] This work has been done during the Ph.D. dissertation studies but was published after graduation.