

Towards Distributed Coordination for Fog Platforms

Tobias Pfandzelter, Trever Schirmer, David Bermbach
 Technische Universität Berlin & Einstein Center Digital Future
 Mobile Cloud Computing Research Group
 {tp,ts,db}@mcc.tu-berlin.de

Abstract—Distributed fog and edge applications communicate over unreliable networks and are subject to high communication delays. This makes using existing distributed coordination technologies from cloud applications infeasible, as they are built on the assumption of a highly reliable, low-latency datacenter network to achieve strict consistency with low overheads. To help implement configuration and state management for fog platforms and applications, we propose a novel decentralized approach that lets systems specify coordination strategies and membership for different sets of coordination data.

Index Terms—fog computing, edge computing, service orchestration

I. INTRODUCTION

To leverage fog and edge computing, fog application and data distribution platforms have been proposed that shift the burden of managing the heterogeneous, distributed infrastructure. As shown in Figure 1, these platforms manage the replication and orchestration of services, e.g., function instances in a FaaS platform [1]–[4], and data, e.g., application state [5]–[7]. A key requirement here is the coordination among fog nodes for exchange of configuration and management data, e.g., access control, monitoring data, naming, or routing, not unlike coordination in a distributed cloud system. Existing distributed application coordination methods and systems, e.g., *etcd*¹ or *Apache Zookeeper*², rely on highly available datacenter networks to achieve equally highly available, strictly consistent coordination with low overhead. In contrast to cloud data centers, fog resources are typically distributed over wide geographical areas and communicate over the Internet instead of dedicated (private) networks, making the existing cloud coordination approaches infeasible: As a result of CAP [8] and PACELC [9], the frequent network partitions lead to service disruption, making the platform unavailable. In addition, the high communication delay between distant fog and edge resources would lead to considerable latency for both centralized and decentralized strictly consistent coordination, either as a result of the long path between a client and central server, or of the path distance between participants. Finally, fog systems usually comprise thousands of servers in the edge, cloud, and in-between – orders of magnitudes more than in cloud systems. These sites are also heterogeneous in their capabilities, from single-board edge computers to clusters of cloud virtual machines [10].

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 415899119.

¹<https://etcd.io/>

²<https://zookeeper.apache.org/>

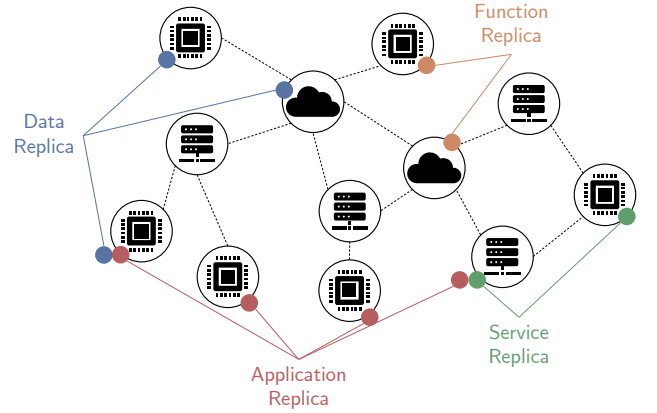


Fig. 1. A fog platform manages the replication of application components by abstracting from the underlying geo-distributed and heterogeneous fog infrastructure.

Two factors control different quality of service (QoS) aspects of coordination approaches, such as availability and access latency: (i) *how* coordination state is exchanged, i.e., synchronously or asynchronously, or which participant may start an update [11], and (ii) *who* participates in the update exchanges, i.e., the set of nodes that have copies of the coordination data or must reach consensus on an update. While these factors also apply in cloud applications, their impact in a fog environment is decidedly different, e.g., node membership must also take network distance into account.

The tradeoffs at play here make a one-size-fits-all solution infeasible for fog environments. The optimal choice depends, among others, on the type of data, data access patterns, the (geo-)distribution of clients accessing the data, and QoS requirements. In this paper, we thus propose a fog coordination approach that lets developers make these choices per set of coordination data, requiring few changes to the platforms and applications.

II. RELATED WORK

With an increased research interest in fog computing, numerous fog platforms for application and data management have been proposed that mostly adapt global, strictly consistent coordination from cloud platforms. *FBase* [6], [7] uses a centralized naming service for configuration and access control, similarly to the use of *Chubby* [12] in the *GFS* [13] and *BigTable* [14] data store systems. In the *FBase* implementation,

this naming service is based on Apache Zookeeper. *Fogernetes* [15] extends Kubernetes to the fog using a centralized master that handles configuration. While these approaches allow strictly consistent configuration management without a performance impact for constrained edge devices, the long delay for reads and writes can impact access latency, and any network partition can make the system unavailable.

An alternative approach is a decentralized configuration management. Eberhardt et al. [16] propose using conflict-free replicated data types (CRDTs) to distribute configuration data for Docker containers. Similarly, Jeffery et al. [17] propose replacing central configuration in Kubernetes with an eventually consistent, distributed approach. Using CRDTs, write latency is reduced as conflicts are resolved lazily. Although decentralized approaches scale well, they also require global synchronization and data exchange, which can be bandwidth-intensive. Furthermore, the eventual consistency can be problematic for some use-cases, e.g., access control, where a user might expect a permission change to be final.

III. DISTRIBUTED CONFIGURATION

As we have motivated, in fog platform coordination, different data types require a different tradeoff between latency and consistency. In contrast to choosing only *either* strict or eventual consistency, our approach lets systems specify how a set of configuration or state data is managed. Additionally, we propose the notion of coordination memberships, where data is only coordinated within a subset of all fog sites in order to limit update dissemination in the geo-distributed fog environment. Here, we use the concept of *fog nodes*, a virtual group of machines all running in the same fog location. Our plan is to implement this approach with a middleware.

A. Coordination Strategy

Fog platforms require coordination on different types of control data: Naming data that globally identifies different fog nodes, nodes that replicate a certain service, coordination on data sharding in light of high node churn, or client permissions. For these different types of control data, it is desirable to be able to specify a coordination strategy to increase consistency where necessary to ensure correct system operation, or to decrease consistency to achieve a lower latency. We propose between two such strategies, namely:

1) *Eventual Consistency with CRDTs*: Using CRDTs to manage configuration data allows nodes to write changes without consulting other nodes in the system. This reduces request latency and keeps the system available in case of network partitions. While the observed configuration state is guaranteed to converge in the absence of errors and updates, data will often be stale. This may be achieved with lazy, intermittent synchronization or gossip among nodes.

2) *Strict Consistency with Consensus*: Reaching consensus on configuration updates incurs a significant communication delay in a geo-distributed system and might even be impossible when the network is partitioned. This is desirable in cases where the correct operation of the system depends on strictly

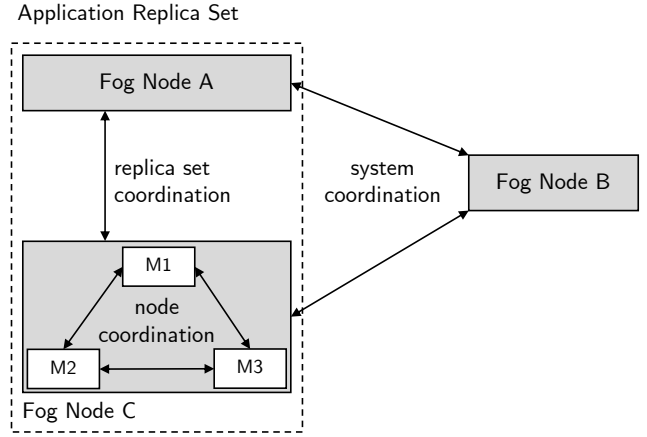


Fig. 2. We identify three levels on which fog nodes require coordination: (i) global system coordination among all nodes, required, e.g., for naming; (ii) coordination within a replica set to coordinate, e.g., access control for a service replica; and (iii) node coordination among multiple machines of the same node, e.g., a cluster of datacenter instances.

consistent coordination. We may achieve strict consistency by using a consensus protocol among nodes or even with a central coordinator (essentially, a primary copy approach [11]).

B. Coordination Levels

Limiting the number of participants for both CRDT message distribution and majority quorums can improve performance and reduce network usage. We thus introduce different levels of coordination, as shown in Figure 2: (i) the system level, (ii) the replica set level, and (iii) the node level. With replica set, we here refer to the group of fog machines jointly managing an application-level replica, e.g., all copies of a data item in the case of a fog storage system, or all instances of a function in case of a FaaS platform.

1) *System Coordination*: Some data, such as naming data, which must be globally unique and known by all nodes in the system, require coordination among the entire fog system. As a result, a high communication delay can be expected that imposes a high overhead for consensus with higher write ratios and increases the message dissemination cost for update broadcasts.

2) *Replica Set Coordination*: Other data is only relevant for members of a replica set, e.g., replica control data. Such data could be access control lists for a replicated application or membership data for the replica set. In some cases, where a replica set comprises geographically close nodes, limiting the coordination for this data to a specific set of fog nodes can even reduce the network delay for this coordination group. Note that the membership of a replica set can change, e.g., when a data replica is migrated to a different location.

3) *Node Coordination*: A fog node comprises one or more physical machines running in the same location, e.g., a micro-datacenter on the edge. Within such a fog node, there will often be coordination needs, e.g., when individual machines run different application code. In this case, the node machines must coordinate in order to appear as one node to other nodes.

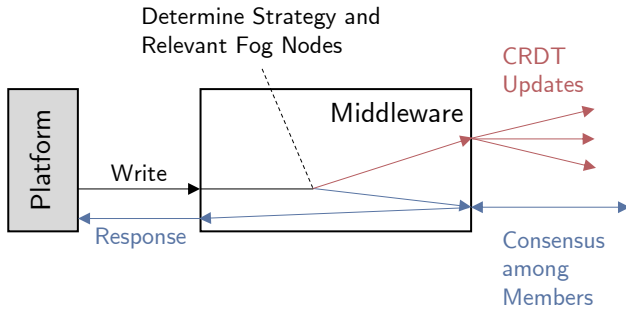


Fig. 3. The coordination middleware takes write requests from the application and determines the relevant fog nodes based on the configured level for a data type. Depending on the configured consistency, a CRDT update is sent immediately (red) or consensus is reached with other members (blue).

C. Architecture

We propose to implement our approach as a coordination middleware that can be used by different fog applications or platforms. Initially, the application's data types must be configured, i.e., coordination strategies and levels are defined for each coordination data set. Figure 3 illustrates the update process: When a write is performed, the middleware first determines the relevant fog nodes for that update depending on membership. We assume that replica set membership is also stored in this coordination middleware or can be inferred from system configuration data. For data types that are configured with eventual consistency, the update is propagated to those nodes as a CRDT update in an asynchronous manner. Where strict consistency is required, a consensus is reached with the relevant nodes before the update is persisted. The middleware will serve read requests from a local cache, where, e.g., CRDT updates from other nodes are applied. If required, a read quorum for strictly consistent data items may also be established. Alternatively, only a subset of the set of member nodes may be chosen, e.g., only the cloud nodes for system coordination, or following a leader-follower approach for node coordination.

IV. CONCLUSION & FUTURE WORK

In this paper, we have proposed a new architecture for distributed coordination in fog platforms. We have motivated why different types of data require different tradeoffs between consistency and latency. Further, we have shown different levels of coordination, from system to node configuration and state management, which limits message passing and consensus participants for geo-distributed fog platforms. As a middleware, an implementation of this approach can be used by different fog and edge platforms without changes to their architecture.

It also opens a number of interesting future research directions: We plan to evaluate the impact of consistency tuning on both correctness and performance of different fog platforms to find out which kinds of data can benefit from relaxed

consistency guarantees. Of course, the middleware may be extended for other coordination strategies. Furthermore, fog overlay networks may be used for deduplication of CRDT update messages to limit the network cost of an update broadcast. Finally, we also plan to explore inferring coordination strategies and levels from data access to allow an integration of the coordination middleware without explicit configuration of data types.

REFERENCES

- [1] B. Cheng, J. Fuerst, G. Solmaz, and T. Sanada, "Fog function: Serverless fog computing for data intensive IoT services," in *Proceedings of the 2019 IEEE International Conference on Services Computing (SCC 2019)*, Jul. 2019, pp. 28–35.
- [2] L. Baresi and D. Filgueira Mendonça, "Towards a serverless platform for edge computing," in *Proceedings of the 3rd IEEE International Conference on Fog Computing (ICFC 2019)*, Jun. 2019, pp. 1–10.
- [3] L. Baresi, D. F. Mendonça, and G. Quattrocchi, "PAPS: A framework for decentralized self-management at the edge," in *Proceedings of the 17th International Conference on Service-Oriented Computing (ICSOC 2019)*, Oct. 2019, pp. 508–522.
- [4] A. Bocci, S. Forti, G.-L. Ferrari, and A. Brogi, "Secure FaaS orchestration in the fog: how far are we?" *Computing*, vol. 103, no. 5, pp. 1025–1056, 2021.
- [5] S. H. Mortazavi, B. Balasubramanian, E. de Lara, and S. P. Narayanan, "Pathstore, a data storage layer for the edge," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobySys 2018)*, Jun. 2018, p. 519.
- [6] J. Hasenburger, M. Grambow, and D. Bernbach, "Towards a replication service for data-intensive fog applications," in *Proceedings of the 35th ACM Symposium on Applied Computing, Posters Track (SAC 2020)*, Mar. 2020, pp. 267–270.
- [7] —, "FBase: A Replication Service for Data-Intensive Fog Applications," Tech. Rep., 2019.
- [8] S. Gilbert and N. Lynch, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services," *ACM SIGACT News*, vol. 33, no. 2, pp. 51–59, 2002.
- [9] D. Abadi, "Consistency tradeoffs in modern distributed database system design: CAP is only part of the story," *Computer*, vol. 45, no. 2, pp. 37–42, 2012.
- [10] D. Bernbach, F. Pallas, D. G. Pérez, P. Plebani, M. Anderson, R. Kat, and S. Tai, "A research perspective on fog computing," in *Proceedings of the 15th International Conference on Service-Oriented Computing (ICSOC 2017) Workshops*, Aug. 2018, pp. 198–210.
- [11] F. Pedone, M. Wiesmann, A. Schiper, B. Kemme, and G. Alonso, "Understanding replication in databases and distributed systems," in *Proceedings of the 20th IEEE International Conference on Distributed Computing Systems (ICDCS 2000)*, Apr. 2000, pp. 464–474.
- [12] M. Burrows, "The chubby lock service for loosely-coupled distributed systems," in *Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI 2006)*, Sep. 2006, pp. 335–350.
- [13] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," *Operating Systems Review*, vol. 37, no. 5, pp. 29–43, 2003.
- [14] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems*, vol. 26, no. 2, pp. 4:1–4:26, 2008.
- [15] C. Wöbker, A. Seitz, H. Mueller, and B. Bruegge, "Fogernetes: Deployment and management of fog computing applications," in *Proceedings of the 2018 IEEE/IFIP Network Operations and Management Symposium (NOMS 2018)*, Apr. 2018, pp. 1–7.
- [16] J. Eberhardt, D. Ernst, and D. Bernbach, "Smac: State management for geo-distributed containers," in *Proceedings of the 2nd IEEE International Workshop on Container Technologies and Container Clouds (WoC 2016)*, Apr. 2016.
- [17] A. Jeffery, H. Howard, and R. Mortier, "Rearchitecting kubernetes for the edge," in *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking (EdgeSys 2021)*, Apr. 2021, pp. 7–12.