Computer Networks 76 (2015) 191–206

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

On the tradeoff of availability and consistency for quorum systems in data center networks

Xu Wang, Hailong Sun *, Ting Deng, Jinpeng Huai

School of Computer Science and Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form 15 September 2014
Accepted 10 November 2014
Available online 15 November 2014

Keywords:
Quorum systems
Availability
Data center networks

ABSTRACT

Large-scale distributed storage systems often replicate data across servers and even geographically-distributed data centers for high availability, while existing theories like CAP and PACELC show that there is a tradeoff between availability and consistency. Thus eventual consistency is proposed to provide highly available storage systems. However, current practice is mainly experience-based and lacks quantitative analysis for identifying a good tradeoff between the two. In this work, we are concerned with providing a quantitative analysis on availability for widely-used quorum systems in data center networks. First, a probabilistic model is proposed to quantify availability for typical data center networks: 2-tier basic tree, 3-tier basic tree, fat tree and folded clos, and even geo-distributed data center networks. Second, we analyze replica placements on network topologies to obtain maximal availability. Third, we build the availability-consistency table and propose a set of rules to quantitatively make tradeoff between availability and consistency. Finally, with Monte Carlo based simulations, we validate our presented quantitative results and show that our approach to make tradeoff between availability and consistency is effective.

1. Introduction

Data centers as a mainstay for large-scale distributed storage systems, often face the challenges of high availability and scalability [1,2]. They typically replicate data across multiple servers and even geographically-distributed data centers [3] to tolerate network partition and node crashes. However, the CAP [4] and PACELC [5] theorems show that there is a tradeoff between availability and consistency in data replication. High availability is regarded as an important property in service level agreements for cloud services. For example, Amazon EC2 claims the availability of 99.95% [6]; Google Cloud Storage's service level commitment provides 99.9% availability [7]; and the availability of Microsoft Windows Azure is promised as 99.9% [8]. In order to provide extremely high availability, systems such as Amazon Dynamo [9] and Apache Cassandra [10] often eschew strong consistent replicas and instead provide eventual consistency [11,12]. But eventual consistency may lead to inconsistency such as read staleness [13] and write conflicts [9], since the newest value of a data item is eventually returned and the same data item may be updated on different, potentially disconnected replicas. Therefore it is necessary to understand how much availability can be obtained in different consistency levels such that we can quantitatively make tradeoff between them.

Distributed quorums [14] are widely used for eventually consistent storage systems in data centers [9,10,15,16]. With quorum replication, these storage systems write a data item by sending it to a set of replicas (write quorums) and read from a possibly different set of replicas (read quorums). Strong consistency is provided

* Corresponding author. Tel.: +86 13811909091.
E-mail addresses: wangxu@act.buaa.edu.cn (X. Wang), sunhl@act.buaa.edu.cn (H. Sun), dengting@act.buaa.edu.cn (T. Deng), huaijp@buaa.edu.cn (J. Huai).

http://dx.doi.org/10.1016/j.comnet.2014.11.006
1389-1286/

by guaranteeing the intersection of write quorums and read quorums; otherwise, read requests may return stale values, thus resulting in inconsistency. In practical quorum systems, any W replicas are defined as write quorums and any R replicas as read ones. With N replicas, if $W + R > N$, strong consistency is achieved since the overlapping of write and read quorums is always guaranteed; and if $W + R \leqslant N$, the overlapping condition may not be satisfied, therefore it can cause inconsistency. The former is a strict quorum system [17] and the latter is a probabilistic quorum system [18]. Almost all practical quorum systems like Dynamo and Cassandra provide the configuration capability on W and R. It is obvious that greater W and R will result in stronger consistency, but will lower system availability since greater W and R require more live replicas to be accessed for available write and read requests during failures. How to configure (W, R) to get the best solution for both consistency and availability? The current practice is mainly experience-based and lack quantitative analysis for identifying a good tradeoff between them.

The quantitative consistency analysis for quorum systems has been well studied [13,18]. However, availability is network topology-aware because network partition is the main cause of system unavailability according to the CAP theorem [4], and data center networks are complex [19–21] such as basic tree, fat tree, folded clos network and even geo-distributed data centers, therefore the quantitative analysis on availability of quorum systems in data center networks is challenging. There are several bodies of work [18,22–26] on quantifying the availability of quorum systems, but they have two limitations: (1) they usually assume simple network topologies such as the fully connected network and ring topology; and (2) they define the availability of quorum systems as the probability that at least one quorum is alive, but in practice live quorums may be inaccessible due to failures of switches in networks.

In this work, we focus on evaluating the availability of quorum systems in four typical data center networks: 2-tier basic tree, 3-tier basic tree, fat tree and folded clos network, and even geo-distributed data centers [19–21]. At first, we propose a system model $QS(DCN, PM, W/R)$ for quorum systems, where DCN represents the data center network topologies containing core switches, aggregation switches, ToR (Top of Rack) switches, servers and their links, PM is a vector describing the placement of replicas on servers, and W/R are the sizes of write/read quorums. Since write requests should reaches at least W live replicas and read requests must wait for responses from at least R replicas, they are equivalent to each other for our availability analysis and we only consider write requests. Based on $QS(DCN, PM, W)$, the write availability is defined as the probability that a write can reach one live replica (i.e. the coordinator) from live core switches and at least other $W - 1$ live replicas from the coordinator. Although typical data center networks are complex especially for the fat tree and folded clos network, they are essentially tree-like. Therefore, after calculating the write availability for a simple 2-tier basic tree, we use super nodes to represent subtrees or groups of nodes for 3-tier basic tree, fat tree and folded clos network which are logically equivalent to each other and obtain simplified network topologies. Then the write availability for the 3-tier basic tree, fat tree and folded clos network is computed by conditional probability and reusing the result of 2-tier basic tree. In addition, we extend our quantitative write availability from one single data center to geo-distributed data centers. We also analyze the impact of replica placement on maximizing write availability. Based on the quantitative results of availability, we build an availability-consistency table filled with values of $\langle Availability, Consistency\rangle$ for quorum systems, and propose a set of rules to choose the best (W, R) to make tradeoff between availability and consistency. Finally, through Monte Carlo based event driven simulations, we validate our quantitative results and show the effectiveness of our method for balancing availability and consistency.

We make the following contributions:

- We propose a system model $QS(DCN, PM, W/R)$ for quorum systems in typical data center networks. Unlike earlier work, our system model considers practical complex data center network DCN where switches may fail, and the placement of replicas on servers PM;
- On the basis of $QS(DCN, PM, W)$, we build a probabilistic model for the write availability of quorum systems in four typical data center networks: 2-tier basic tree, 3-tier basic tree, fat tree and folded clos network. We also extend the write availability from one single data center to geo-distributed data centers;
- We analyze how to place replicas on network topologies to maximize write availability and discuss special cases when the network topology is a 2-tier basic tree and the number of replicas is a popular value 3;
- We build an availability-consistency table filled with $\langle Availability, Consistency\rangle$, and then propose a set of rules to choose the best (W, R) configuration for a specific quorum system to balance availability and consistency;
- With Monte Carlo based event driven simulations, we validate our quantitative results and show the effectiveness of our proposed approach to quantitatively make tradeoffs between availability and consistency.

The remainder of this paper is structured as follows. Section 2 introduces the background. Section 3 presents our system model. Section 4 describes how to quantify the write availability of quorum systems. Section 5 propose the impact of replica placements on availability. How to make tradeoffs between availability and consistency is shown in Section 6. Section 7 provides our experimental results. Related work and discussion are presented in Section 8 and Section 9, respectively. Finally, Section 10 concludes the work.

2. Background

In this section, we present the background, including the preliminary knowledge of quorum systems and data center networks.