ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

DAC: Improving storage availability with Deduplication-Assisted Cloud-of-Clouds

Suzhen Wu [a], Kuan-Ching Li [c], Bo Mao [b,*], Minghong Liao [b]

[a] Computer Science Department, Xiamen University, Xiamen, 361005, China

[b] Software School, Xiamen University, Xiamen, 361005, China

[c] Department of Computer Science and Information Engineering, Providence University, Taiwan

HIGHLIGHTS

- From the investigations and the preliminary performance evaluations, we show the diversity characteristics of cloud storage systems from the cost and performance aspects.
- A deduplication-assisted cloud storage system is proposed to improve the storage efficiency and network bandwidth.
- Exploiting the data reference characteristics to place data blocks among multiple cloud storage providers.
- The availability of cloud storage system is improved by incorporating both the replication and erasure code schemes.

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form
22 January 2016
Accepted 6 February 2016
Available online xxxx

Keywords:
Cloud-of-Clouds
Data deduplication
Availability
Data distribution
Cloud storage

ABSTRACT

With the increasing popularity and rapid development of the cloud storage technology, more and more users are beginning to upload their data to the cloud storage platform. However, solely depending on a particular cloud storage provider has a number of potentially serious problems, such as vendor lock-in, availability and security. To address these problems, we propose a Deduplication-Assisted primary storage system in Cloud-of-Clouds (short for DAC) in this paper. DAC eliminates the redundant data blocks in the cloud computing environment and distributes the data among multiple independent cloud storage providers by exploiting the data reference characteristics. In DAC, the data blocks are stored in multiple cloud storage providers by combing the replication and erasure code schemes. To better utilize the advantages of both replication and erasure code schemes and exploit the reference characteristics in data deduplication, the high referenced data blocks are stored with the replication scheme while the other data blocks are stored with the erasure code scheme. The experiments conducted on our lightweight prototype implementation show that DAC improves the performance and cost efficiency significantly, compared with the existing schemes.

1. Introduction

With the increasing popularity and cost-effectiveness of cloud storage systems, many companies and organizations have migrated or plan to migrate data from their private data centers to the cloud. However, solely depending on a particular cloud storage provider has a number of potentially serious problems. First, it can cause the so-called vendor lock-in problem for the customers [1,2], which results in prohibitively high cost for clients to switch from one provider to another. Second, it can cause service disruptions, which in turn will lead to SLA violation, due to cloud outages, resulting in penalties, monetary or other forms, for the service providers. Examples include a series of high-profile cloud outages in the year of 2013 for cloud providers, such as Amazon, Microsoft and Google [3], from a 5-min failure that costed half a million dollars to a week-long disruption that costed an immeasurable amount of brand damage. From January to March 2014, DropBox has experienced twice service outages [3]. More seriously, Nirvanix filed for Chapter 11 bankruptcy protection on October 1, 2013 [4]. The company gave customers two weeks' notice to retrieve their data. Some users had petabytes of data with single copy stored in Nirvanix. Third, solely depending on a particular cloud storage provider can also result in possible increased service costs and data security issues, such as the data

* Corresponding author. Tel.: +86 5922580033.
E-mail address: maobo@xmu.edu.cn (B. Mao).

http://dx.doi.org/10.1016/j.future.2016.02.001
0167-739X/

Please cite this article in press as: S. Wu, et al., DAC: Improving storage availability with Deduplication-Assisted Cloud-of-Clouds, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.02.001

leakage problem [5]. Therefore, using multiple independent cloud providers, called Cloud-of-Clouds, is an effective way to provide better availability for the cloud storage systems.

In a Cloud-of-Clouds storage system, the data redundancy is introduced to judiciously distribute the data among the clouds. Thus, the redundant data distribution scheme is critically important for the storage availability, performance, cost and space efficiency. Several systems have been proposed for Cloud-of-Clouds. RACS [1] uses the erasure coding to mitigate the vendor lock-in problem encountered by a user when switching the cloud vendors. It transparently stripes the data across multiple cloud storage providers with RAID-like techniques. HAIL [6] provides integrity and availability guarantees for the stored data. It allows a set of servers prove to a client that a stored file is intact and retrievable by the approaches adopted from the cryptographic and distributed-systems communities. NCCloud [7] achieves the cost-effective repair for a permanent single-cloud provider failure to improve the availability of cloud storage services. It is built based on network-coding-based storage schemes called regenerating codes with an emphasis on the storage repair, excluding the failed cloud in repair.

The above three systems are all based on the erasure code or the network code. In contrast, DuraCloud [8] utilizes replication to copy the user content to several different cloud storage providers to provide better availability. Moreover, it ensures that all copies of the user content remain synchronized. However, users should pay more money for the additional storage space and bandwidth required by DuraCloud. DEPSKY [2] improves the availability and confidentiality of the commercial storage cloud services by building a Cloud-of-Clouds on top of a set of storage clouds, combining the Byzantine quorum system protocol, cryptographic secret sharing, replication and the diversity of different cloud providers. Different from these approaches, HyRD [9] integrates both replication and erasure code to the Cloud-of-Clouds. It takes the workload characteristics and the diversity of cloud storage providers, specially the file sizes, into the design of the redundant data distribution strategy so that the advantages of both the replication and erasure code can be exploited while their disadvantages can be hidden. As a result, both the performance and storage efficiency are improved with the availability guarantee. However, the redundant data blocks over the network are not eliminated.

On the other hand, previous studies on the workload characteristics have shown that the data redundancy is moderate to high in the cloud storage environments [10–12]. These studies have shown that by applying the data deduplication technology to large-scale data sets, an average space saving of 30%, with up to 90% in VM and 70% in HPC storage systems, can be achieved. The recent studies, such as RACS [1], DuraCloud [8], DepSky [2], NCCloud [7] and HyRD [9], indicate that the replication-based schemes are performance-friendly to the hot data blocks while the erasure-code-based schemes are cost-efficient to the cold data blocks [9,13]. It suggests that a sensible data distribution scheme in the Cloud-of-Clouds should dynamically utilize the replication and erasure codes based on the hotness characteristics of data blocks. To address the important storage availability issue in the Cloud-of-Clouds, we propose a deduplication-assisted data reduction and data distribution approach, called DAC, by exploiting the data redundancy characteristics of applications. DAC utilizes the replication scheme to store the data blocks with high reference count, and utilizes the erasure codes to store the other data blocks on multiple cloud storage providers. By exploiting the data redundancy characteristics and the diversity of cloud providers, both the advantages of erasure codes and replication are exploited and their disadvantages are alleviated. The extensive trace-driven experiments conducted on our lightweight prototype implementation of DAC show that DAC significantly outperforms RACS, DuraCloud and HyRD in the I/O performance measure of average response times. Moreover, our evaluation and analysis results also show that DAC achieves significant cost and space efficiency.

The rest of this paper is organized as follows. The background and motivation are presented in Section 2. We describe the DAC architecture and design in Section 3. The performance evaluation is presented in Section 4. We present the related work in Section 5 and conclude this paper in Section 6.

2. Background and motivation

In this section, we present some important observations drawn from previous and our analysis of the vendor lock-in problem of cloud storage, the diversity characteristics of cloud storage providers, and the data redundancy in primary storage systems to motivate the DAC study.

2.1. The vendor lock-in problem

The services provided by the cloud storage are diverse [1,18]. The cloud storage providers offer different pricing and different performance characteristics, including extra features such as geographic data distribution, access through mountable file systems and specific APIs. Changes in these features, or the emergence of new providers with more powerful and attractive characteristics, might compel some users to switch from one provider to another. However, moving from one provider to another one may be very expensive because the switching cost is proportional to the amount of data that has been stored in the original provider [1]. The more data has been stored in the original provider, the higher switching cost will be paid to the data migration. It puts the users at a disadvantage, that is, when the cloud storage provider that has stored the user's data raises the prices or negotiates a new contract less favorable to the user, the user has no choice but to accept because of the high switching cost, which is called *vendor lock-in problem* [1,2].

Besides the possible increased prices or pressed unfavorable new contract, the vendor lock-in can also lead to possible data loss or unavailability for users if their cloud storage provider goes out of business or suffers a service outage. Despite of the strict Service-Level Agreements (SLAs) between the cloud provider and the user, the service failures and outage occur and are almost unavoidable [9,19]. The cloud outages in 2013, although infrequent, showed that the service unavailability may last up to several hours and even several days [3]. A study conducted by the ESG (Enterprise Strategy Group) research shown that about 58% of professionals in SMBs (Small and Medium Businesses) can tolerate no more than four hours of downtime before experiencing significant adverse effect [20,21]. More seriously, EMC's Disaster Recovery Survey in 2013 [22] observed that the average cost per hour of downtime is much higher than ever before and 54% of users suffered from lost data or service downtime, which further stresses the importance of the service/data availability in cloud storage systems.

To address the vendor lock-in problem induced by a single individual cloud provider, a Cloud-of-Clouds solution is proposed in the literatures [1,2,7,8]. It redundantly distributes the data across multiple providers by means of the data redundancy schemes, such as replication and erasure codes. As a result, users can maintain their mobility while insuring against the outages of a single individual cloud provider.

Please cite this article in press as: S. Wu, et al., DAC: Improving storage availability with Deduplication-Assisted Cloud-of-Clouds, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.02.001