Computer Standards & Interfaces 44 (2016) 18–27

Contents lists available at ScienceDirect

Computer Standards & Interfaces

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

A file-deduplicated private cloud storage service with CDMI standard

Xiao-Long Liu [a], Ruey-Kai Sheu [b,*], Shyan-Ming Yuan [a], Yu-Ning Wang [a]

[a] Dept. of Computer Science, National Chiao Tung University, Taiwan
[b] Dept. of Computer Science, Tunghai University, Taiwan

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form 19 August 2015
Accepted 16 September 2015
Available online 25 September 2015

Keywords:
Cloud storage
Private cloud
Data deduplication
CDMI
DFS

ABSTRACT

The emergence of cloud environments makes users convenient to synchronize files across platform and devices. However, the data security and privacy are still critical issues in public cloud environments. In this paper, a private cloud storage service with the potential for security and performance concerns is proposed. A data deduplication scheme is designed in the proposed private cloud storage system to reduce cost and increase the storage efficiency. Moreover, the Cloud Data Management Interface (CDMI) standard is implemented in the proposed system to increase the interoperability. The proposed service provides an easy way to for user to establish the system and access data across devices conveniently. The experiment results also show the superiority of the proposed interoperable private cloud storage service in terms of data transmission and storage efficiency. By comparing with the existing system Gluster Swift, the proposed system is demonstrated much suitable for the service environment where most of the transmitted data are small files.

1. Introduction

Cloud storage is a service model where data is maintained, managed and backed up for users over network. Different from traditional local device storage, in general, cloud storage is an on-demand self-service which can be easily accessed via standard Internet APIs and communications protocols [1]. Since cloud storage service provides users with abundant storage space and gives users convenience in synchronizing files across platform and devices, it is now adding its appeal for many citizens. The different forms of cloud storage design are public cloud storage and private cloud storage. Public cloud storage, such as Dropbox [2] and Google Drive [3], are provided by service providers and the storage infrastructure is hosted by the cloud vendor at the vendor's premises. The charge of the storage is depended on how many resources the user used [4] and it can scale storage space up or down elastically in accordance with the request. The data of different customers are likely to be stored at the service provider site and mixed together on the cloud storage systems. The public cloud storage makes immediate acquiring of data more convenient for the user. However, the data security and privacy are still critical issues in public cloud environments, especially for the exclusive clients who care about the privacy and ownership of files. Chen and Zhao point out that data security and privacy issues are the top concerns of consumers, especially for large enterprise [5]. It is difficult to set powerful security boundaries and protect data privacy in cloud storage. After surveying with 402 participants, Ion et al. [6] also indicate that most of the cloud storage users are worried about data privacy issue. Hence, the private cloud storage appeared on the scene.

Different from public cloud storage, private cloud storage [7] is built for the exclusive use of one person or organization with low bandwidth environment. It is suitable for the users who want to have customization service or who don't trust the public cloud storage vendors. The users of private cloud storage must establish the infrastructure personally and they just need to download the software from private cloud service provider and install it on their own hardware. Consequentially, private cloud storage is equipped with the attributes of public cloud storage and protects the security of private files meanwhile. However, there are still some aspects of private cloud storage service that need concern. Firstly, since private cloud storage users have the responsibility to establish the infrastructure, the space of storage system has to be used in an efficient way. According to the research of Symantec [8], there are 37% duplicated data in Taiwan enterprises and 42% duplicated data in global enterprises. The duplicate copies of data would result in redundant consumptions of storage space and network bandwidth. Moreover, most of the files transmitted in enterprise are images or documents with small size. Therefore, how to efficiently reduce the redundant cost of storage spaces, especially dealing with small files, has already been a complex and challenging issue for enterprises [9]. Secondly, elasticity is also an essential factor for a private cloud storage system. An elastic system means that the potential system must have the ability to dynamically add or reduce resource according to the requirements.

* Corresponding author.
E-mail addresses: shallen548@gmail.com (X.-L. Liu), rickysheu@thu.edu.tw (R.-K. Sheu), smyuan@cs.nctu.edu.tw (S.-M. Yuan), wang790222@gmail.com (Y.-N. Wang).

http://dx.doi.org/10.1016/j.csi.2015.09.010
0920-5489/

The private cloud storage owner would easily manage storage and get resource benefit if the underlying infrastructure provided primitives for elasticity. Thirdly, the interoperability of a storage system with international standard interface should be treated as another important characteristic. The unique interface released by a single vendor would remain under the change control of the vendor. It may be essentially locking customers into that service. Compared to the storage system with unique interface, the storage system with a standard interface which accommodates requirement from multiple vendors and can be extended for proprietary functions would be more convenient and less restrictive for users.

To solve the issues mentioned above, this research constructs a data deduplication private cloud storage system with Cloud Data Management Interface (CDMI) [10] standard based on the fundamental distributed file system (DFS). A data deduplication scheme which can detect and remove the redundant data in the storage is designed in the proposed system to reduce cost and increase the storage efficiency. The deduplication algorithm is also designed to be suitable for the service environment where most of the transmitted data are small files. Moreover, the international standard interface CDMI is implemented in the proposed system to increase the interoperability. Finally, Gluster [11] is chosen as the fundamental DFS in our proposed private cloud storage system. Integrating with distributed file system as back end storage makes the proposed system have the elasticity, good performance at backend storage, and be managed conveniently [12]. By integrating these characteristics, the proposed private cloud storage service can provide users an easy way to establish the system and access data across devices conveniently. The experiment results also demonstrate the superiority of the proposed system in terms of data transmission and storage efficiency.

The remainder of this paper is as follows. Section 2 provides background and related works of the proposed system; Section 3 elaborates on the proposed system's architecture, components and the details of working; Section 4 describes the experimental results; and, finally, we give concluding remarks and future works in Section 5.

2. Background

2.1. Data deduplication

Data deduplication is a technique to optimize the storage space. This technique keeps only one copy of all identical part in the storage system without saving the redundant part. Since the amount of data is growing with astonishing speed, the enterprise would not only save cost on physical devices by data deduplication but also manage data efficiently. From the side of where data is processed, deduplication can be divided into target deduplication and source deduplication [13]. Target deduplication happens after front-end user, which is client, transmits files to back-end storage server and redundant data is eliminated at the back-end storage side, which is server. The storage device has the responsibility for processing the redundant data without influencing client's operating. It doesn't consume client's resource and they don't know the deduplication is occurring. On the other hand, source deduplication does data deduplication before it is transferred. There are some calculations on client side, like calculating hash value. Source deduplication consumes client's resource but it has the advantage of saving the network traffic bandwidth [14].

Deduplication can be further classified into two approaches by the unit of comparison. One is chunk level deduplication and the other is file level deduplication. Data Domain Deduplication file System (DDDFS) [15] is one of the file system which performs chunk level deduplication. It divides the file into variable sized chunks and uses Secure Hash Algorithm (SHA-1) [16] to find the hash value of each chunk. Each chunk is checked with a set of chunk indices maintained in chunk store for duplicate detection. Hence, it eliminates duplicate chunks of data even if the corresponding files are not identical. With performing file-level deduplication, Xu et al. [9] designed a file deduplication framework on Hadoop system, where Secure Hash Algorithm 2 (SHA-2) [16] was utilized to conduct the data mapping through the whole file. Therefore, it eliminates duplicate copies of the same file instead of chunks. Lokeshwari et al. [17] further proposed an optimized cloud storage with high throughput deduplication approach, where chunk level and file level deduplication methods were both implemented. They discussed the approach of deduplication from two dimensions, i.e. efficiency and throughput, in private cloud storage environment. Chunk level deduplication has the advantage of better efficiency, which means the degree of how many space is saved. However, chunk level deduplication has lower throughput than file level deduplication. The throughput means the overhead generated by the process of deduplication. Besides, in the study of practical deduplication, Meyer [18] found that file level deduplication should be a highly efficient way to lower storage consumption in the situation with sparseness.

In public cloud environment, the stored data are usually encrypted, since security and privacy are the main concerns from users' perspective for data outsourcing. However, traditional deduplication approaches are incompatible with the encrypted data in the public cloud. To secure the data in public cloud while realizing deduplication, Douceur et al. presented a convergent encryption approach [19], where identical data copies will generate the same convergent key and the same cipher text. Therefore, it allows the deduplication procedure performed on the cipher texts. To further achieve efficient and reliable secure deduplication, Li et al. [20] proposed a new construction called Dekey for convergent key management on both user and public cloud storage sides. They utilized the secret sharing techniques and apply deduplication to the convergent keys, which significantly limited the storage overhead in realistic environments. Later. Li et al. also presented a secure authorized deduplication approach based on [21]. In their approach, the private cloud is regarded as a proxy to allow users to securely perform duplicate check with differential privileges. The data operation is managed in private cloud while the users only outsource their data storage by utilizing public cloud.

The studies of secure deduplication are mainly presented on account of the data outsourcing in public cloud environment. If the users' data was stored in the private cloud system and managed by themselves, the security and privacy will not be the main concerns from users' perspective. Instead the designed storage system should give the priority to the convenience and reliability. The goal of the designed proposed system can be concluded as providing a convenient and efficient private cloud storage system. Therefore, with the comparison and analysis of different deduplication strategies mentioned before, source deduplication with file level deduplication strategy will be adopted in our proposed system.

2.2. Cloud data management interface

The Cloud Data Management Interface (CDMI) [10] defines the functional interface that can be used by applications to create, retrieve, update and delete data elements from the Cloud. A CDMI client has the ability to find the capabilities of the cloud storage offering and this interface allows a CDMI client manages containers and the data. Further, this interface defines the rule about how to set metadata on containers and the data which is contained in them. The administrator also can use the interface to manage containers, accounts, security access and monitoring/billing information [22]. CDMI is used in a cloud by defining Representational State Transfer (REST) HTTP operations. The basic flow of CDMI is illustrated in Fig. 1. CDMI client communicates with CDMI server upon RESTful protocol and the methods include PUT, GET, DELETE, and so on. After CDMI server receives the request, it sends response according the status of work.

REST is an efficient software architectural style which was introduced and defined in 2000 by Fielding and Taylor [23]. There are some