A Study on Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage

Dhirajsing Thakur[1], Prakhar Yadav[2], Shital Bhosale [3] and Prof. Pr. Patil[4]

PICT, University of Pune

ABSTRACT

Data is the heart of any organization; hence it is necessary to protect it. For doing so, it is the needed to implement a good backup and recovery plan. But the redundant nature of the backup data makes the storage a concern; hence it is necessary to avoid the redundant data present in the backup. Data de-duplication is one such solution that discovers and removes the redundancies among the data blocks. In this paper, a deduplication scheme is proposed that improves data deduplication efficiency by exploiting application awareness. The application proposed is motivated on personal storage.

Keywords:- Application awareness, Cloud backup service, Chunking schemes, Data redundancy, Data deduplication, Deduplication efficiency.

1. INTRODUCTION

According to the present scenario, the backup has become the most essential mechanism for any organization. Backing up files can protect against accidental loss of user data, database corruptions, hardware failures, and even natural disasters. However, the large amount of redundancies which is found in the backups makes the storage of the backups a concern, thus utilizing a large of disk space. Data de-duplication comes as a rescue for the problem of redundancies in the backup. It is a capacity optimization technology that is being used to dramatically improve the storage efficiency. Data de-duplication eliminates the redundant data and stores only unique copy of the data. Here instead of saving the duplicate copy of the data, data de-duplication helps in storing a pointer to the unique copy of the data, thus reducing the storage costs involved in the backups to a large extent. It need not be applied in only backups but also in primary storage, cloud storage or data in flight for replication, such as LAN and WAN transfers. It can help organizations to manage the data growth, increase efficiency of storage and backup, reduce overall cost of storage, reduce network bandwidth and reduce the operational costs and administrative costs. The five basic steps involved in all of the data de-duplication systems are evaluating the data, identify redundancy, create or update reference information, store and/or transmit unique data once and read or reproduce the data. Data de-duplication technology divides the data into smaller chunks and uses an algorithm to assign a unique hash value to each data chunk called fingerprint. The algorithm takes the chunk data as input and produces a cryptographic hash value as the output. The most frequently used hash algorithms are SHA, MD5. These fingerprints are then stored in an index called chunk index. The data de-duplication system compares every fingerprint with all the fingerprints already stored in the chunk index. If the fingerprint exists in the system, then the duplicate chunk is replaced with a pointer to that chunk. Else the unique chunk is stored in the disk and the new fingerprint is stored in the chunk index for further process.

2. EXISTING SYSTEM

In paper [1] the cloud computing is a technology which is used to provide resources as a service. There are many services provided by cloud provider. Such as SAAS, IAAS, PAAS. The cloud computing provides the Storage-as-a-Service which is used to backup the users data into cloud. The Storage-as-a-Service is provided by Storage Service Provider or Cloud Service Provider. This service is provided by Cloud Service Provider which is effective, reliable and cost-effective. The existing backup scheduling provides the reliability by maintaining the same copy of the data twice. The existing backup scheduling provides the reliability and backup speed, but the redundancy of data is not considered. The existing backup scheduling not considers much of the security issues. The limitations of the existing backup scheduling algorithm is improved by proposing a backup scheduling algorithm(IBSD) which aims at reducing redundancy without compromising on availability. The IBSD algorithm reduces redundancy by deduplication techniques. The deduplication is a technique which is used to identify the duplicate data. The de-duplication identifies the duplicate data and eliminates it, by storing only one copy of the original data. If the duplicate occurs then the link will be added to the existing data. Also paper [2] tells us that, Data Deduplication describes approach that reduces the storage capacity needed to store data or the data has to be transfer on the network. Source Deduplication is useful in cloud backup that saves network bandwidth and reduces network space Deduplication is the process by breaking up an incoming stream into relatively large segments and deduplicating each segment against only a few of the most similar previous segments. To identify similar segments use block index technique The problem is that these schemes traditionally require a full chunk index, which indexes every chunk, in order to determine which chunks have already been stored unfortunately, it is impractical to keep such an index in RAM and a disk based index with one seek per incoming chunk is far too slow. In this paper we describes application based deduplication approach and indexing scheme contains block that preserved caching which maintains the locality of the fingerprint of duplicate content to achieve high hit ratio and to overcome the lookup performance and reduced cost for cloud backup services and increase dedulpication efficiency. In paper [3], to improve space utilization and reduce network congestion, cloud backup venders (CBVs) always implement data deduplication in the source and the destination. Towards integrating source and destination data deduplication, we mainly propose two proposals in this area. One of the important things of this is benefit-cost model for users to decide in which degree the deduplication executes in client and in cloud, and let the data centre to decide how to handle the duplications. This will give better reliability, quality of service etc. Combining caching and prefetching, and the requirements of different cloud backup services, the read performance in the cloud backup systems can be improved.

3.STUDY OF SYSTEM ARCHITECTURE AND ALGORITHM

Following diagrams shows the system architecture and the activity diagram from which the flow and the working of the system can be explained.

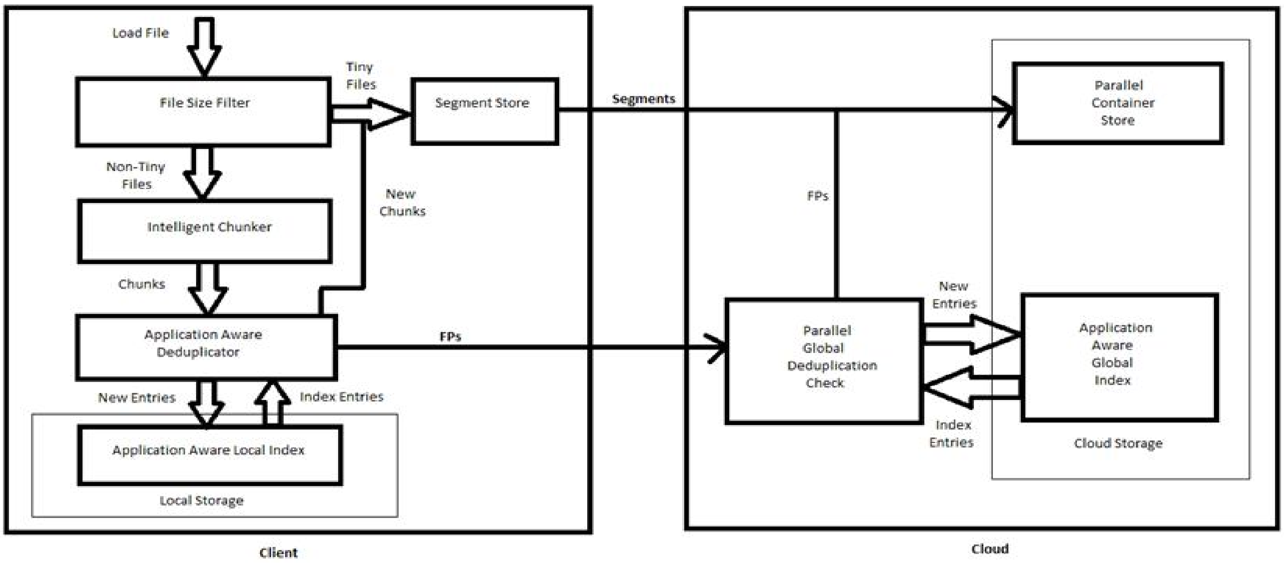

Fig. 3.1 System Architecture

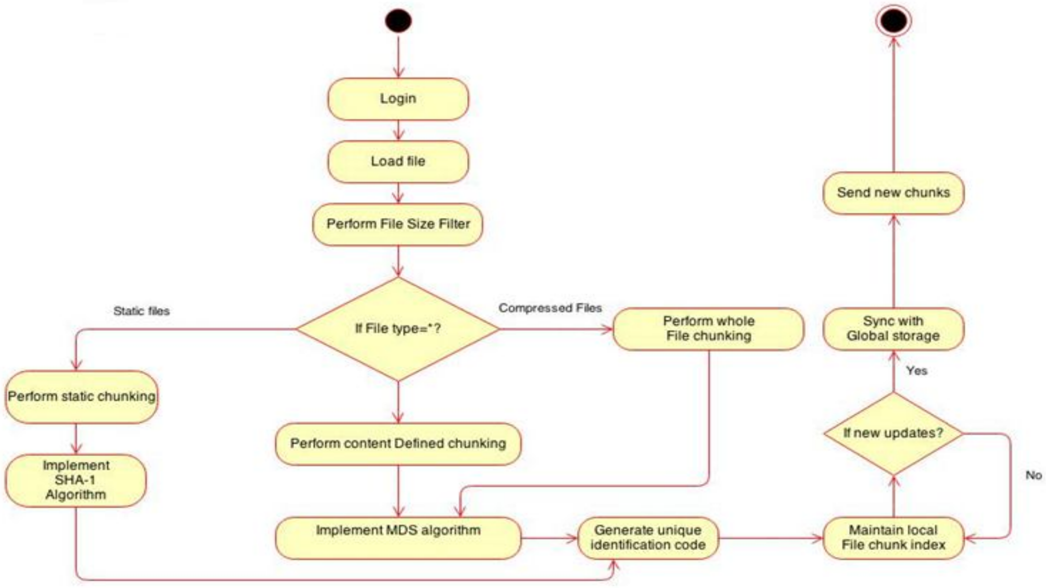

Fig 3.2 Activity Diagram

Studied Algorithm: In a system which we are going to implement after studying this all, we follows divide n conquer strategy. In this strategy the main problem is Sub-divided into number of Sub-problems(say n).Now the solution for each Sub-problem is devised and all such n solutions are integrated together to form a common solution which returns the required results. We are using 1) MD5 (Message-Digest 5): It is a widely used cryptographic function with a 128-bit hash value. MD5 has been employed in a wide variety of security applications, and is also commonly used to check the integrity of files. An MD5 hash is typically expressed as a 32-digit hexadecimal number.2) SHA-1(Secure Hash Algorithm-1): In cryptography, SHA-1 is a cryptographic hash function 160-bit hash value. A SHA-1 hash value is typically rendered as a hexadecimal number, 40 digits long. As we discuss the existing system contains two types of source deduplication strategies, 1] Local source deduplication :Detects redundancy in backup dataset from the same device at the client side and only sends the unique data chunks to the cloud storage.2] Global source deduplication :Performs duplicate check in backup datasets from all clients in the cloud side before data transfer over WAN. But we are going to implement something different.

In our system we are considering following points:

1] An ALG-Dedupe, an Application aware Local-Global source deduplication scheme is proposed that not only exploits application awareness, but also combines local and global duplication detection.

2] The system is proposed to achieve high deduplication efficiency by reducing the deduplication latency to as low as the application-aware local deduplication while saving as much cloud storage cost as the application-aware global deduplication.

3] The application design is motivated by the systematic deduplication analysis on personal storage.

4] The basic idea of ALG-Dedupe is to effectively exploit the application difference and awareness by treating different types of applications independently and adaptively during the local and global duplicate check processes.

5] This will help to significantly improve the deduplication efficiency and reduce the system overhead.

4. CONCLUSION

By studying all the previously work done on deduplication we summarize that, ALG-Dedupe is an application aware local-global source-deduplication scheme for cloud backup in the personal computing environment to improve deduplication efficiency is proposed. Also an intelligent deduplication strategy in ALG-Dedupe is designed to exploit file semantics to minimize computational overhead and maximize deduplication effectiveness using application awareness. It combines local deduplication and global deduplication to balance the effectiveness and latency of deduplication.

REFERENCES

[1] Improved Backup Scheduling With Data Deduplication Techniques For Saas In Cloud 1tamilselvi.T, 2k.Saruladha 1Department of Distributed Computing Systems (CSE), Pondicherry Engineering College, Pondicherry 2Department of computer science and engineering, Pondicherry engineering college, Pondicherry.

[2] A Novel Way of Deduplication Approach for Cloud Backup Services Using Block Index Caching Technique Jyoti Malhotra1 ,Priya Ghyare2 Associate Professor, Dept. of Information Technology, MIT College of Engineering, Pune, India1 PG Student [IT], Dept. of Information Technology, MIT College of Engineering, Pune , India2.

[3] Data Deduplication in Cloud Backup Systems Xiongzi Ge, Zhichao Cao CSci 8980, Fall 2013, Final Report Computer Science and Engineering, University of Minnesota, Twin Cities {xiongzi, zcao}@cs.umn.edu.

[4] Hemant Palivela, Chawande Nitin P, Sonule Avinash, Wani Hemant, "Development of servers in cloud computing to solve issues related to security and backup", NJ, USA: IEEE Computer Society, 158-163, 2011.

[5] Dirk Meister, Jürgen Kaiser," Block Locality Caching for Data Deduplication". In Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST). USENIX, February 2013.

[6] Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu,'' Application-Aware local global Source Deduplication for Cloud Backup Service of personal storage " IEEE International Conference on Cluster Computinges in the Personal Computing-Environment(2012).

[7] Mao B, Jiang H, Wu S, et al. SAR: SSD Assisted Restore Optimization for Deduplication-Based Storage Systems in the Cloud[C]//Networking, Architecture and Storage (NAS), 2012 IEEE 7th International Conference on. IEEE, 2012: 328-337.

[8] Tan Y, Jiang H, Feng D, et al. CABdedupe: A causality-based deduplication performance booster for cloud backup services[C]//Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International. IEEE, 2011: 1266-1277.

[9] Ng C H, Ma M, Wong T Y, et al. Live deduplication storage of virtual machine images in an open-source cloud[C]//Proceedings of the 12th International Middleware Conference. International Federation for Information Processing,2011:80-99.

[10] Fu Y, Jiang H, Xiao N, et al. Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage[J]. 2013.