Data deduplication for cloud backup services

Ms.Nidhi Panpaliya
Department. Of Information Technology
RMD Sinhgad school of Engineering
Pune, India
e-mail: npanpaliya25@gmail.com

Ms.Prachi Sorte
Department Of Information Technology
RMD Sinhgad school of Engineering
Pune, India
e-mail:sorte.prachi@gmail.com

Abstract—**Data Deduplication describes an approach which reduces the data storage capacity to store data and also reduces the data transmission traffic on the network. In today's era due to increasing volume and the value of the digital information of personal computing devices have raised a serious and growing requirement for data protection in personal computing environment. A coming up challenge faces source deduplication for cloud backup services low down deduplication efficiency due to combination of resources in demanding nature of deduplication and the limited system resources. An Application-aware Local-Global source deduplication method improves the data deduplication effectiveness by exploiting application awareness with content defined chunking method, and reduces deduplication time. Application-aware mechanism helps to improve the deduplication efficiency with little system overhead and low down the cost for cloud backup services of personal storage.**

Keywords-**Cloud backup, data deduplication, application awareness, deduplication efficiency.**

I. INTRODUCTION

Nowadays, the personal computing systems, such as desktops, laptops, tablets, smart phones have become crucial platforms for several users, increasing the importance of data on these devices. To evade data loss due to hardware failure, unintentional deletion of data, or device theft/loss, individuals have improved their use of data protection and recovery tools in the personal computing devices. Due to the virtually infinite storage resources that are available on demand and charged according to usage, the cloud storage services (e.g., Amazon S3 and Google Storage) take considerable economic advantages to both cloud providers and cloud users. As shown in Figure 1, data backup for personal storage has emerged to be a particularly attractive application for outsourcing to cloud storage providers because users can manage data much more easily without having to worry about maintaining the backup infrastructure. This is feasible because the centralized cloud management has created effectiveness and cost variation point, and the cloud offers simple offsite storage for disaster recovery, which is always a critical concern for data backup.

There are various data reduction techniques, like data deduplication with delta encoding, and Lempel-Ziv (LZ) compression, can significantly improve the storage efficiency by exploiting data redundancy and similarity.

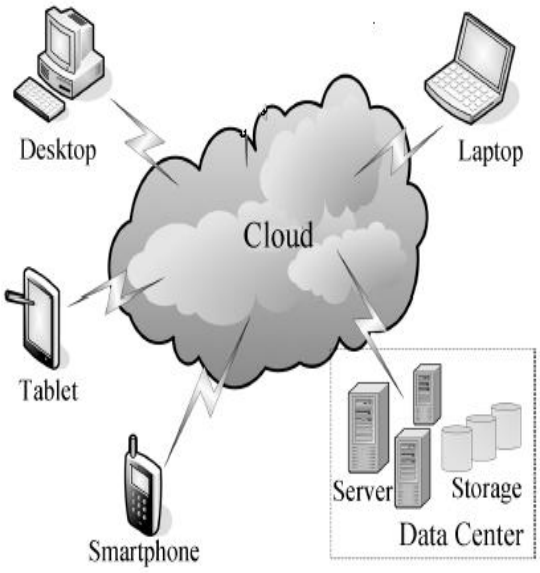

Figure 1: Cloud backup platform for personal computing devices.

Data backup for personal storage in the cloud storage environment implies a geographic division between the client and the service provider. Cloud storage that is usually bridged by wide area networks (WANs), data deduplication is an effective data compression approach that exploits data redundancy, divides the large data objects into smaller parts, called chunks, these chunks can be represented by their fingerprint (i.e., usually a cryptographic hash of the chunk data), replace the replica chunks with their fingerprints once chunk fingerprint index lookup, and only transfers or stores the distinctive chunks for the purpose of communication or storage

efficiency. Source deduplication (i.e Client-side data reduction) that eliminates redundant data at the client site is obviously preferred to target deduplication due to the former's ability to significantly decrease the amount of data transferred over wide area network (WAN) with low communication bandwidth. It has been adopted by various cloud backup services including Dropbox.

The existing source deduplication strategies can be separated into two categories: local source deduplication and Global source deduplication. Local source deduplication[5] that only detects redundancy in backup dataset from the same device at the client side and simply sends the unique data chunks to the cloud storage.In global source deduplication it performs duplicate check in backup datasets from all clients in the cloud side. The local deduplication scheme only eliminates intra-client redundancy with low duplicate removal ratio by low-latency client-side duplicate data check, where global deduplication scheme can contains both intra-client and inter-client redundancy with high deduplication effectiveness by performing high-latency duplication detection on the cloud side.

An application-aware data deduplication scheme is based on different data deduplication techniques such as data reduction ratio and data processing speed with different types of applications[9]. To achieve high data reduction efficiency with high data reduction ratio and shorten the backup window by leveraging both limited local resources on personal computing clients and plentiful global resources of cloud computing.

In cloud storage to provide the data privacy protection and cloud storage cost saving, the main contributions includes:

- Analysis of data reduction techniques on various application datasets.
- Design an application-aware data reduction scheme to reduce cloud storage capacity and shorten backup window based on interpretation on the application-oriented data reduction process.

II. EXISTING SYSTEM

In the traditional storage file systems and storage hardware, every layers contains different kinds of information about the data they handle and such information in one layer is usually not available to any other layers. Code sign for storage and application is probable to optimize deduplication based storage system [1], when the lower-level storage layer has broad knowledge about the data structures and their access characteristics in the higher-level application layer.

Deduplication approach reduces the storage capacity needed to store the data or to transfer the data on network. In cloud backup data storage resources are available on demand which helps to reduce the network space by breaking up incoming stream of data into small segments. To identify such segments block index technique [7] is used. Data deduplication techniques for data reduction are the most effective behavior to promote data storage efficiency by deleting data redundancy. Data reduction technique includes data compression, delta encoding and deduplication.

Data compression eliminates redundancies contained by data objects to characterize original information using fewer bits. This can be either lossy or lossless. Lossless compression reduces bits by identifying and eliminating statistical redundancy in data. The LZ compression methods [2] is most accepted algorithms for lossless storage. It is universal lossless data compression algorithm and It is simple to implement and probable for very high throughput in network implementation.

Lossless compression technique reduces bits by identifying slightly important information and removing it. It gives an equivalent substitution between information loss and the size reduction. In some well-liked applications, including images, audio and video, some loss of information is acceptable. Data compression [4] only achieves a partial data reduction ratio due to its intra-object data reduction nature.

Chunk-based storage system utilizes the file similarity instead of chunk locality [6], Index reside in RAM and index kept on disk. Extreme Binning exploits file likeness instead of locality to make only one disk access for chunk lookup per file instead of per chunk, thus alleviating the disk bottleneck problem. The new data structure in application-aware deduplication [3], application-aware index structure can extensively mitigate the disk index lookup bottleneck by dividing a central index into many independent small indices to optimize lookup performance.

Data de-duplication is performed through hybrid cloud architecture which does not eliminate the redundant data completely. For the data security purpose when the data is stored over the cloud only, and if some data loss occurs it is unable to recover the data. The convergent encryption technique used for encrypting the data is inefficient. The same privilege key is used by the user for storage and retrieval of data for every time. The same privilege key is easily predictable by the hackers or intruders.

III. PROPOSED SYSTEM

The main purpose of data deduplication scheme is to reduce the computation overhead and also to utilize cloud resources to reduce the computational transparency. It can be done by using an intelligent chunking scheme. This system is based on an application-awareness with the use of hash function [1]. To improve the efficiency of system

with low system overhead on client side it combines the local and global source deduplication with application-awareness.

An architectural overview of proposed system is illustrated in Figure 2, where tiny files are first filtered out from file size filter and then non-tiny files are processed in application-aware data deduplicator to eliminate data redundancy from the data stream.

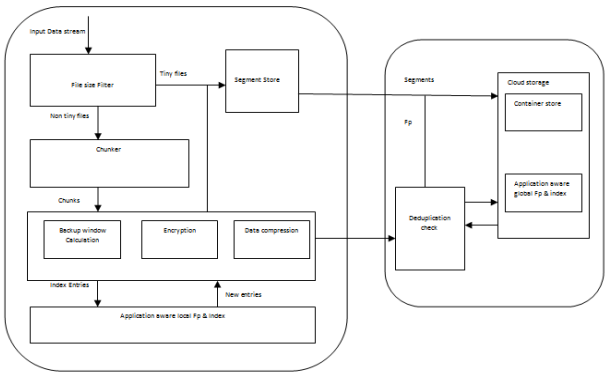

Figure 2: Architectural overview of the Data deduplication design.

Non-tiny files are broken into chunks by an intelligent Chunker using application-aware chunking strategy in Local-Global deduplication scheme. The data chunk from same type of files are Deduplicated in the application-aware deduplicator by generating the chunk fingerprint in hash engine and performs the redundancy check in application-aware indices in both local client and remote cloud. Chunk fingerprints are first looked up in an application-aware local index for local redundancy check. If match found, the data for the file containing that chunk is updated to point to the location of the existing chunk. When there is no match then fingerprint will be sent to the cloud for global duplication check on an application-aware global index. If a match is found in the cloud, the corresponding file data is updated for duplicate chunks, else new chunk is added. Fingerprints will be transferred in batch and new chunk will be packed into large unit called segment in segment store with tiny files before their transfer to improve the network bandwidth efficiency over WAN. The deduplication process explained in more detail in the rest of this section.

A. File Size Filter For Input Data Stream

A backup data contains most of the tiny files which holds a negligibly small percentage of the storage capacity. To minimize the data overhead the proposed system filter out these tiny files in file size filter before starting deduplication process. It forms the group of all tiny files together into large unit in the segment store and such data will be stored as segment in segment store.

B. Intelligent Chunker For Data Chunking

Data chunking scheme has the great impact on the efficiency of data deduplication which taking the data backup. There are different data deduplication algorithms [6] such as Static chunking (SC) and Content defined chunking (CDC).

1. *Static chunking (SC)*

In this technique file is divided into number of fixed sized blocks then apply hash function to create the hash key of the block. This chunking scheme has write once policy since no other block can be found with the same address. The address of multiple writes of the same data are identical, so the duplicate data is easily identified and the block is stores only once. This system has the limitation of boundary shifting problem as all data is shifted but boundary is fixed.

2. *Content defined chunking(CDC)*

In content defined chunking technique each file is partitioned by anchoring based on their data patterns. It prevents the boundary shifting problem of static chunking. In content defined chunking all data is shifted with their variable boundary.

C. Application-aware Deduplicator

The deduplication of the data chunks will be performed in application aware deduplicator after performing the data chunking in Chunker. Application-aware deduplicator contains the hash engine and data compression module. Hash engine generates the Fingerprint (FP) of data chunks and detects the replica of chunk in both local client and remote cloud. When the chunk is a unique data chunk and the compressed chunk is sent to the selective encryption module while the uncompressed chunk is further processed by delta encoding and LZ compression[2] in the application-aware deduplicator. While performing the deduplication check on client side and on global cloud it requires two types of application-aware indices such as application-aware local index on client side and application-aware global index on cloud side.

D. Segment Store

To avoid the higher overhead of network protocol due to small file transfer, application-aware deduplication will group the duplicate data of various tiny files and chunk into large units that will know as segment. This group of duplicate data will be stored in segment store before transferring data over the network.

E. Container Store

Container store is maintained for each arriving backup data stream. A container is nothing but the self describing data structure in chunk descriptor for stored chunk. The data stored at container is nothing but the segments send over cloud with respective its fingerprint.

For providing security to the data stored in the cloud environment the Ramp Secret Sharing Scheme (RSSS) is

used. In existing system Blowfish Encryption algorithm is used for encryption, Blowfish is a symmetric block cipher that can be used as a drop-in replacement for DES. It takes a variable-length key, from 32 bits to 448 bits. The data stored in the cloud are encrypted by the secret key of Ramp Secret Sharing Scheme [8]. The secret key for the service requester is generated at the time of data storage and retrieval from the cloud environment. Data De-duplication and encryption are performed through Convergent Key Management.

Ramp secrete sharing scheme can be performed in two different ways i.e. strong ramp sharing and weak ramp sharing. In the strong secrete sharing key data integrity maintained in high level than the weak ramp sharing key. To store the sensitive basic requirements about how data should be stored in secure manner must consider the Diversity, security and cost effectiveness.

Convergent key provides high level of confidentiality to the users data in order to maintain the security in cloud computing. In this scheme data owner can obtain the convergent key from the each original data copy for encrypting the data. And this can be achieved by obtaining the hash value of the plain text.

IV. CONCLUSION

An Application-aware local-global source-deduplication scheme for cloud backup in the personal computing environment is used to improve deduplication efficiency. Deduplication scheme avoids the redundancy of data over the cloud storage by utilizing application awareness and combines local and global deduplication detection. It hits the good balance between cloud storage capacity saving and deduplication time reduction with the secure deduplication process. This scheme helps to improve deduplication efficiency with low system overhead. The proposed scheme works on the secure deduplication in cloud backup services of the personal computing environment. Application-aware local-global de-duplication system architecture improves power efficiency, shorten the backup window size and save cloud cost for cloud backup service.

V. ACKNOWLEDGMENT

I would like to take this opportunity to express our profound gratitude and deep regard to my guide prof. Prachi Sorte and our Head of dept. Prof. Dhara Kurian for her exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project. Her valuable suggestions were of immense help throughout my project work. Her perceptive criticism kept me working to make this project in a much better way. Working under her was an extremely knowledgeable experience for me.

REFERENCES

[1] Yinjin Fu, Hong Jiang, "Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage," IEEE Transactions On Parallel And Distributed Systems, VOL. 25, NO. 5, MAY 2014.

[2] Yin-Jin Fu, Nong Xiao, "Application-Aware Client-Side Data Reduction and Encryption of Personal Data in Cloud Backup Services," journal of computer science and technology 28(6): 1012-1024 Nov. 2013.

[3] J. Malhotra, P. Ghyare, "A Novel Way of Deduplication Approach for Cloud Backup Services Using Block Index Caching Technique," IJAREEIE Vol.3, Issue 7, July 2014.

[4] A. ElShimi, R. Kalach, A. Kumar, J. Li, A. Oltean, and S. Sengupta, ''Primary Data Deduplication Large Scale Study and System Design,'' in Proc. USENIX ATC, 2012, pp. 285-296.

[5] Y. Fu, H. Jiang, N. Xiao, L. Tian, and F. Liu, ''AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment,'' in Proc. 13th IEEE Int'l Conf. CLUSTER Comput., 2011, pp. 112-120.

[6] K. Eshghi, H. Khuern Tang, "A Framework for Analyzing and Improving Content-Based Chunking Algorithm," Hewlett-Packard Laboratories palo Alto, CA Feb 25,2005.

[7] D. Bhagwat, K. Eshghi, "Extreme Binning: Scalable, Parallel Deduplication For Chunk-based File Backup," 17th IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS' 2009), London, UK, September 2009.

[8] J. Li, X. Chen, M. Li, "Secure Deduplication With Efficient And Reliable Convergent Key Management," Ieee Transactions On Parallel And Distributed Systems, Vol. 25, No. 6, June 2014

[9] A. Katiyar and J. Weissman, \ViDeDup: An Application-Aware Framework for Video De-Duplication," in Proc. 3rd USENIX Workshop Hot-Storage File Syst., 2011, pp. 31-35