Genetic Algorithm based Local-Global Deduplication for Cloud Backup Services

K. Karthick[1], P. Neelaveni[2]
[1]*M. E. Department of Computer Science and Engineering, GKM College of Engg and Tech.*
[2]*Associate Professor, GKM College of Engineering and Technology.*

***Abstract*- Data growth and storage management are the two major challenges faced by both the organizations and the domestic users. The personal computing devices for accessing the data from anywhere have to rely on the cloud storage services. Cloud backup services is the core technology of the cloud storage, but while relying on cloud storage duplicates in the storage space becomes another issue. To overcome this, deduplication is the best solution in order to have a better utilization of the storage space available. Here a Genetic based approach is being introduced for removing the duplicates that are present in the storage environment. This Genetic Algorithm based approach significantly reduces the computational overhead and increases the data transfer efficiency.**

***Index Terms*- Deduplication, Genetic Algorithm, Cloud storage, Cloud Backup Services.**

I. INTRODUCTION

With the development of the trend in smart devices with their ubiquitous networks have led to the increase in the use of the personal computing devices among people and organizations. These type of the personal computing devices are mobile and depend on the online storage for their backup and retrieval. Data and the information come from user generated contents all around the world. Over the past few years cloud computing has become one of the top industries in information technology. Cloud computing can combine physical and virtual resources to cope with a great deal computing services including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). These type of storage environments are not only needed by companies but also for personal use and for educational environments with latter not as rich as former. Therefore, without purchasing the infrastructure, the latter will have to consider other solutions. Hence, they have to efficiently use the space provided to them in a more efficient way, the prerequisite of this paper is this condition. With the explosively growing data, the cloud backup services can ensure data reliability, but the backup space required stands as the burden. For efficient use of these cloud backup spaces, multiple copies of the data should be avoided, such as more storage space could be saved. Data deduplication, an effective data compression approach that exploits data redundancy, partitions large data objects into smaller parts, called chunks.

These chunks of data are represented by their fingerprints, replaces the duplicate chunk with their fingerprints after chunk fingerprint index lookup, and only transfers or stores the unique chunks for the purpose of communication or storage efficiency. The overall architecture of the Cloud backup service is shown in the fig. 1. This data deduplication is a resource intensive process and needs high computational environment for identifying and eliminating duplicate data. Such resources could not be possibly made available for the personal computing devices. So it is necessary to achieve the deduplication efficiency and the system overhead for personal computing devices with limited system resources.

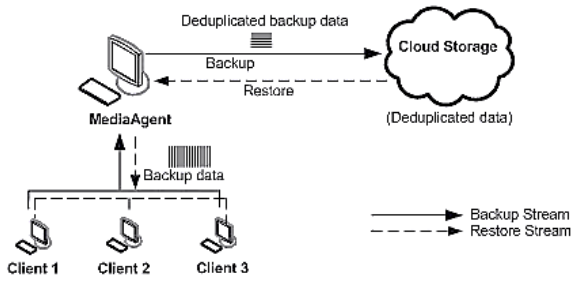

Fig. 1. Architecture of Cloud Storage

But the available deduplication solutions focus only on the effectiveness of the deduplication to remove more redundant data without considering the system overhead and available resources. Thus in order to ensure the data deduplication process in the personal computing devices with low computational efficiency, and to make the data deduplication more efficient, we develop a deduplication model based on genetic algorithm. The genetic algorithm is the best optimization problem available. Thus, applying the genetic algorithm in the cloud environment not only reduces the deduplication efficiency, but also increases the backup and restores performance of the cloud backup servers. Due to the increased deduplication efficiency, the bandwidth of the backup and restore functions are also reduced. Thus the implementing the deduplication using the genetic algorithm in the cloud backup environment increases the overall performance of the cloud storage and backup services.

II. CLOUD BACKUP SERVICE

Even though cloud is engineered to prevent data loss, maintaining the recent backups of our data is still to be considered as a fundamental practice. Cloud backup is file-based backup solution that uses compression and encryption techniques in order ensure your data is protected and recoverable. The data backup is normally done at particular time period, or manually by the user whenever it is necessary. The backup data are chunked and are stored in the storage space provided by the cloud backup service provider. The backup data is sent as a stream, so as to make the chunking easy and efficient. Various methods are available for chunking the data stream into the smaller data units. Thus the cloud backup more or less becomes an everyday routine depending upon how frequently the data backup is necessary. The cost of the cloud backup service depends on the size of the storage space utilized. Thus the utilization of the cloud space is an important factor in case of the cloud backup services.

III. DATA DEDUPLICATION

There are different storage optimization techniques that are faced while using data backup service. Data deduplication is the key technique among the techniques available for optimizing the available storage space. This deduplication technology identifies the duplicate data, eliminate redundancy and reduce the need to transfer or store the data.

IV. DESIGN AND IMPLEMENTATION

A. Architecture

The data backup stream from the computing device is sent to the application aware module in which chunks the data stream is chunked into the various chunks depending upon the type of the data that it belongs to. Then the files are taken to similarity function depending on the application or the type of the file. Then the corresponding similarity function is applied to the files. Then as shown in the fig. 2**.,** the similarity function checks for the similarity between the files in the cloud and the data being uploaded. If the similarity of the files defines to be duplicate of the file already present in the cloud backup, then the data will generated as pairs and sent to the genetic programming module for deduplication process.

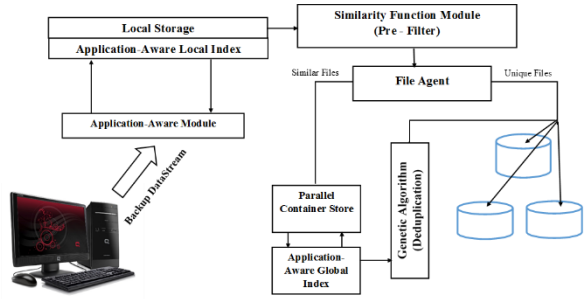

Fig. 2. Architecture of the Proposed System

The feature extraction module is responsible for the process of extracting the files that are announced to be similar by the similarity functions used in the previous stages of the process.

B. Application Aware Chunking

The deduplication efficiency of data chunking scheme among different applications differs greatly. Depending on the type of file, the static chunking or dynamic chunking could be performed. The dynamic uncompressed files are always editable, while the static files are inevitable in common. Depending on the type of files used, various techniques are applied to generated hash values.

C. File Similarity Search

File similarity search concept is widely used in data processing system area. The main idea is to extract hash keys from a file. Usually, Rabin hash function calculates hash key from a file and stores the

hash key in a queue. By shifting one byte step by step, Rabin hash function repeatedly generates hash key and insert the hash key to the queue. The queue contains only several number of keys in ascending order or in descending order by configuration of the system. If Rabin hashing is finished, there remains several hash keys whose value is maximum or minimum. These key value is used for file similarity search. When A file and B file have duplicated hash keys, this means that the file have duplicated region of data.

D. File Agent

File Agent is a software program that provides a functional interface (file backup/restore) to users. It is responsible for gathering datasets and sending/restoring them to/from Storage Servers for backups/restores. To apply SAM to the system, File Agent adds two key functional modules, Incremental Backup and Local Chunk-level Deduplication (consisting of local chunk-fingerprint identification and local chunk-fingerprint store).

E. Genetic Algorithm

We present a genetic programming (GP) approach to deduplication. Our approach combines several different pieces of evidence extracted from the data content to produce a deduplication function that is able to identify whether two or more entities in a repository are replicas or not. Our aim here is to foster a method that finds a proper combination of the best pieces of evidence, thus yielding a deduplication function that maximizes performance using a small representative portion of the corresponding data for training purposes. Then, this function can be used on the remaining data or even applied to other repositories with similar characteristics. Moreover, new additional data can be treated similarly by the suggested function, as long as there are no abrupt changes in the data patterns, something that is very improbable in large data repositories.

It is worth noticing that this function, which can be thought as a combination of several effective deduplication rules, is easy and fast to compute, allowing its efficient application to the deduplication of large repositories.

A simple GA works as follows

1. Start with a randomly generated population.
2. Evaluate the fitness of each individual in the population
3. Select individuals to reproduce based on their fitness
4. Apply crossover with probability Pc
5. Apply mutation with probability Pm
6. Replace the population by the new generation of individuals
7. Go to step 2.

The architecture of the Genetic Algorithm is shown in the fig. 4.2. Thus, we generalize our previous results in by showing that our GP-based approach is also able to automatically find effective deduplication functions, even when the most suitable similarity function for each record attribute is not known in advance.

This is extremely useful for the non-specialized user, who does not have to worry about selecting these functions for the deduplication task. This distinguishes our approach from all existing methods, since they require user-provided settings frees the user from the burden of choosing the replica identification boundary value, since it is able to automatically select the deduplication functions that better fit this deduplication parameter.

V. RESULTS AND DISCUSSION

The process of uploading the data from the local storage to the cloud backup storage has been initiated. The perquisites for the deduplication process are made through, that is the files to be deduplicated are also uploaded. The intermediate process of finding the similarity between the files in the backup data stream with the data in the cloud as also validated through the file similarity search and file pattern search modules. The results provide the set of similar files, that are found to be matching either with the files from the backup stream or from the cloud. These similar files are generated as pairs and are sent to the Genetic Programming module for the process of Deduplication. The Genetic algorithm with the active pairs looks for the duplicate files and then eliminates them if found any. Indexes are maintained in each and every part of the operation, such that further operations on same type of data could be eliminated. The indexes found here will contain the complete details of where the duplicates are found and also the details of the similarities among the files found be pattern and similarity search modules. Thus, the process of deduplication is highly efficient as the duplicates are found at faster rate

when comparing to traditional deduplication mechanisms.

REFERENCES

1. Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu, and Lei Xu: "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. T. Yujuan, et al., "CABdedupe: A Causality-Based Deduplication Performance Booster for Cloud Backup Services,"IPDPS IEEE International, 2011, pp. 1266-1277.
3. Abdullah Gharaibeh, Cornel Constantinescu, Maohua Lu, Anurag Sharma, Ramani R Routray, Prasenjit Sarkar, David Pease, Matei Ripeanu, "CloudDT: Efficient Tape Resource Management using Deduplication in Cloud Backup and Archival Services"
4. Bo Mao, Hong Jiang, Suzhen Wu, Yinjin Fu, Lei Tian "SAR: SSD Assisted Restore Optimization for Deduplication-based Storage Systems" in IEEE Seventh International Conference on Networking, Architecture, and Storage, 2012, 328-337
5. F. Yinjin, et al., "AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment", IEEE International Conference on Cluster Computing, 2011, pp. 112-120.
6. J. Wei, H. Jiang, K. Zhou, and D. Feng, "MAD2: A scalable high throughput exact deduplication approach for network backup services," in Mass Storage Systems and Technologies (MSST), 2010 IEEE 2^{6th} Symposium on, Incline Village, NV, USA 2010 pp. 1 – 14
7. Kiatchumpol Suttisirikul, Putchong Uthayopas, "Accelerating the Cloud Backup using GPU based Data Deduplication", 2012 IEEE 18th International Conference on Parallel and Distributed Systems
19. to genetic algorithms" in Springer, NewYork, 2008.
8. Lei Xu, Jian Hu, Stephen Mkandawire and Hong Jiang, "SHHC: A Scalable Hybrid Hash Cluster for Cloud Backup Services in Data Centers", 2011 31st IEEE International Conference on Distributed Computing Systems Workshops
9. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, and I. Stoica, "Above the Clouds: A Berkeley View of Cloud Computing," UC Berkeley, Tech. Rep. 2009-28, Feb. 2009.\
10. P. Mell, and T. Grance. The NIST Definition of Cloud Computing, Draft by The National Institute of Standards and Technology (NIST). United States Department of Commerce Version 15., July2009,
11. Quinlan, S., and Dorward, S. Venti: a new approach to archival storage. In Proceedings of the FAST 2002 Conference on File and Storage Technologies (2002), vol. 4.
12. S. Quinlan and S. Dorward. Venti: A New Approach to Archival Storage. In *FAST '02:* Proceedings of the Conference on File and Storage Technologies, pages 89–101, Berkeley, CA, USA, 2002. USENIX Association.
13. S. Rhea, R. Cox, and A. Pesterev, "Fast, inexpensive content addressed storage in Foundation," in *USENIX'08*, Jun. 2008.
14. Shu-Ching Wang, Kuo-Qin Yan, Shun-Sheng Wang, Bo-Wei Chen "LDMCS: a Lightweight Deduplication Mechanism under Cloud Storage", Business and Information 2013, E32-E40.
15. Syncsort Backup Express and NetApp, "http://www.syncsort.com."
16. T. Clements, I. Ahmad, M. Vilayannur, and J. Li, "Decentralized deduplication in SAN cluster file systems," in *USENIX'09*, Jan. 2009.
17. Akshara k., Soorya P. "Replica free repository using genetic programming with decision tree" in International Journal of Advanced Engineering Applications, Vol.1, Iss.2, pp.62-66 (2012).
18. S. N. Sivanandam and S. N. Deepak "Introduction