IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11815

A Multi Path Novel De-duplication-based Cloud-of-Clouds Storage Service

Satish K[1], Dr. Divakar Harekal[2], Ms. Veena G.S[3]

Dept of Computer Science, Ramaiah Institute of Technology, Bangalore, India[1,2,3]

Abstract: The massive expansion of online digital information has increased the demand for storage solutions. The Total Cost of Ownership, which includes storage infrastructure costs, management expenses, and human administration costs, grows in tandem with the volume of data. Reducing the quantity of data that has to be transported, stored, and maintained becomes critical in large-scale distributed archival storage systems, and it also helps application performance, storage costs, and administrative overheads. De-duplication is a storage-saving method that has shown to be extremely effective in business backup setups. A same data block in a file system may be saved numerous times across different files; for example, several copies of a file that are substantially similar may exist. It localizes data replication and eliminates redundancy; by storing data just once, all files that employ identical areas refer to the same unique data. In this project we extend a existing cloud of clouds service to make the process of routing and storing of data chunks on a cloud network more efficiently.

I. INTRODUCTION

Cloud computing is essentially the process by which users can outsource their computing needs and save data or information on the cloud via the Internet. The data that is saved in the cloud mostly confidential and needs to be protected, such as social networks and medical records. Thus, privacy and security are key challenges in cloud computing. The important issue is that the user should check oneself before starting any transaction, and it should be protected so that other users should be unaware user's congruity. The cloud holds the user liable for storage for various purposes, and the cloud itself is also responsible for its services. The user's honesty in storing the information is also certified. Cloud is prone to server collusion attacks and data manipulation.In a server scheme attack, the rivals can attack the storage servers. Data encryption is necessary for the safety of the data. However, data is frequently modified, and this aspect must be considered when creating a safe and secure storage systems. Accountability is not accountable for the clouds, or should users refuse any requested or completed actions. It is necessary to keep a record of all transactions.Parallel and distributed systems specify a collection of processing units that interact and collaborate to achieve a goal by connecting a number of nodes in a network.Many firms are migrating their data to the cloud as cloud storage becomes more popular. Putting all data in one cloud, on the other hand, presents issues as vendor lock in,data availability and service costs.

In this project,a Cloud-of-Clouds storage service is designed to help businesses outsource its data to be stored on the cloud. We utilise three main strategies to attain the goal of low cost and high availability. First, using an applicationaware chunking mechanism, it reduces superfluous data at the client side to decrease storage costs. Secondly, for high availability, adopts an inner-chunk based erasure coding mechanism to distribute unique chunks across many clouds. Finally, a container-based share management mechanism is used to optimise performance.

II. BACKGROUND AND MOTIVATION

DCStore is a Cloud-of-Clouds storage solution that allows businesses to outsource their data to the clouds. We utilise three main strategies to attain the aim of low cost and high availability. First, using an applicationaware chunking mechanism, DCStore reduces superfluous data at the client side to decrease storage costs. Second, for high availability, DCStore employs an inner-chunk based erasure coding approach to distribute unique pieces across several clouds. Finally, for performance optimization, a container-based share management method is implemented

A. De-Duplication

To split the input file, there are two main ways. The first method, known as static chunking, divides data into fixed-size pieces. The key advantage of this technology is that it has a very high throughput and is simple to install. The advantage of this technology is that it has a high throughput and is simple to apply. While employed effectively by Dropbox [1], this strategy has a fundamental drawback: when bytes are added at the beginning of a file, all subsequent chunk boundaries are changed [2], i.e., any chunks after the addition will have different hash sums and will become obsolete for that file.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11815

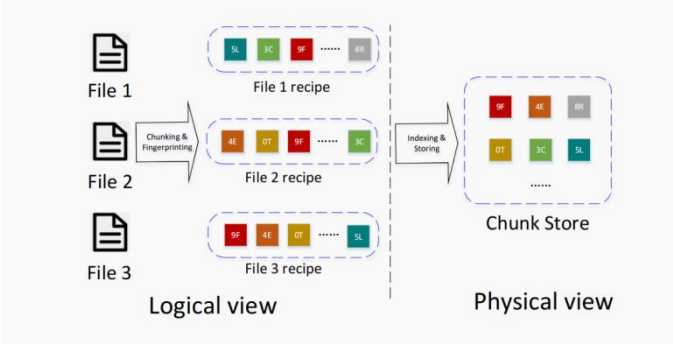

Fig 1. Overview of De-duplication proces.

The second chunking approach, known as content defined chunking (CDC), eliminates this problem. However, the chunking algorithms described above do not make advantage of the underlying data's file attributes. If the chunking approach understands the file's data stream, the deduplication method can provide.

B. Erasure Coding

Erasure coding in a Cloud-of-clouds storage by offering chunk redundancy among many providers, cloud storage services enable clients to conceal transitory disruptions and boost data availability. Erasure coding encrypts k data blocks and generates r parity blocks, so any k of the total (k + r) blocks is enough to decode the original k data blocks.

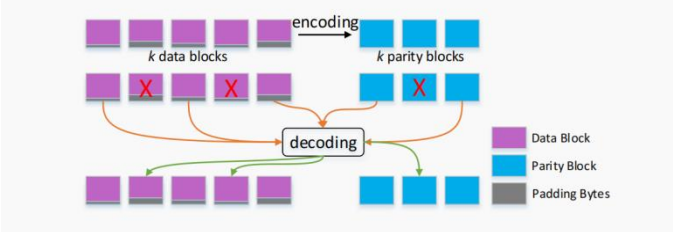

Fig2: Overview of Erasure Coding.

Erasure coding is used across fixed-size items in various storage systems [3]: k objects are encoded to produce r new objects. Unless the original object is absent, read requests to an object are serviced from it. Parity reconstruction incurs considerable bandwidth overheads if the item is missing.Furthermore, de-duplicated storage must pack various size chunks into fixed-size objects, which will result in zero-byte padding[4].

III. LITERATURE SURVEY

LWen Xia [5] in this paper after reviewing the background and key features of data de-duplication he found that in secondary storage system is shown a data reduction of 5% to 40% and 40% to 60% on primary storage.

Chuanyi Liu [6] ADMD suggested an application driven metadata aware de-duplication system that employed metadata information at various stages of I/O pipeline to make file partitioning into more relevant data chunks, decreasing inter-file level duplications even more.

Chuanyi Liu [7] in this paper he establishes a de-duplication technique know as R-ADMD and compares them with RAID like schemes, It demonstrates that R-ADMAD may provide comparable dependability to replication based schemes with significantly more redundancies while giving the same storage consumption as RAID like schemes. Because RADMAD provides a distributed and dynamic recovery mechanism, the typical recovery time of R-ADMAD-based systems is 2-5 times that of RAID like schemes.

Sujata V. Randad [8] proposed an intelligent deduplication strategy ALG-Dedupe to use file semantics to minimise computational cost and increase deduplication efficacy. To balance the efficacy and latency of deduplication, it mixes local and global deduplication. The suggested application-aware index structure, which divides a central index into many independent tiny indices to maximise lookup efficiency, can greatly alleviate the disc index lookup bottleneck.

IV. IMPLEMENTATION

A. Architecture Overview.

In the below Fig 3.we can see the architecture used for this project.It mainly vanilla javascript for the forntend view of the project.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11815

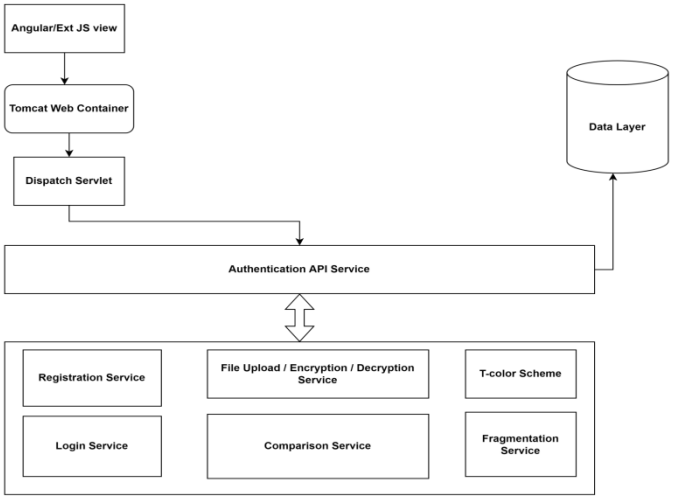

Fig3. Overview of the System.

The server side of things are handled by Tomcat webserver with the use of java servlets.The API service make sure to carry data back and forth securely. Most the services are accessed after valid authentication through API's.

B. Novel De-duplication algorithm.

The optimized algorithm is show in the below fig4.Source and Destination Node along with its Transmission Range and TTL are taken.Neighbor Nodes to these are founds along with the length of each neighboring nodes are found.All the above is done form the first node to all the neighbor node using the Novel-DeDuplication Method to discover Individual Paths between the Source and Destination Nodes and all the paths are cached then it is repeated until all paths are found.

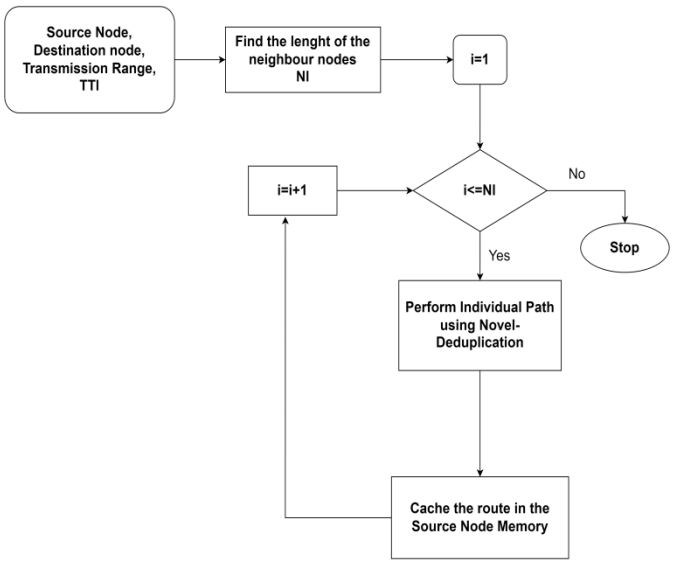

Fig 4.Novel De-Duplication Algorithm.

All the data is encrypted using AES encryption for secure file Storage.

V. RESULTS

In the Novel De-Duplication all the possible routes are found using its own Individual path algorithm and shortest path algorithm by consider the transmission range and TTL of each node on the network and the among the multiple paths that have been found the best path has been selected.The algorithms performance is then measured as shown in the using the time taken for it to calculate all the path and then select the best path among them.The total number of hops is the number of nodes it had to jump to finally conclude the best path among them.According to the threshold value given the number of dead and alive nodes are found after the algorithm has selected the best path.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11815

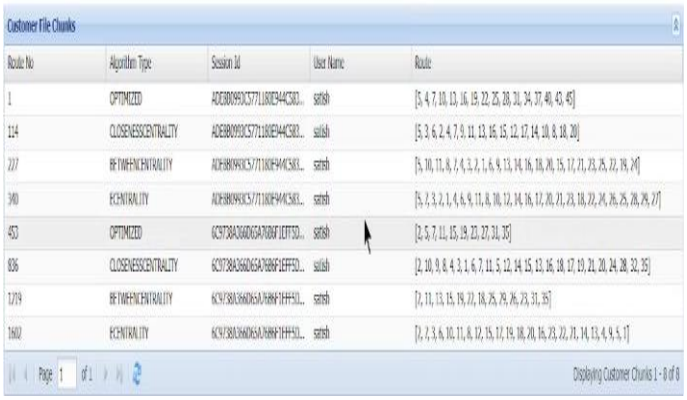

Fig 5: Algorithm Path Discovery.

From the Fig 5.we can the nodes that was chosen as the best path to traverse through the network and store the data chunks in the them.

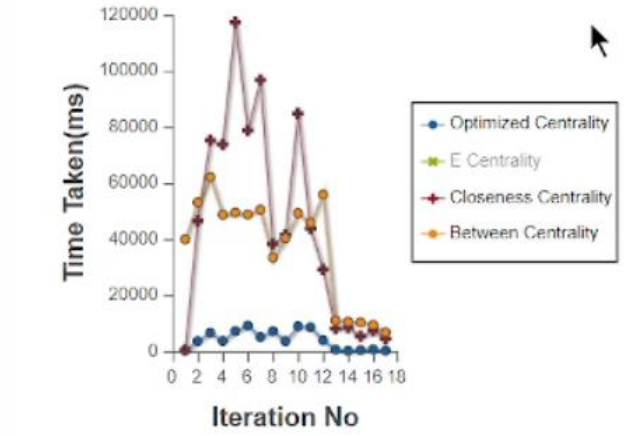

Fig 6: Performance Overview

As we can see in both of the experiment concluded the optimized algorithm has the shortest path with less routing overhead and higher throughput when compared to the other algorithms as Shown in Fig 6.

VI. CONCLUSION

In this work, we extend cloud of clouds Storage using a deduplication-based storage solution that addresses the dependability of cloud storage services. From the experiment we can conclude that the optimized Novel De-duplication algorithm was able to find the multi path to the destination in a fewer amount time by also keeping almost all of the nodes alive that were present on the cloud network.It also performed at low routing overhead and with higher throughput when compared with other algorithms depicted on the base paper.

REFERENCES

1. Drago, M. Mellia, M. M Munafo, A. Sperotto, R. Sadre, and A. Pras, “Inside Dropbox: understanding personal cloud storage services,” in Proceedings of the 2012 Internet Measurement Conference. ACM, 2012, pp. 481–494.
2. C. Policroniades and I. Pratt, “Alternatives for detecting redundancy in storage systems data.” in USENIX Annual Technical Conference, General Track, 2004, pp. 73–86.
3. C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, “R-ADMAD: High reliability provision for large-scale de-duplication archival storage systems,” in Proceedings of the 23rd international conference on Supercomputing. ACM, 2009, pp. 370–379.
4. W. Chen, Y. Hu, S. Yin, and W. Xia, “EEC-Dedup: Effificient erasure coded deduplicated backup storage systems,” in Ubiquitous Computing and Communications (ISPA/IUCC), 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on. IEEE, 2017, pp. 251–258.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11815

5. Yucesoy, Wen Xia ,Dan Feng,Fred Douglis,Philip Shilane,Min Fu,Yucheng Zhang,yukun Zhou ”A Comprehensive Study of the Past, Present, and Future of Data Deduplication” in Proceedings of the IEEE (Volume: 104, Issue: 9, September 2016)
6. Chuanyi Liu, Yingping Lu, David Du and Dongsheng Wang. ADMAD: Application-Driven Metadata Aware De-duplication Archival Storage System. International Workshop on Storage Network Architecture and Parallel I/Os (SNAPI 2008) held in conjunction with the 25th IEEE Conference on Mass Storage Systems and Technologies (MSST 2008).
7. Chuanyi Liu, Yu gu,Linchun Sun,Bin Yan,Dongshemg Wang “R-ADMAD: High Reliability Provision for Large-Scale De-duplication Archival Storage Systems” ICS '09 Proceedings of the 23rd international conference on SupercomputingJune 2009.
8. Sujata V. Randad,V. R.Chirchi “Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage using ALG-Dedup Algorithm” Journal of Emerging Technologies and Innovative Research (JETIR) 2019 JETIR June 2019, Volume 6, Issue 6.