International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Neoteric Geo-Distance Based 2- Replica Placing Algorithms on Cloud Storage System

Dr. S. Annal Ezhil Selvi[1]

Submitted: 24/04/2023 **Revised**: 24/06/2023 **Accepted**: 06/07/2023

Abstract: In today's digitalized world, all of us have moved to Cloud technology for cloud services to reduce the burden of maintenance issues and to incur lower storage costs than traditional methods. The main reason for the movement of people to the cloud is, it includes its 24/7 service, reliability in all situations, and suitability for large amounts of data storage. Meet out the requirements of high availability and reliability, it adopts a replication system concept. In replication systems, objects are replicated multiple times, and each copy resides in a different geo- location on a distributed computer. It is vulnerable to threats to the Cloud Storage System (CSS). So, this research seeks to explore the mechanisms to rectify the issues mentioned above. Thus, this research work has proposed an algorithm named as 2-Replica Placing (2RP) algorithm which is used to reduce the storage cost, maintenance cost; and maintenance overheads as well as increase the available storage spaces for the providers. This proposed algorithm is placing the data files on two locations based on Geo-Distance and it is used to store only 2-replications with one original file far away from each other in data centers. The future direction of the research is to maintain the 2-replica concept forever even during a disaster occurring time.

Keywords: *Cloud Storage, Data Availability, Data Replication, Edge Computing, Reliability, Storage Cost and 2- Replica Placing.*

1. Introduction

It is no exaggeration to say that today's world is built on data. In such a situation, all are forced to have all the data in our hands every minute. Keeping all the data at hand is not possible. At the same time, if the user digitizes all the data, the storage facility is the biggest question mark and maintenance of that data is also a very big difficult and challenging task for everyone. So, everyone has moved to a great technology called Cloud computing where they can access all the data from anywhere anytime they have an internet enabled device.

Cloud storage is a system that operates on a "on demand" or "pay per use" basis. In cloud computing, all computational resources (such as storage and data) are shared among users [1, 4]. The user and the service provider are linked by a service level agreement (SLA) which QoS parameters. High performance and fault tolerance are crucial considerations [2] which are achieved through replication concept [3].

Users will want a service provider that guarantees maximum demand for data storage. Consequently, replication is used to achieve high availability [4 and 5]. But at the same time, the benefits of replication do not necessarily outweigh the costs. Therefore the cost of replication is an essential consideration [5 and 6].

[1] *Associate Professor, PG. Department of Computer Science, Bishop Heber College (Affiliated to Bharathidasan University), Trichy, TamilNadu, India- 620017.*
** Corresponding Author Email: annalabel.cs@bhc.edu.in*

This study suggested an algorithm for cloud storage systems which algorithm determines where the replica will be stored. The location selection is done based on the geographical distance. This proposed algorithm keeps the data file at three different locations based on geographical distance which will be discussed in the forthcoming sections in detail.

The remaining portion of the paper is structured as follows. Section 2 contains a review of the literature. Section 3 describes the existing system. Section 4 describes the proposed Geo-Distance-based 2-Replica placing (2RP) algorithm. Section 5 discusses the interpretation and comparison of Expected results. Finally, the work is going to concludes by discussing future plans in section 6.

2. Related Works

Two main strategies were used to implement a replication system in cloud storage in the predecessor of this research work which has done by Annal et al.[5 and 16]. They have optimized the storage cost by reducing the number of replication in 2 ways they are, a static method and a dynamic method. The benefits of a static replication system are reliability, scalability but the drawbacks are listed as unnecessary storage usage, neither any flexibility, nor any scalability, as well as increased storage costs. Dynamic methods give better result of high availability but the reliability concerns it is a very big puzzle.

Aral et al. [7] proposed a distributed data propagation algorithm that depends on dynamic creation, alteration, update or deletion of replicas steered by constant

monitoring of data requests from network infrastructure of the edge nodes. And their methodology took advantage of the geographic location of data during widespread processing because of common data requests originating from close customers. This method illustrate, using both actual and synthetic data that even a decentralized replication placement approach can give significant cost savings over client-side caching, which is widely used in old-fashioned distributed systems.

Changsong et al. [8] formalised the replica placement problem as a classical multi-knapsack problem, and two heuristic measurements were initiated to gain a suboptimal solution. They stated that numerous experiments were carried out to investigate the performance of the proposed algorithm, and that their experimental results outperformed many other existing approaches while also significantly improving the throughput for data-intensive applications. Zhen et al. [9] used a storage allocation mechanism that reduces data duplication while retaining a high level of data reliability. To that end, they introduced a new design based on function generation, exhibiting that the storage scheduler can reduce redundant information. They claimed that the insightful presentation of the proposed solution resulted in several benefits, most notably the reduction of the search space and the acceleration of the computation. The progress is evaluated in redundant data savings by continuing to follow the availability indications collected from the physical world, one that revealed that data redundancy was reduced by up to 30% once compared to the mechanism.

The two major concerns of cloud storage systems, according to Wenhao et al. [10], are data reliability and storage costs. As a result, they described PRCR, a low-cost data reliability management process focused on a standard data reliability prototype. Proactive replica checking has served as a cost-effective benchmark for replication-based approaches. According to their simulation results, PRCR reduced cloud storage. According to their simulation results, PRCR scaled back cloud storage space utilization from one-third to two-thirds when compared to the traditional three-replica tactic, significantly lowering cloud storage expense.

Abdenour et al. [11] suggested a dynamic replication strategy. They claim that their methodology still improves provider profitability without neglecting customer satisfaction.

Daya et al. [12] addressed some significant page ranking algorithms before proposing theirs, dubbed user priority oriented page ranking. This algorithm is effective in terms of significance since that uses investigators to determine the relevance of results pages, as well as user behaviour, and user choice-based page ranking tries to make users' search result interaction easier and even more fulfils when ranking web pages to discover the needed data.

To enhance the availability of cloud-based data storage systems, Annal et al [14] suggested a replication automated system based on the popularity (hit rate).

Chunlin et al [15] wanted to experiment with and recommended the grey Markov chain-based dynamic replica artistic style, stating that if the percentage of replicas is required to be increased, the recently introduced replicas must be positioned on the Nodes in the cluster. They stated that their Fast Non-dominated Sorting Genetic algorithm-based replica placement strategy took into account the problem of data replica synchronization as well as the backup and recovery of failed DataNodes in the edge cloud system.

2.1. Existing System

Instead of homogeneous architecture, Navneet et al. [14] used heterogeneous architecture. The original copy of the each data file was stored in Super data centers. In order to meet out the availability concern, the data centres were widely used replicas mechanisms. Whenever the user initiates request of the file, the access request send to the intermediary centres, and it communicates internally with the replication table which contains information about where copies of the requested data files are kept then the request completed or aborted based on suitation.

The brokers will review the list of available centres received and route the request to the adjacent data centre. The fundamental storage unit is referred as a Data Block or Data Unit. Because a file can be replicated across multiple data centres, distinct blocks of a file may be available in various types of data centres. Because super data centres have high-cost, dependable hardware, they have such a massive number.

In order to optimize the replication cost this work was designed Dynamic Cost- Aware Re-replication and Rebalancing Strategy (DCR2S). If the need for availability is increased this algorithm is to switch the location of the replicas from super to main data centers and/or main to ordinary data centers rather increase the number of replications for optimizing the replication cost.

3. Proposed System

As stated in the introduction, the replication framework is used to accomplish the major characteristics of CSS also including high availability, high reliability, and high performance. However, the cost of replication storage and administration may be slightly higher. As a result, users may be hesitant to switch to cloud storage. Despite this, the number of cloud storage customers has dramatically upped. As a result, the proposed system's focus shifts to cloud storage.

Figure 1 depicts the Life Cycle of a Cloud Storage System. In order to reduce the storage space, cost, maintenance cost and overhead due to replication, this research have proposed an algorithm called 2-Replica Placing (2RP) algorithm which is also used increase the available storage space without affecting the key features of CSS that will be discussed in next section.

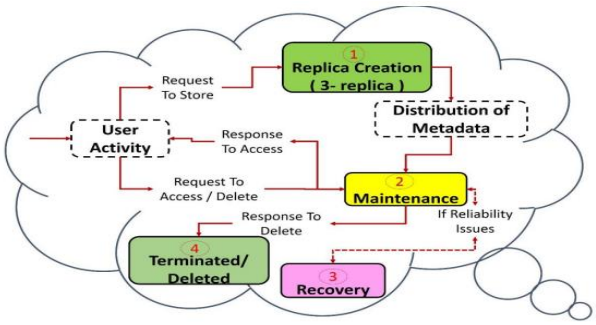

Fig. 1. Magnetization as a function of applied field. Note that "Fig." is abbreviated. There is a period after the figure number, followed by two spaces. It is good practice to explain the significance of the figure in the caption.

The data files are maintained in various data centres across the globe in distributed data storage. There are additional data centres in each Data Center (DC), such as backup or ordinary data centres. And all the DC's have different managers and logs to maintain the data storage. The replication manager and the replication scheduler is the very important role to connect the all other data centers. This work doesn't change anything in the existing system models [7]. But the only difference, there is no variation in storage cost. It may store in super, primary, secondary or ordinary data centres but the storage cost is the same. The mathematical notations used are described in table 1.

Table 1. Symbols and Notations

Symbol	Notations
M	Number of uploaded Data Files
N	Number of Data Center (DC)
Di	Distance between ith DC to jth DC, where i ← 1,…,n Λ j ← 1,…,m
DFj	jth file, j ← 0,1 Λ 2 (DF0 ← Original Data File
DCi	ith Data Centres where i ← 1,2…n
D	Selected distance ith DC to jth DC, where i ← 1,…,n Λ j ← 1,…,m.
S	Selection variable which is used to select the far geo-distanced DC to store the next DF.
Di,j	Distance from i^{th} DC to j^{th} DC. It is shown in the following matrix (n X n) representation

$$D_{i,j}: \begin{array}{c|cccc} & DC_1 & DC_2 & \cdots & DC_n \\ \hline DC_1 & D_{1,1} & D_{1,2} & \cdots & D_{1,n} \\ DC_2 & D_{2,1} & D_{2,2} & \cdots & D_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ DC_n & D_{n,1} & D_{n,2} & \cdots & D_{n,n} \end{array}$$

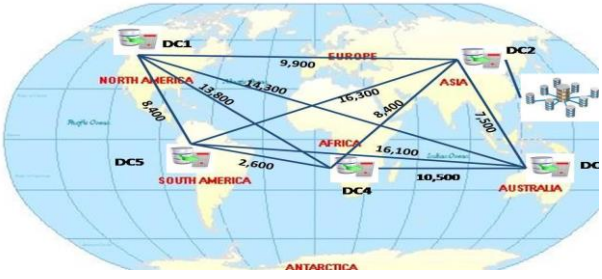

Fig. 2. The System Model

The 2-Replica Placing (2RP) algorithm is used to work based on Geographical Distance (GD). This algorithm plays when the user makes the request to store their data file on a cloud storage system. For an example, the sample system model configuration has shown here the following figure 2.

3.1. Proposed Geo-Distanced based 2RP Algorithm

1: {DC (1), DC (2) … DC (n)}

2: {Data File (DF) is initiated to store on Cloud Storage (CS)}

3: DF1=DF and DF2=DF {Create Two Copies of DF}

4: Store DF on DC(i) { DC(i) which is nearest DC}

5: Calculate Distances {D(1), D(2) …D(n) which is from DC(i) to all other DC's}

6: Arrange that Distances in Descending order

7: D ← D(1).

8: Store DF1 on DC(j) {D(1)'s data Centre which is longest distance from DC(i)}

9: s←2

10: D(s)'s Data centre DC(k) is Selected

11: Select the distance 'd' from DC(j) to DC(k)

12: if (d>= D(1) then Goto 17

13: else s ←s+1

14: while (s<=n/2) then Goto 10

15: D(1) ←D(1)-500

16: if (D(1)>= D/2) then Goto 9

17: Store DF2 on DC(k) {Which far from DC(i) and DC(j)}

The 2RP stores the original Data File (DF) on the nearest Data Centre (DC) once the DF is requested by users to store on Cloud Storage. Before that the replication system replicates two more copies of that DF such as DF1 and DF2. After that the replication system calculates the distances from the original data file residing in the data centre to all other data centres. Find the longest distance DC from the original DF residing DC and store DF1 on that location. Then find and store the DF2 in the next longest distance DC which is far from the DF and DF1 residing DC's.

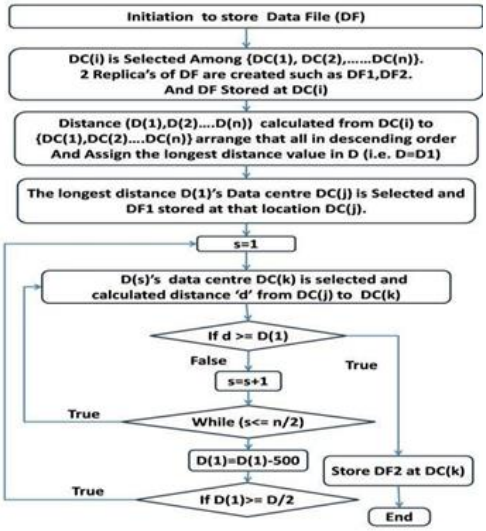

Fig. 3. Flow chart of 2RP Algorithm

Even if that process is also not possible to select the new DC. It will undertake another way to select the new DC. Again it will take the actual distance into consideration. It will satisfy at least half the distance of the actual distance from both the DCs. Then it will store the copy of the DF on that new DC. The above mentioned two ways of the process to select the new actual distance reduce by 500 kilometers until to reach the half distance of the actual distance which is depicted in the figure 2. The working flow of the 2RP has displayed in the figure 3.

Distances (D) =

	DC(1)	DC(2)	DC(3)	DC(4)	DC(5)
DC(1)	0	9,900	14,300	13,800	8,400
DC(2)	9,900	0	7,500	8,400	16,300
DC(3)	14,300	7,500	0	10,500	16,100
DC(4)	13,800	8,400	10,500	0	2,600
DC(5)	8,400	16,300	16,100	2,600	0

Fig. 4. Data Centre's Distances in matrix Representation

4. Interpretation and Comparison of the Expected Results

The System Model DC(1), DC(2), DC(3), DC(4) and DC(5) are data Centre's located in North America, Asia, Australia, Africa and South America respectively. And the DC's connected like a complete graph. The distance between the DC's are represented in Kilometers' (KM) as a metrics form as follows. The Figure 4 shows the instance of an assumption that if DF is initiated to store on cloud storage from North America. Then that DF is stored at the nearest DC which is North America (DC(1)) and makes copies of DF such as DF1,DF2. Then arrange the distances from DC(1) to all other Data Centre's in descending order. That Distances D(1), D(2), D(3),D(4) and D(5) are 14300 (DC(3)), 13800 (DC(4)), 9900 (DC(2)), 8400 (DC(5) and 0 (DC(1)) respectively. Then select the long distanced DC

form DC(1) which is Australia(DC(3)(D(1)=14,300)). Then, the DF1 stores at DC(3).

Then it selects the next long distanced DC which is Africa (DC(4)(13,800)). After that the distance'd' will be calculated from Africa to Australia. The distance 'd' is 10,500. The distance'd' is not greater than or equal to D(1)(14,300). So, the next distanced DC is selected which is DC(2)(9900). Next the system would find the distance 'd' between DC(2) to DC(3). Thus, the distance 'd' is 7,500 that is also not greater than or equal to D(1)(14,300).

Like that the new DC selection process goes until the 'S' reach round of (n/2). Because the distances from DC(1) to all other DC's were arranged already in descending order. So, the next half set of DC's are closed to the first DF residing DC which is DC(1).

Therefore the next step is to reduce the D(1)'s value by 500 kilometers. Then the same process is going on until the desired DC is found which is at least half the distance far from the DC(1). In this example the round of(5/2) is 3. So D(4) and D(2)'s data centers are only involved in the new DC selection process and their distances from DC(3) is 10,500 and 7,500. So after eight reductions of D(1) by 500 kilometers, it's value is 10,300. Then DC(4) satisfies the condition. So DF2 is stored at DC(4) which is Africa.

Thus the DF, DF1, and DF2 are stored in North America, Australia and Africa respectively which are far from one another.

Table 1. Comparison of Existing Algorithm Vs Proposed algorithms (2RP)

COMPARISON FACTORS	**DCR2S [15]**	**2RP**
Size of the Data Center	1. Super DC 2. Main DC 3. Ordinary DC	Size May vary for elasticity
Configuration of the Data Centre	Each has different Configuration	May have Different Configuration (Based on Service Provider's Specifications)
Cost of the Data Center	(Varies) Super High Cost Main little bit lesser than Super Ordinary low cost	**Same cost**
Numbers of Replication	Depends on Users needs, Based on the service offering by the providers and SLA, etc….	**Only 2 replications inclusive of Original Data file.**
Storage Cost	Vary (Based on SLA and Type of Service)	**Common to all-No variant cost** **(Because 2 Replica and 1 original totally 3 files only stored)**
Maintenance cost	High(Based on SLA and Type of Service)	**Less(Because 2 Replica and 1 original totally 3 files only stored)**
Maintenance Overhead	High	**Low**
Reliability	Doesn't mention strongly	**will be consider to ensure in Future enhancement work**
Available Storage Spaces	Decidable based on the service	**Increased**
Dynamic Replication	Created based on DF Popularity and usage requirements.	**Temporally** Created based on DF Popularity and usage requirements after the usage it will be **discarded**.

5. Conclusion

The proposed Geo-distance based 2-Replica Placing (2RP) algorithms has reduced the replication storage, maintenance cost and increase the storage space availability for the provider's concern as well. The Geo-distance 2RP is used to store only two replications at two different geographical locations that means the replicas are stored far from each other. As the result of these algorithms both the users as well as the providers can be benefited. That is, this algorithms is not only minimizing the cost, but also to increase the available storage spaces and reduce maintenance overheads for the service providers. The providers can provide an efficient Cloud Storage System (CSS) because of this proposed algorithm is proven. In future, this algorithms will be implemented

and tested in alpha testing environment and the main key feature of CSS will be assured through the enhancement of 2RP with availability and reliability concerns forever even the DC's destroyed by the natural disaster.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] N. Mansouri, M. M. Javidi, M.M. and B. M. H. Zade, "Hierarchical data replication strategy to improve performance in cloud computing," *Front. Comput. Sci.* vol. 15, pp. 152-501, 2021.

[2] V. Hadzhiev, "SWOT Analysis of a Hybrid Model for Structuring, Storing and Processing Distributed Data on the Internet," 2021 13th International Conference on Electrical and Electronics Engineering (ELECO), pp. 585-588, 2021.

[3] S. Kianpisheh, M. Kargahi and N. M. Charkari, "Resource Availability Prediction in Distributed Systems: An Approach for Modeling Non-Stationary Transition Probabilities," in IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 8, pp. 2357-2372, 1 Aug. 2017.

[4] A. Mohammad H. A. Haque, Z. Daka, "On Reliability Management of Energy-Aware Real-Time Systems Through Task Replication," IEEE Transactions on Parallel & Distributed Systems, vol. 28, no.3 , pp. 813-825, 2017.

[5] S. Annal Ezhil Selvi and Dr. R. Anbuselvi, "Optimizing the Storage Space and Cost with Reliability Assurance by Replica Reduction on Cloud Storage System", International Journal of Advanced Research in Computer Science (IJARCS),ISSN: 2394-3785,Vol. 8, No. 8, pp. 327-333,2017 (ICI).

[6] Y. Mansouri, A. N. Toosi and R. Buyya "Cost Optimization for Dynamic Replication and Migration of Data in Cloud Data Centers," IEEE Transactions on Cloud Computing, Vol. pp, No. 99, 2017.

[7] A. Aral and T. Ovatman, "A Decentralized Replica Placement Algorithm for Edge Computing," in IEEE Transactions on Network and Service Management, vol. 15, no. 2, pp. 516-529, June 2018.

[8] C. Liu, "A novel replica placement algorithm for minimising communication cost in distributed storage platform," International Journal of Networking and Virtual Organisations, Inderscience Enterprises Ltd, vol. 22(2), pages 147-161, 2020.

[9] Z. Huang, J. Chen, Y. Lin, P. You, and Y. Peng, "Minimizing data redundancy for high reliable cloud storage systems," Computer Networks, 81, 164-177, 2015.

[10] W. Li, Y. Yang, and D. Yuan, "Ensuring cloud data reliability with minimum replication by proactive replica checking," *IEEE Transactions on Computers*, vol. 65, no.5, 1494-1506, 2015.

[11] A. Lazeb, R. Mokadem, and G. Belalem, "Towards a new data replication management in cloud systems," *International Journal of Strategic Information Technology and Applications (IJSITA)*, *10*(2), 1-20, 2019.

[12] D. Gupta, D and D. Singh, "User preference based page ranking algorithm. In *2016 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 166-171), April 2016.

[13] S. Selvi, S and R. Anbuselvi, "Popularity (Hit Rate) Based Replica Creation for Enhancing the Availability in Cloud Storage," *International Journal of Intelligent Engineering & Systems*, *11*(2), 2018.

[14] N. K. Gill, and S. Singh, "Dynamic cost-aware re-replication and rebalancing strategy in cloud system," In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications, Vol. 2* (pp. 39-47). Springer International Publishing, 2015.

[15] C. Li, M. Song, M. Zhang, Y. Luo, "Effective replica management for improving reliability and availability in edge-cloud computing environment," *Journal of Parallel and Distributed Computing*, vol. 143, 107-128, 2020.

[16] S. Annal Ezhil Selvi., "Geo-Distance Based 2-Replica Maintaining Algorithm for Ensuring the Reliability forever Even during the Natural Disaster on Cloud Storage System". *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(7s), 01–07. 2023.

[17] Thomas Wilson, Andrew Evans, Alejandro Perez, Luis Pérez, Juan Martinez. Machine Learning for Anomaly Detection and Outlier Analysis in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/207

[18] Sherje, N. P., Agrawal, S. A., Umbarkar, A. M., Kharche, P. P., & Dhabliya, D. (2021). Machinability study and optimization of CNC drilling process parameters for HSLA steel with coated and uncoated drill bit. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1070