An Analysis of Data Replication Issues and Strategies on Cloud Storage System

S. Annal Ezhil Selvi[1]
Department of Computer Science
Bishop Heber College
Trichy, TamilNadu

Dr. R. Anbuselvi [2]
Department of Computer Science
Bishop Heber College
Trichy, TamilNadu

Abstract: **Recently the number of cloud storage users has increased abundantly. The reason is that the Cloud Storage system reduces the burden of maintenance and it has less storage cost compared with other storage methods. It also provides high availability, reliability and also it is most suitable for high volume of data storage. In order to provide high availability and reliability, the systems introduce redundancy. In replicated systems, the objects are copied several times and each copy residing on different locations in distributed computing. Therefore, the Data Replication is rendering little bit threat about the Cloud Storage System for the users and for the providers it is a big challenge to provide efficient Data Storage. Thus, this paper analyzed the various existing data replication strategies and pointed out the major issues affected by the Data Replication. In future, the direction of this research work is aimed to reduce the number of replication without affecting the availability and reliability of the data.**

Keywords: Cloud Storage, Data Replication, Replication Strategies, Data Availability and Reliability, Storage Cost.

I. INTRODUCTION

Cloud computing is an emerging practice that offers more flexibility in infrastructure and reduces cost than our traditional computing models. Cloud computing offers everything as a service on demand basic through internet. Cloud computing software frameworks manage cloud resources and provide scalable and fault tolerant computing utilities with globally uniform and hardware-transparent user interfaces. The cloud provider takes the responsibility of managing the infrastructural issues.

A. Cloud Computing Architecture

There are 3 blocks involved in cloud computing architecture [3, 4, 6 and 8- 11]. They are,

Service Delivery Model: Sofware as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS), Data as a Service (DaaS), Storage as a Service (StaaS), Network as a Service (NaaS), and Everything as a Service (Xaas).

Service Deployment Model: Services deployed based on different accessing methods like public cloud, community cloud, private and Hybrid cloud.

Cloud Entities: They are providers, Brokers, Reseller, Consumers, Auditors, and Carriers.

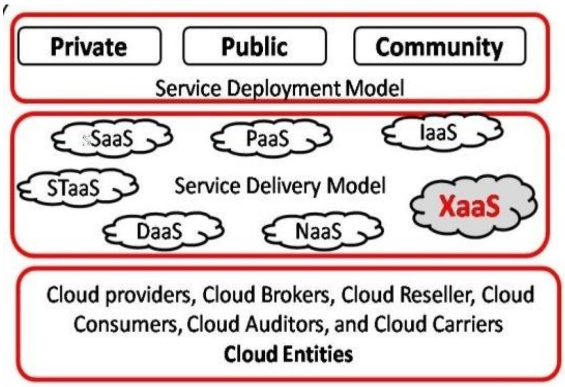


Figure 1: Cloud Architecture

B. Cloud Storage Service

Now-a-days Cloud storage is one of the most needed services and it becoming a popular business paradigm, such as Amazon S3, Elephant Drive, Gigaspaces and small concerns also that offer large Web applications can avoid large capital expenditures in infrastructure by renting distributed storage and pay per use. The storage capacity employed may be large and it should be able to further extent. However, as data scales up, hardware failures in current data centre’s become regular; for example overheating, power (PDU) failures, network issues, hard drive problems, network re-wiring and maintenance. Also, geographic proximity notable affects of data availability; such as in case of a PDU problem 500-1000 machines suddenly disappear, or in case of a rack failure 40-80 machines instantly go down. Furthermore, data may be lost due to natural disasters, such as tornadoes destroying a complete data center, or various attacks [15]. On the other hand, as [7] suggests, Internet availability varies from 95% to 99.6%. Also, the query rates for Web applications data are highly irregular and an application may become temporarily unavailable. So, to avoid these problems and to increase the data availability and Reliability of cloud storage system, data replication has been widely used. Data replication refers to duplicating multiple copies for the data, and these copies are stored in the cloud storage system on different data nodes. When users access the data from one certain node, it will access the replica of the data which is present in the nearest adjacent node, and therefore it can improve the data availability. Similarly, when unwanted events affecting one location where the data resides occur, data can be recovered from another location to provide continued service [17]. There are fault tolerance techniques available that replicates data at different location to tolerate data losses and ensures high reliability of service. Thus, Replication is an important key mechanism to achieve scalability, availability and reliability. But at the same it is a critical task to maintain copies of data as it flows through the network.

The rest of the paper is designed as follows. The literature review is presented in section 2. Various dynamic replicated System models analyzed in section 3. Discussion is done in section 4. Finally, section 5 concludes the work with future scope.

II. LITERATURE REVIEW

Many individuals and organizations are outsourcing their data to the remote cloud service providers (CSP) for reducing the maintenance cost as well as to decrease the burden of large data storage management. The data replication plays a vital role to increase the reliability and availability of the information. At the same time in order to replicate the data, it increases the maintenance overheads for the providers as well as for that backup purpose they charged more fees from the customer to hold the multiple copies over the multiple hosts. But sometimes, they may not keep that copy in all servers. Likewise, the data update the request of the customers also not executed properly.

In [1 and 14] Integrity problems are discussed because the cloud server not trusted since it is accessed through public network such as the internet. And the data integrity is a question mark because they stored in multiple places (distributed storage). Barsoum et al. [2] proposed two Dynamic Multi-Copy Provable Data Possession (DMCPDP) schemes for preventing the CSP from cheating and using less storage by maintaining few copies. Support dynamic operation of data copies (i.e. Modification, insertion, deletion and append).

Prasad et al. [5] discussed the cloud storage system storing of data in a third party cloud system that causes serious concern about data confidentiality and data protection. For that the authors analyzed some of the existing solutions through this article. Finally, they proposed some encryption techniques, but that not enough to protect the data. So the authors proposed the following ways to get solution. First to developing a cloud based business solution for an organization that includes, trusted user verification server that checks the user authentication. Secondly they were planning to design SSAP (Secure Storage Authentication and Privacy) System that is the Re-Encryption Scheme that formulates the secure distributed storage system which was named as Erasure Code-Based Cloud Storage System. Next the Proxy Re-Encryption Scheme, which encrypts the information and forwarding operation over encrypted data. Finally SUV (Straight Unsigned Verification) Scheme that Improved Secrecy ID (ISID) scheme uses address at the receiving end. The advantages of this proposed work is more flexible adjustment between the storage server and authentication user and ISID scheme provides efficiency and provable secure Cloud System.

Vinod Kumar Paidi et al. [12] investigate, how the multi-cloud deployments can reduce security risk and have an impact on usage of the cloud computing technology. So that, in this research article the authors were built a Prototype Application (Architecture) to simulate the advantages of using multi-cloud to improve security, and reduce security Risks. The merits of this proposed solution have improved service availability and reduce the of data loss.

Yanzhi Wang et al. [14] pointed out the recourse allocation problem which is one of the important challenges in cloud computing system in the provider’s side. Especially, when the clients have some service SLAs, the total profit depend on how the system can meet these SLAs. If the providers meet out the SLAs request, then the operation cost will be high and the profit will be low. That is, the operation cost depends on the average service request and the response time. To reduce the request and response meet out time the operation cost will be reduced, then the profit automatically increased. The resource allocation plays the major role to decide the request and response meet out time calculation. So that, the authors suggested optimum recourse allocation by using the following techniques: Hungarian Algorithm for the resource assignment problem and Convex Optimization technique to maximize the total profit for the Service Providers.

III. ANALYSIS OF EXISTING REPLICATION STRATEGIES

There are two major strategies used to obtain a replication system in cloud storage. They are Static mechanism and Dynamic mechanism. In static method of replication the availability and reliability is high. But unwanted use of storage, no flexibility, no scalability and high amount of cost received from the user for storage. In dynamic method the issues are same but the percentage of problem is less compare with static methods. There is number of dynamic heuristic methods used to reduce the percentages.

Priya et al. (2013) [16] discussed the different data replication strategies with Hadoop Distributed File System which provides MapReduce Framework for data replication and consistency maintenance in cloud computing, to accomplish high performance, consistency, availability and partial tolerance and discuss the performance evaluation of these different techniques, frameworks like cloud MapReduce, Integrated data replication, consistency maintenance, MapReduce with Adaptive Load balancing for Heterogeneous and Load imbalanced cluster (MARLA).

Dhananjaya et al. (2013) [17] implemented an automatic replication of data from local host to cloud System. Data replication is implemented by using HADOOP which stores the data at different data centres. If one node goes down then data can be getting from other places seamlessly.

Wenhao et al (2011) [18] designed a novel cost-effective dynamic data replication strategy which facilitates an incremental replication method to reduce the storage cost and meet the data reliability requirement at the same. This strategy works very well especially for data which are only used temporarily and/or have a relatively low reliability requirement. The simulation result shows that replication strategy for reliability can reduce the data storage cost in data centers significantly.

Yaser Mansouri et al. (2013) [19] suggested an algorithm that determines the minimum replication cost of objects such that the expected availability for users is assured. And they also developed an algorithm to optimally select data centers for striped objects such that the expected availability under a given budget is maximized. Navneet et al. (2014) [20] developed an algorithm named as Dynamic Cost-aware Re-replication and Re-balancing Strategy (DCR2S). This algorithm optimizes the cost of replication using the knapsack problem concept and this algorithm is evaluated using CloudSim.

Rajalakshmi et al. (2014) [21] proposed an algorithm for dynamic data replication in cloud. A replication management system allows users to create, register and manage replication and update the replication if the original datasets are modified. This proposed algorithm is suitable for optimal replication selection and placement to increase availability of data in the cloud. Replication method used to increase availability of resources, low access cost and shared bandwidth usage. This approach is based on dynamic replication that adapts replica creation continuously changing network connectivity and users. Their systems developed under the Eucalyptus cloud environment. Cloud storage services often provide key-value store (KVS) functionality for accessing a collection of unstructured data items. Every data set is associated with a key that serves as identifier to access the data set. On the other hand, many existing schemes for replicating data with the goal of enhancing resilience (e.g., based on quorum systems) associate logical timestamps with the stored values, in order to differentiate multiple versions of the same data item.

Cachin et al. (2012) [22] used the agreement number of a shared storage abstraction as a meter for its owner to help the implementation of data replication. It is verified the key-value store (KVS) is a very simple primitive, and it is not different from read/write registers. And that a replica capable of the typical operations on time stamped data is basically more potent than a KVS. Hence, data replication schemes over storage providers with a KVS interface are inherently more complicated to realize than replication schemes over providers with richer interfaces.

John D. Cook et al. (2014) [23] examines the trade-offs in cost and performance between replicated and erasure-encoded storage systems. Data replication placement mechanism had analyzed by Zhang et al. (2013) in [24] and they developed a heuristic algorithm for data replica placement. The simulation evaluates that the algorithm has a better performance whether in a storage sensitive environment or not.

Mohammad et al. (2013) [25] developed a distributed and replicated Dynamic Provable Data Possession (DPDP) which is transparent to clients. It allows for real scheme where the cloud storage provider (CSP) may hide its internal structure from the client, flexibly manage the resources, while still providing provable service to the client. The CSP can decide on how many and which data centre will store the data. Since the load is distributed on multiple data centres, this work monitored one-to-two orders of magnitude better performance in simulation tests, whereas availability and reliability are also improved through replication system. And also, this work use persistent rank-based authenticated skip lists to create centralized and distributed variants of a dynamic version control system with optimal complexity.

IV. DISCUSSIONS AND ANALYSIS

Based on the above analysis this work listed out the following pros and cons of replication system.

A. *Advantages of replicated system on cloud storage:*

High Availability: Data availability is one of the important key factors for cloud storage system. Data replication has been commonly used for increasing the availability in distributed storage

High Reliability: In distributed storage systems, failures of services are treated as reliability issue. It is clear that the reliability of a system will generally increase as the number of replication. Since more replicas will be able to mask more failures.

Minimum Data Loss: If there is any technical problem or inconvenience in provider side the user can loss their information or go for offline. The replication system overcomes the above stated problem.

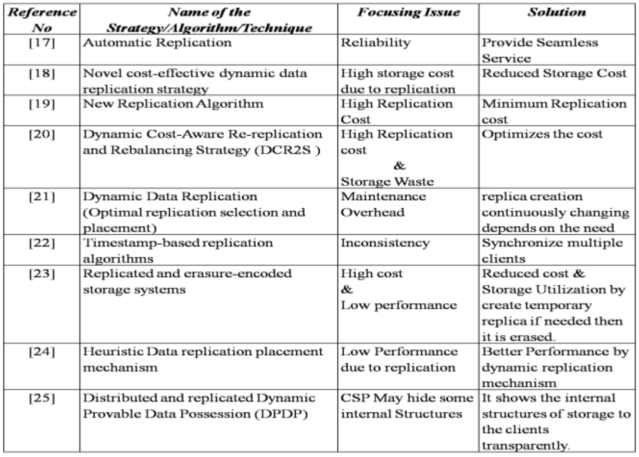
B. *Issues of replicated system on cloud storage:*

In order to provide high availability and reliability the cloud service providers placed that data in different data centers and they put an agreement with the user according to their need. Due to this replication system, the following inconvenience may occur.

a. High Replicated Cost
b. Maintenance overhead
c. Integrity and Consistency

The following table clearly shows the existing replication system strategies, focusing problems and their solution.

Table 1. Analysis of various Replication Strategies



Reference No	*Name of the Strategy/Algorithm/Technique*	*Focusing Issue*	*Solution*
[17]	Automatic Replication	Reliability	Provide Seamless Service
[18]	Novel cost-effective dynamic data replication strategy	High storage cost due to replication	Reduced Storage Cost
[19]	New Replication Algorithm	High Replication Cost	Minimum Replication cost
[20]	Dynamic Cost-Aware Re-replication and Rebalancing Strategy (DCR2S)	High Replication cost & Storage Waste	Optimizes the cost
[21]	Dynamic Data Replication (Optimal replication selection and placement)	Maintenance Overhead	replica creation continuously changing depends on the need
[22]	Timestamp-based replication algorithms	Inconsistency	Synchronize multiple clients
[23]	Replicated and erasure-encoded storage systems	High cost & Low performance	Reduced cost & Storage Utilization by create temporary replica if needed then it is erased.
[24]	Heuristic Data replication placement mechanism	Low Performance due to replication	Better Performance by dynamic replication mechanism
[25]	Distributed and replicated Dynamic Provable Data Possession (DPDP)	CSP May hide some internal Structures	It shows the internal structures of storage to the clients transparently.

V. CONCLUSION AND FUTURE SCOPE

This research article addresses the major issues of replicated systems that are affecting the Cloud storage efficiencies in the distributed computing. This paper pointed out the following issues like, replication cost, maintenance cost, overhead of maintenance, consistency and integrity are the threatening areas of cloud Storage due to replication system. And also this paper analyzed different existing replication strategies. Based on the analysis, this research work concludes that the existing methods are not enough to provide highly efficient cloud storage system. Future direction of this research work is focused to provide the novel dynamic strategy to reduce the number of replication without affecting the availability and reliability of the data.

REFERENCES

[1] Anuradha.R and Dr. Y. Vijayalatha, "A Distributed Storage Integrity Auditing for Secure Cloud Storage Services", International Journal of Advanced Research in Computer Science and Software Engineering, August 2013.

[2] Ayed F. Barsoum and M. Anwar Hasan, "On Verifying Dynamic Multiple Data Copies Over Cloud Servers",August 15, 2011.

[3] Gurudatt Kulkarni, Ramesh Sutar and Jayant Gambhir", Cloud Computing – Storage as a Serivice", International Journal of Engineering Research and Applications (IJERA), Feb-2012.

[4] Gurudatt Kulkarni, Rani Waghmare, Rajinkant Palwe, "Cloud Storage Architecture", International Journal of Engineering Research and Applications, 2012.

[5] Prasad.T, Aravindhu.N and Rampriya.D, "SSAP: Augmentation based Secure Storage Authentication and Privacy System in Cloud Environment", International Journal of Cloud Computing and Services Science, June 2013.

[6] Poonam devi, Tirlok Gaba,"Cloud Computing", International Journal of Advanced Research in computer Science and Software Engeneering,Volume 3, Issue 5, ISSN: 2277 128X, May 2013.

[7] Priyanka Ora, P.R. Pal,"Security: A Major Concern in Cloud Computing",International Journal of Advanced Research in computer Science and Software Engeneering, Volume 3, Issues 7, ISSN 2277 128X, July 2013.

[8] Sheih Mahbub Habib, Sascha Hauke, Sebastian Ries and Max Muhlhauser, "Trust a facilitator in cloud computing: a survey", Journal of cloud computing, Springer open journal, 2012.

[9] Sunitha and Prachi, "Introduction to Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering, May 2013.

[10] Subashini.S and V. Kavitha, "A Metadata Based Storage Model for Securing Data in Cloud Environment",Ameerican Journal of Applied Sciences9(9) , ISSN: 1546-9239, 2012.

[11] Talasila Sasidhar, Pavan Kumar Illa and Subrahmanyam Kodukula, "A Generalized Cloud Storage Architecture with Backup Technology for any Cloud Storage Providers", Issue 2, Volume 2, ISSN 2250-1797, April 2012.

[12] Vinod Kumar Paidi and P.Varaprasad Rao, "Multi-Cloud Architecture to Reduce Security Risks in Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering, August 2013.

[13] Venkateswara Rao. N and SK. MeeraSaheb, "A survey of Cloud Computing: Cloud Computing Concerns and Issues", International Journal of Engineering Research and Technology, Volume 2, Issue 4, April 2013.

[14] Yanzhi Wang, Shuang chen and Massoud Pedram, " Service Level Agreement –based Joint Application Environment Assignment and Resource Allocation in Cloud Computing System", IEEE Green Technologies Conference, 2013.

[15] Nicolas Bonvin, Thanasis G. Papaioannou and Karl Aberer "A Self-Organized, Fault-Tolerant and Scalable Replication Scheme for Cloud Storage", SoCC'10, Indianapolis, Indiana, USA, June 10–11, 2010.

[16] Priya Deshpande, Aniket Bhaise, Prasanna Joeg ,"A Comparative analysis of Data Replication Strategies and Consistency Maintenance in Distributed File Systems", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

[17] Dhananjaya Gupt, Mrs.Anju Bala, "Autonomic Data Replication in Cloud Environment", International Journal of Electronics and Computer Science Engineering, ISSN 2277-1956/V2N2-459-464, Volume2, Number 2, 2013.

[18] Wenhao LI, Yun Yang and Dong Yuan, "A Novel Cost-effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centers", IEEE International Conference, 2011.

[19] Yaser Mansouri, Adel Nadjaran Toosi and Rajkumar Buyya, "Brokering Algorithms for Optimizing the Availability and Cost of Cloud Storage Services", Cloud Computing Technology and science (CloudCom), IEEE 5th International Conference, Volume: 1, 2-5 Dec. 2013,pp 581 – 589.

[20] Navneet Kaur Gill and Sarbjeet Singh, "Dynamic Cost-Aware Re-replication and Rebalancing Strategy in Cloud System" © Springer International Publishing Switzerland 2015 S.C. Satapathy et al. (eds.), Proc. of the 3rd Int. Conf. on Front. of Intell. Comput. (FICTA) 2014 – Vol. 2, Advances in Intelligent Systems and Computing 328, DOI: 10.1007/978-3-319-12012-6_5.

[21] A.Rajalakshmi, D.Vijayakumar, Dr. K .G. Srinivasagan, "An Improved Dynamic Data Replica Selection and Placement in Hybrid Cloud" International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.

[22] Christian Cachin, Birgit Junker and Alessandro Sorniotti ,"On Limitations of Using Cloud Storage for Data Replication" Proc. 6th Workshop on Recent Advances on Intrusion Tolerance and reSilience , WRAITS 2012," Boston, MA (IEEE), June 2012.

[23] John D. Cook , Robert Primmer and Ab de Kwant, "Compare Cost and Performance of Replication and Erasure Coding", WHITE PAPER , Hitachi Review Vol. 63, July 2014.

[24] ZHANG Tao, "Data Replica Placement in Cloud Storage System", International Workshop on Cloud Computing and Information Security (CCIS), 2013.

[25] Mohammad Etemad and Alptekin, "Transparent, Distributed, and Replicated Dynamic Provable Data Possession", April 15, 2013.