International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 128 – 133

A Novel Reliability Management Technique by Reduced Replication in Cloud Storage Applications

[1]C. Afzal, [2]M. Atheequllah Khan,[3] M. Sudhakar
[1] PG Scholar, [2] Associate Professor
[1,2]SSITS, JNTUA University.
[3] Research Associate, VIT, Chennai Campus
[1]*Afzal.ck2@gmail.com*
[2]*atheequllah@gmail.com*
[3]*mallasudhakar.cse@gmail.com*

Abstract :- In the recent years, cloud computing has become popular among various fields like information technology and various business enterprises. The fundamental use of the cloud now a day is for storing information and sharing the resources. Cloud is another way to store the expensive measure of information. Cloud provides the storage space and offering to use this information to various clients. Additionally, it is a technique for pay according to we utilize. The two principle concerns in current cloud storage systems are providing the reliable data and storage cost. To ensure data reliability, current cloud systems uses multi-replica strategy (typically three replicas), which incurs a huge storage space, on the effect it leads to more storage cost for data-intensive applications in the cloud. To minimize the storage space, in this paper we proposed a cost effective data reliability mechanism called Proactive Replica Checking for Reliability (PRCR) by using a generalized data reliability model. PRCR guarantees the reliability with the reduced replication factor, which can likewise reduce the storage cost for replication based methodologies. Contrasting to the traditional three – replica strategy, PRCR can diminish the storage space utilization by one-third to two-thirds of the current storage space, thus fundamentally bringing down the storage cost.

Keywords – *Reliability Management, Proactive Replica Checking, Data generalizability model.*

*****

I. INTRODUCTION

In recent years, cloud computing has turned out to be prominent among various fields like information technology and different business ventures because the clients need to pay just as indicated by their utilization of storage. Cloud makes the utilization of cloud specialist organization which provides different services to their users according to their need. Cloud computing has turned out the popular in both scholarly community and industry. Because of the quick advance of innovation and large scale utilization of web has brought about the era of enormous measure of information and cloud storage is growing at dramatic speed.

The data which are stored in the cloud ought to initially me the reliability criteria with high efficiency and cost effectiveness. High reliability is essential for the typical cloud systems. Be that as it may, accomplish the reliability with a sensible cost can be challenging one, particularly in large scale systems.

The size of the cloud storages is growing at a dramatic speed. Under the analysis, the data storage in the cloud almost reached 1 ZB, while more data are stored or processed in their journey. Interim, with the development of the cloud computing paradigm, cloud-based applications have advanced their demand the cloud storage [1]. While the necessity of the data reliability should be met in the first place and the data in the cloud should be put away with higher cost-viability.

Data reliability is the main concern in the cloud and it is defined as the likelihood of one data item available in the cloud system for a specific period. It is the critical issue in the cloud storage systems, which shows the capacity of keeping the data consistently and accessible fully by the client/user. Because of the quick growth in the cloud data, providing the data reliably has turned in to a challenging task. At present, data replication is a standard method for providing the data reliability. Nonetheless, the consumption of the storage space with the help of replication methodology brings about the immense cost, and it is trusted that cost would be passed on to the clients in the end as a result.

In this paper, the authors present a new cost-effective data reliability management mechanism based on the proactive replica checking called PRCR for decreasing the storage space consumption, hence subsequently reducing the storage cost for data-intensive applications in the cloud systems. The main goal of PRCR algorithm is to decrease the number of replicas stored in the cloud while meeting the data reliability requirement. Contrasted with the traditional data reliability mechanisms in the cloud application, the PRCR algorithm contains the following features:

1. As the First level, the PRCR algorithm is guaranteed to provide the data reliability.
2. PRCR is capable of providing data reliability management in a cost-effective manner. By applying PRCR, data reliability is guaranteed with at most two replicas stored in the cloud.

By applying benchmarks from Amazon web services, we assess the efficiency of PRCR nodes and simulate the reliability management process utilizing PRCR with the data created by data-intensive storage applications. The outcomes are contrasted with the traditional three-replica mechanism. It is observed that the PRCR mechanism can reduce the storage space from 33% to 66% of the storage space than the existing methods used in the cloud, subsequently, it also decrease the storage cost also.

The rest of the paper is organized as follows. The related works on data reliability, data replication and cost-effective data storage are addressed in Section II. The concepts proposed on the mechanism Proactive Replica Checking for Reliability (PRCR) and its high level design are presented in Section III. The working procedure of PRCR is described in Section IV. Results and Discussions are discussed in Section V. Finally, conclusions and future work are summarized in Section VI.

II. LITERATURE SURVEY

There is a huge research is going on to reduce the replication factor and many works have already been done in the past. Some of the existing works are illustrated here. S. Ramabhadran et.al. [6]. have studied and assumed that the failure of each disk is constant in the exponential data reliability model. For example, the existing studies analyze that the Markov chain model assumes that the failure rate of each disk in the cloud or any other storage system is the same. In reality, the constant disk failure rate cannot explain the overall phenomena. It has been observed the mostly the failure rate of the hard disk drives follows a "bath tub" curve model as shown in figure 1. In the figure, it is shown that the failure rate is higher in the disk early life, and drops during some period, and remains constant for the remainder of the disk life span and increase again at the end of the disk's lifetime.

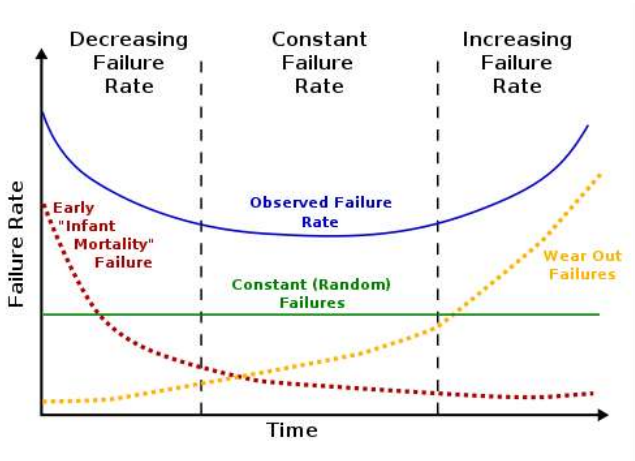

Figure 1. Bathtub Curve for Disk Failure rate

The disk failure probability does not follow an exponential distribution. By fixing this inconsistency IDEMA viz. International Disk Drive Equipment and Materials Association proposed a satisfactory presentation for disk failure rates that, the life span of each disk are divided into various life stages with various failure rates. [10]. In a nine-month investigation Google also gets the results inconsistent with this model. This paper concentrates the disk failure rates with IDEMA style [11], in which the lifespan of each disk is varied accordingly. Many efforts and research are performed for providing the data reliability as within the case of software aspect also.

A.Gharaibeh, B.Balasubramanian[4] studied and analyzed that Data replication is considered as the dominant approach in the cloud storage system for providing the data reliability and recent works on the large-scale cloud storage systems followed the multi-replication strategy to ensure the data reliability. In the field of cloud computing, data replication has been the widely adopted technique in the commercial distributed storage systems. Amazon S3, Google File System (GFS) [11], HDFS are the typical examples that are following three replica strategy.

Although the data replication strategy is used by many commercial cloud storage service providers, Amazon S3 proposed Reduced Redundancy Storage (RRS) to address the data redundancy issue to decrease the storage consumption [10]. The problem with RRS is it sacrifices the data reliability. With the help of low level of reliability only provided. An erasure coded storage system also presented and implemented by K.V Rashmi [3]. In the approach, the data is divided into various blocks and store them with additional erasure coding blocks. By using this kind of technique reliability is assured but there is a computation overhead in encoding and decoding the data. Hence, In the case of data intensive applications,

erasure coding methods are not the best solution compared with the replication-based mechanisms.

III. PROPOSED WORK

The PRCR runs on the virtual machines (VM) in the cloud environment. This VM is responsible for running the user interface, PRCR node and to conduct the proactive replication checking separately. The main responsibility of the user interface is to determine the minimum replication factor. It additionally makes the duplicate (replicas) if necessary and distributes the metadata of records. At whatever point the first replica of the file is created or uploaded in the cloud stage it finds the minimum number of replica depending on the storage duration that is either short period or long period of existence of the file in the cloud. With this principle, it makes the replication either one or two. If one replica is not sufficient for one particular file to store the data then the user interface calls the cloud to create the second replica to the corresponding file. When the second replication of that file is created, it also creates one metadata file and this will be distributed to the suitable PRCR node. The metadata has added up to six sorts of attributes such as file Identifier, time stamp, data reliability required, expected storage duration, checking interval and the address of the replica. The record of file ID and the replica address is automatically given when the first or second copy of the files are made.

Proactive Replica Checking for Reliability (PRCR)

The principle thought to use proactive replica checking is to propose a cost-effective data reliability management mechanism for cloud data storage. The PRCR utilize the well-known property of exponential distribution called the memory-less property. In PRCR the information in the cloud is organized into various types according to its expected storage time and reliability management. For those data items, which are required for short term storage are sufficient to maintain a single copy to provide the data reliability. For those data items, which are required for long term or regularly used, maintains three replica strategy for providing the reliability. Consider an example, there are 500 files each of 1 Mb size. Typically, with the conventional strategy we should need to maintain three replicas for each type, it requires totally 1500 MB storage space (500 * 3 copies of each), results from a huge storage cost. As with two kinds of data types, assume there are 300 files among 500 are critical and 200 files are for the short term use. If we consider this scenario 300 * 3+ 200 * 1 = 1100 MB of storage space required, which will save 400 MB compared with the conventional three-replica strategy. The proactive replica checking is done at regular intervals to check the presence of the replicas. This task must be done before the reliability assurances drop below the reliability requirement. If any single replica is missed in the observation it should quickly recover under a strategy. For example, replica maintenance for distributed cloud storage system. If in some cases both the replicas may be lost, the probability of this situation is incorporated in the data reliability model in its earliest. PRCR ensures that the data loss rate is maximum of 0.01 percent of the total data per year

PRCR keeps running on the top virtual layer of the cloud and overseas data which are stored as the file in various data centers. Figure 2. Demonstrates the architecture of the Proactive Replica Checking for Reliability (PRCR). There are two noteworthy parts in the architecture are PRCR node and UI (User Interface)

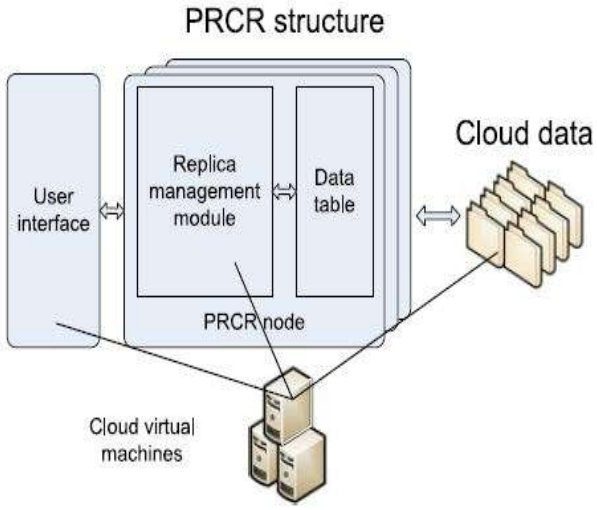

Figure 2: PRCR Architecture

PRCR node: The main segment in the PRCR architectures is the PRCR node. The main responsibility of this node is the management of the replicas. As will be specified later, as per the number of files in the cloud, PRCR may contain at least one or more PRCR nodes and each one is independent together. The PRCR node is made of two noteworthy components inside. One is data table and the second is replica management module.

i . Data Table: It keeps track the metadata of all data that each PRCR node manages. For each file in the data storage, four metadata attributes are maintained in the data table. They are File Identifier (file ID), scan interval, a timestamp value (time stamp) address of the replica (replica address). File ID is used to uniquely identify the file. The scan interval is the time between the two replica checking of each file. Timestamp records the time whenever the replica checking is done on one

file. PRCR can decide the files that need to be checked depending on the scan interval and timestamp and based on the replica address all the replicas are found. Here each round of the scan is considered as a scan cycle. In each scan cycle, metadata is sequentially examined once.

ii. Replica Management Module: It is the control component of the PRCR node, which is responsible for managing the metadata in the data table and cooperating with the Cloud computing instance2 to process the checking tasks. In each scan cycle, the replica management module scans the metadata in the data table and determines whether the file needs to be checked. For files that need to be checked, the replica management module extracts their metadata from the data table and sends these metadata to the Cloud computing instances. After the Cloud computing instances have finished the checking tasks, the replica management module receives the returned results and conducts further actions accordingly. In particular, if any replica is not available, the replica management module sends a request for replication to the Cloud which creates a new replica of the data and updates the data table accordingly.

User interface: It is a very important component of PRCR, which is responsible for justifying reliability management requests and distributing accepted requests to different PRCR nodes. A reliability management request is a request for PRCR to manage the file. In the request, the metadata of the file is required. Despite the four types of metadata attributes in the data table, a metadata attribute which is called the expected storage duration is needed. It is an optional metadata attribute of the file, which indicates the storage duration that the user expects. According to the expected storage duration, the user interface is able to justify whether the file needs to be managed by PRCR or not. In particular, if such management is unnecessary, the reliability management request is declined and the file is stored with only one replica in the Cloud.

IV. WORKING PROCEDURE OF PRCR

We illustrate the working process of PRCR in the Cloud by following the life cycle of a file.

1. When the original file of the replica is created, the actual process begins at this time. Based on the issues such as disk failure rate, storage space, data reliability, the user interface will determine to store the file with one copy or two copies (replicas).
2. According to the calculation in the user interface, if one replica cannot satisfy the data reliability and storage duration requirements of the file, the user interface creates a second replica by calling Cloud services and calculates the checking interval(s) of the file. Its metadata is then distributed to the appropriate PRCR node (2). If one replica is sufficient, only the original replica is stored and the metadata of the file is not created (9).
3. The attributes related to metadata are stored in the data table of the PRCR node.
4. Metadata is scanned periodically according to the scan cycle of the PRCR node. According to file's time stamp and the current checking interval, PRCR determines whether proactive replica checking is needed.
5. If proactive replica checking is needed, the replica management module obtains the metadata of the file from the data table.
6. The replica management module assigns the proactive replica checking task to one of the Cloud virtual machines for proactive replica checking. The Cloud virtual machine executes the task, in which both replicas of the file are checked.
7. The Cloud virtual machine conducts further action according to the result of the proactive replica checking task: if both replicas are alive or lost, go to step 8; if only one replica is lost, the virtual machine calls the Cloud services to generate a new replica based on the replica that is alive.
8. The Cloud virtual machine returns the result of the proactive replica checking task, while in the data table, the time stamp and checking interval(s) are updated. Specifically, step (1) if both replicas are not lost, the next checking interval is put forward as the current checking interval; and step(2) if a replica is lost and recovered on a new disk, the new replica address is stored and all the checking interval(s) are recalculated. Otherwise, further steps could be conducted, for example, a data loss alert could be issued.
9. Steps 4 to 8 form a continuous loop until the expected storage duration is reached or the file is deleted. If the expected storage duration is reached, either the storage user could renew the PRCR service or PRCR could delete the metadata of the file and stop the proactive replica checking process.

V. RESULTS & DISCUSSION

The work is summarized in three steps. As a first step, the registration process of the user will be done with the credentials. Secondly. If the user is authenticated, the User Interface will be able to create a file. The files which are critical will be replicated as two times and the address of the

replica will be noted, the other files can be replicated only once. The PRCR node is able to monitor the number of replicas as well as when there is a request from the user from the user interface for the additional replicas it will be considered and a replica is created and at the same time the metadata created and stored in the corresponding file. Thirdly, the generalized reliability model implemented for providing the reliability. The implementation divided into three modules. One is user interface design, second Proactive replica checking and finally storage prediction is implemented and the snapshots of a user interface and the categorization of relevant and irrelevant data shown in Figure 3 and Figure 4.

Figure 3: User Interface for taking Input credentials.

Figure 4: Categories of data to replicate file

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented mainly three contributions. First, a generalized data reliability with multiple replicas is proposed. We have implemented the Proactive Replica Checking for Reliability mechanism by considering the reliability management as a first requirement. Based on that demand, the replication copies will take place according to the request from the user interface, i.e. whether for a specific data file, either one replica or two replicas are stored. The results show that this method works well and give the better results compared to the existing methods. It reduces the storage space, in turn, reduce the storage cost which is essential for the current cloud storage applications. The performance considerations will be considered as our future work.

REFERENCES

[1] Wenhao Li, et.al.,” Ensuring Cloud Data Reliability with Minimum Replication by Proactive Replica Checking” in IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 5, MAY 2016

[2] W. Li, Y. Yang, J. Chen, and D. Yuan,”A cost-effective mechanism for cloud data reliability management based on proactive replica checking," in Proc. Int. Symp. Cluster, Cloud Grid Comput., 2012, pp. 564-571.

[3] K. V. Rashmi, N. B. Shah, D. Gu, H. Kuang, D. Borthakur, and K.Ramchandran, A Hitchhikers guide to fast and efficient data reconstruction in erasure coded data centers, in Proc. Conf.SIGCOMM, 2014, pp. 331342.

[4] W. Li, Y. Yang, and D. Yuan. 2013, Oct. 11. An energy-efficient data transfer strategy with link rate control for Cloud, Int. J. Auton. Adaptive Commun. Syst. [Online]. Available: http://www.ict.swin.edu.au/personal/yyang/papers/IJAACS-Li.pdf

[5] D. Yuan, Y. Yang, X. Liu, W. Li, L. Cui, M. Xu, and J. Chen, "A highly practical approach towards achieving minimum datasets storage cost in the cloud," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1234-1244, Jun. 2013.

[6] B. Balasubramanian and V. Garg, Fault tolerance in distributed systems using fused data structures, IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 4, pp. 701715, Apr. 2013.

[7] W. Li, Y. Yang, J. Chen, and D. Yuan, "A cost-effective mechanism for cloud data reliability management based on proactive replica checking," in Proc. Int. Symp. Cluster, Cloud Grid Comput., 2012, pp. 564-571.

[8] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li and S.Yekhanin, "Erasure Coding in Windows Azure storage,” in Proc. USENIX Annu. Techn. Conf., 2012, pp. 2-13.

[9] W. Li, Y. Yang, and D. Yuan,”A novel cost-effective dynamic data replication strategy for reliability in cloud

data centers," in Proc. Int.Conf. Cloud Green Comput., 2011, pp. 496-502.
[10] Amazon. (2011). Amazon Simple Storage Service (AmazonS3).[Online]. Available:http://aws.amazon.com/s3/, 2011.
[11] S. Ghemawat, H. Gobioff, and S. Leung, "The Google File system,"In Proc. ACM Symp. Oper. Syst. Principles, 2003, pp. 29-43. [13] IDEMA, "R2-98: Specification of hard disk drive reliability", IDEMA Standards,1998.
[12] R. Bachwani, L. Gryz, R. Bianchini, and C. Dubnicki,"Dynamically quantifying and improving the reliability of distributed storage systems," in Proc. IEEE Symp. Rel. Distrib. Syst., 2008, pp. 85- 94.
[13] D. Yuan, Y. Yang, X. Liu, W. Li, L. Cui, M. Xu, and J. Chen, "Ahighly practical approach towards achieving minimum datasetsstorage cost in the cloud," IEEE Trans. Parallel Distrib. Syst.,vol. 24, no. 6, pp. 1234–1244, Jun. 2013
[14] D. Borthakur. (2007). TheHadoop Distributed File System: Architectureand Design [Online]. Availablehttp://hadoop.apache.org/common/docs/r0.18.3/hdfs_design.html

Author's Profile:

C Afzal is currently a PG scholar in the school of Computer science and Engineering in SSITS, JNTUA University. He received his Bachelor's degree in computer science and engineering in the year of 2015 from JNTUA University. I am very interested to learn the new technologies and to do various research works and other activities. I have attended various national and international conferences and workshops. My research interests are cloud computing and analytics on the big data.

M Atheequllah Khan is currently working as Associate Professor at the school of Computer Science and Engineering in SSITS, JNTUA. He received his master's degree from JNTUA in 2011 and Bachelor's degree in 2006. He has 10 years of teaching experience and taught various subjects in computer science stream and organized various national conferences and workshops in the organization. His research interests are cloud computing, computer vision and knowledge engineering.

M.Sudhakar, currently working as Research Associate at VIT University, Chennai Campus. He received his master's degree at the school of SCSE from JNTUA University in the year 2012. He received his Bachelor's degree in computer science and engineering in 2010. He has published various national and international journals. His research interests are image analytics, and image enhancement and recommender systems,