IOP Conf. Series: Materials Science and Engineering **263** (2017) 042063 doi:10.1088/1757-899X/263/4/042063

Proactive replica checking to assure reliability of data in cloud storage with minimum replication

Damini Murarka, G Uma Maheswari

Vellore Institute of Technology University, Vellore-632014

g.uma maheswari@vit.ac.in

Abstract. The two major issues for cloud storage systems are data reliability and storage costs. For data reliability protection, multi-replica replication strategy which is used mostly in current clouds acquires huge storage consumption, leading to a large storage cost for applications within the loud specifically. This paper presents a cost-efficient data reliability mechanism named PRCR to cut back the cloud storage consumption. PRCR ensures data reliability of large cloud information with the replication that might conjointly function as a price effective benchmark for replication. The duplication shows that when resembled to the standard three-replica approach, PRCR will scale back to consume only a simple fraction of the cloud storage from one-third of the storage, thence considerably minimizing the cloud storage price.

1. Introduction

The cloud storage size is increasing at a very high pace. By 2015, it was calculated that the information held on within the cloud reached 0.8 ZB. Meantime, when the cloud-computing paradigm arrived, a huge amount of cloud storage was demanded by the cloud based applications. An extremely cost-efficient manner is needed to store the knowledge and information within the cloud. The research done in this paper focuses on decreasing storage of cloud consumption by reducing data replication and also maintaining information reliability requirements. A cost-efficacious data reliability management mechanism called the proactive replica checking for reliability (PRCR) is presented to minimize the storage consumption in cloud by reducing the number of replicas.

PRCR has the following features:

a. It is able to ascertain the data reliability of storage contrivances with variable disk failure rates.
b. It is able to manage immensely colossal amounts of data in the cloud with a negligible running cost.
c. It provides data reliability management in a highly cost-efficacious way.

By applying PRCR, a wide range of data reliability assurance can be provided with the minimum number of replicas, which is no more than two.

2. Proposed Technique

Proactive Replica Checking for Reliability (PRCR) technique can be used to store data in the cloud with minimum number of replication, at the same time meet the reliability of data in a cost efficacious manner. This PRCR technique insures data-reliability of the data stored with minimum number of replicas of data which further results in less cost for data replication technique. In this paper, the analysis is to minimize the consumption of cloud storage by reducing data replication with maintained

information reliability demand. This research presents a cost-efficient data reliability mechanism called as PRCR (Proactive Replica Checking for Reliability) to cut back consumption of cloud storage. When contradicted to the three-replica strategy, PRCR reduces the consumption of storage space from one-third to two-third, eventually lowering the cost for storage.

3. System Architecture

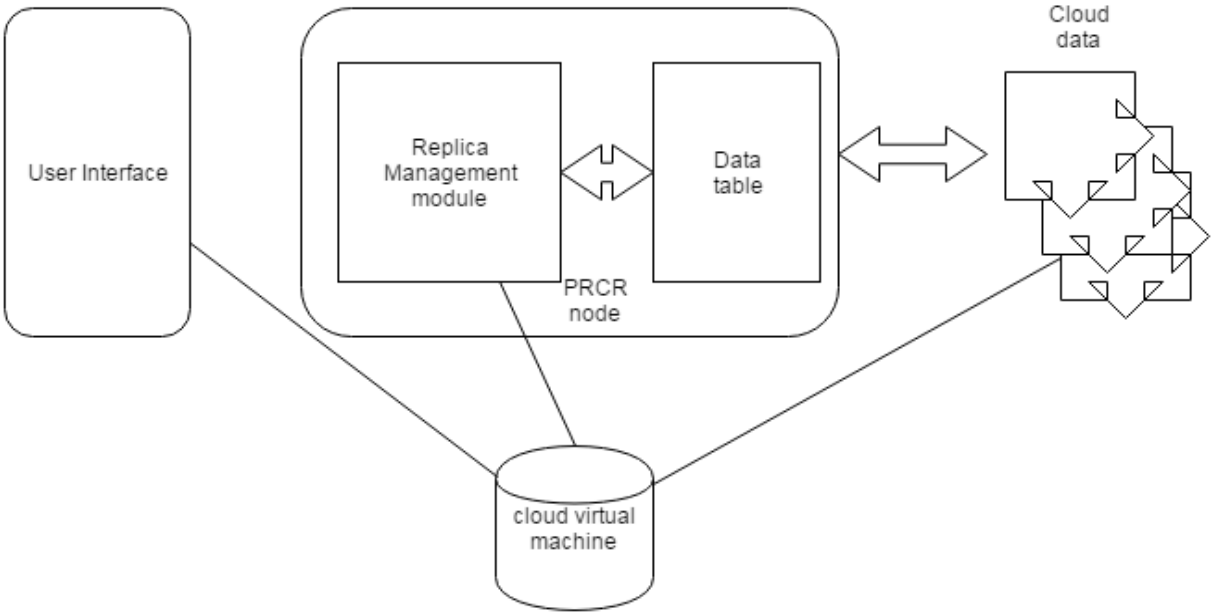

Figure 1 Architecture of PRCR

PRCR uses the reliability requirements and the expected duration for storage in order to manage the data stored in cloud. Single replica is stored in the cloud for data that is to be stored for a short period of time or the data which has a data reliability requirement of only one replica. Two replicas are required for the data which has higher storage duration or a reliability requirement which cannot be fulfilled by a single replica. PRCR can be used to store data in the cloud with minimum number of replication, at the same time meet the reliability of data in a cost effective manner. PRCR is a service given by the Cloud storage providers which acts as a data reliability management technique. Cloud virtual machines are used to run PRCR nodes, user interface and to conduct proactive replica checking.

3.1. Overview of Architecture

3.1.1. User Interface

This component of the architecture is used to create replicas, know the least number of replicas and also to determine metadata of the files. This component gives the least number of replicas as soon as the file is uploaded by the user. The User interface component informs the cloud service to make replicas of the stored data. PRCR node gets the metadata of file with the help of this module.

3.1.2. PRCR Node

This core component of the system manages the replicas and metadata. The PRCR is made up of an user interface and 'n' number of PRCR nodes. These nodes are independent and hence make it easier to create and destroy any number of nodes as per the requirement. The two sub-components of this module are:

a) Data Table, and
b) Replica Management Module

3.1.3. Data Table

Data Table component maintains the metadata attributes of the files managed by PRCR node. The replica management module regularly scans all meta-data to confirm information reliability of files.

3.2. Replica Management Module

This module helps to process the replica checking work by co-operating with virtual machines and also to scan the metadata that is stored in the data table. This module filters the metadata that is stored in data table and decides if document should be checked. On the off chance that a document should be checked, this module acquires the metadata stored in data table and forwards it to a virtual machine for checking. After the checking is done, this module leads additionally activities as indicated by the returned result. Specifically, if any copy is lost, it introduces the recovery procedure for making another replica.

3.3. System Techniques

3.3.1. Proactive Replica Checking

PRCR uses the reliability requirements and the expected duration for storage in order to manage the data stored in cloud. Single replica is stored in the cloud for data that is to be stored for a short period of time or the data which has a data reliability requirement of only one replica. Two replicas are required for the data which has higher storage duration or a reliability requirement which cannot be fulfilled by a single replica.

3.4. Replica Management Module

This module helps to process the replica checking work by co-operating with virtual machines and also to scan the metadata that is stored in the data table.

3.5. Data Reliability Model

Rather than the ordinary three-replica technique, there is an option by which we can give information reliability quality less replicas. A scientific model for information dependability gives the likelihood of decreasing the quantity of replicas while meeting information dependability necessity.

4. LITERATURE REVIEW

In the paper[1], "A cost-effective mechanism for cloud data reliability management based on Proactive Replica Checking", the authors 'Wenhao Li, Yun Yang, Jinjun Chen, Dong Yuan' have proposed that in current Cloud processing situations, administration of information unwavering quality has turned into a test. For information concentrated logical applications, putting away information in the Cloud with the normal 3-copy replication technique for dealing with the information unwavering quality would cause gigantic capacity price. To claim this problem, this paper exhibits a practical information unwavering quality administration component called PRCR that proactively analyses the accessibility of reproductions for keeping up information dependability. Our reproduction shows that, contrasting and the regular 3-imitation replication methodology, PRCR can lessen the storage room utilization by 33% to 66%, consequently decrease the capacity cost altogether in the Cloud.

In the paper[2]," A Novel Cost-effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centers" the authors, 'Li, Wenhao, Yun Yang, and Dong Yuan' have put forward an approach called the Cost-effective Incremental Replication (CIR) strategy which was a cost effective replication approach. When the data to be stored are of lower reliability or are to be stored temporarily, the CIR approach significantly reduces the storage cost. Unlike most approaches, this approach considers the storage cost to have the highest priority. However, storage aging process and dealing with both performance and cost in cloud storage are considered to be some major issues which have to be worked on in future.

In the paper[3], "Dynamically Quantifying and Improving the Reliability of Distributed Storage Systems", the authors, 'Rekha Bachwani, Leszek Gryz, Ricardo Bianchini, and Cezary Dubnicki' contend reliability quality extensive capacity frameworks can be altogether enhanced through utilizing good reliability quality measurements with more effective approaches for recuperating from equipment disappointments. In particular, the author makes three fundamental commitments. To begin with, the author presents NDS (Normalcy Deviation Score), for powerfully evaluating reliability quality of a capacity framework. Also, the author propose Scaled down Minimum Intersection), a novel recuperation planning strategy that enhances dependability by effectively remaking information after an equipment disappointment. At last, the authors assess NDS and MinI for three normal information assignment plans and various diverse parameters. The assessment concentrates on an appropriated stockpiling framework in light of deletion codes. It is located that MinI enhances reliability quality fundamentally, when contrasted with customary strategies.

In the paper[4] "A Cost-Effective Strategy for storing scientific datasets with multiple service providers in the cloud", the authors "Dong Yuan, Lizhen Cui, Xiao Liu, Erjiang Fu, Yun Yang", propose that cloud computing gives researchers stage which conveys calculation in addition to information escalated applications apart from framework venture. With exorbitant cloud assets or a choice emotionally supportive network, extensive created data sets could be adaptably 1) put away in the present cloud locally, 2) erased or re-produced at whatever point reused or 3) exchanged to less expensive cloud benefit for capacity. In this paper, a novel system is proposed that can cost store extensive created datasets with different cloud specialist co-ops.

In the paper[5] “Fault Tolerance in Distributed Systems using Fused Data Structures”, the authors ‘Bharath K. Balasubramanan, Vijay Goyal’ depicts method to endure blames in extensive information structures facilitated on appropriated servers, in light of the idea of melded reinforcements. The predominant answer for this issue is replication. We display an answer, alluded to as combination which has blended with eradication ciphers or specific duplication with endures crash issues utilizing extra melded reinforcements. To represent the down to earth convenience of combination, we utilize intertwined reinforcements for dependability with Amazon's very accessible esteem keeps, Dynam. On the contrary to the present duplication arrangement utilizes 330 reinforcement frameworks, it exhibit an answer which exclusive needs 122 reinforcement frameworks. The outcome with investment funds at area and additionally different assets, for example, control.

In the paper [6] “Disk Infant Mortality in Large Storage Systems” the authors ‘Qin Xin1Thomas J. E. Schwarz, S. J. Ethan L. Miller1’ depicts the impact of baby mortality haul disappointment amounts of frameworks which could protect these information that a considerable length of time. Our disappointment models fuse the notable "bath bend," which mirrors the higher disappointment amounts with recent circle storages, a smaller, consistent disappointment amount amid to rest with outline traverse, expanded disappointment amounts according to parts destroy. Expansive frameworks are powerless against the "accomplice impact" that happens when many plates are at the same time supplanted with recent circles. Many precise plate structures or re-enactments had expectations with framework times which are the sceptical to previous structure which accept consistent circle disappointment amount. Along these lines, bigger framework scale obliges planners to consider plate new-born child mortality.

5. FUTURE ENHANCMENTS

For future enhancements, this work can be continued in two ways. Initially, for more optimization, a more precise design of PRCR can be introduced. Also, future research needs to be done on the performance of cloud data access with location of data as PRCR minimizes the number of replicas on the cloud.PRCR mechanism manages a large amount of data and information in Cloud and it reduces the consumption of cloud storage space at a minor cost.

6. FUTURE TECHNIQUE

The technique that can be used in future is Private Information Retrieval. The databases which are accessible publicly are a key resource to retrieve up-to-date information. Although this is a threat to the privacy of user, the user's queries can be followed and the actions can be assumed for fraudulent purposes. The cases in which the user's information is to be kept are a secret, the database files are accessed. It can be checked whether a better solution can be obtained to retrieve the private information of the user by replicating the database. For this, a scheme can be described that may allow k-replicated-copies of the database to be accessed by users.

7. CONCLUSION

This paper introduces a cost-effective data-reliability-management system, i.e., PRCR technique. This PRCR technique assure data-reliability of the data stored in cloud with minimum replication of data which further results in less cost for data replication technique. An innovative proactive duplicate checking technique to confirm information reliability is employed whereas to maintain data with the

minimum range of replicas (which helps as a value effectiveness benchmark for evaluation), which isn't any quite analysis of PRCR to explain that this approach is in a position of managing massive knowledge within the Cloud, considerably scale back storage space consumption of cloud at a negligible overhead.

REFERENCES

[1] Li W, Yang Y, Chen J and Yuan D 2012 *12th Int. Sym. on Cluster Cloud and Grid Computing* (Canada: IEEE/ACM) pp 564-571
[2] Li W, Yang Y and Yuan D 2011 *9th Int. Conf. on Dependable Autonomic and Secure Computing* (Australia: IEEE) pp 496-502
[3] Bachwani R, Gryz L and Bianchini R and Dubnicki C 2008 *Sym. on Reliable Distributed Systems* (Italy: IEEE) pp 85-94
[4] Yuan D, Cui L, Liu X, Fu E and Yang Y 2016 A Cost-Effective Strategy for Storing Scientific Datasets with Multiple Service Providers in the Cloud *Preprint* arXiv/160107028
[5] Balasubramanian B, Garg V 2013 Fault Tolerance in Distributed Systems Using Fused Data Structures *IEEE transactions on parallel and distributed systems* **24** pp 701-715
[6] Qin X, Thomas J E, Schwarz S J and Miller E L 2005 *13th Int. Sym. on Modeling Analysis and Simulation of Computer and Telecommunication Systems* (Atlanta: IEEE) pp 125-134
[7] Gibson G and Patterson D 1993 Designing Disk Arrays for High Reliability *Journal of Parallel and Distributed Computing* **17** pp 4-27
[8] Chun B, Dabek F, Haeberlen A, Sit E, Weatherspoon H, Kaashoek M F, Kubiatowicz J D, Morris R 2006 *3rd Sym. on Networked System Design and Implementation* (California: NSDI) **6** pp 45-58
[9] Bauer E and Adams R 2012 *Reliability and availability of cloud computing* (New Jersey: John Wiley) pp 1-15
[10] Huang C, Simitchi H, Xu Y, Ogus A, Calder B, Gopalan P, Li J and Yekhanin S 2012 *Proc. of the USENIX conf. on Annual Technical Conference* (Boston: ACM) p 12
[11] Armbrust M, Stoica I, Zaharia M, Fox A and Griffith R et al 2010 A view of cloud computing *Communications of the ACM* **53** pp 50-58
[12] Deelman E, Singh G, Livny M,Berrimen B and Good J 2008 *Int. Conf. for High Performance Computing Networking Storage and Analysis* (Texas: ACM/IEEE) pp 1-12
[13] Yuan D, Liu X and Cui L et al 2013 *9th International Conference on e-Science* (China: IEEE) pp 285-292