A LITERATURE SURVEY ON MULTITHREADING TECHNIQUES TO AVOID DATA LOSS IN CLOUD STORAGE

JEYA SHREE R[1], Dr.S UMA[2], KOUSALYA S[3]

[1] PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

[2]Head of the Department, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

[3]PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract – "Security" is a one of the important issues in cloud computing environment. Cloud computing is used to deliver software, storage and processing for millions of users across the world. In the user environment, it is the cyberspace based service which provides storage for users by using the delivery model. Cloud encompass billions of user data both private and public data. There are many sanctuary challenges in cloud. Data loss is an important issues in a majority of these storage systems. Overwhelmed the data loss problem, data can be massed transversely multiple parts in numerous regions with compound servers. Records which can be misplaced, their segment are stated and that portion is restored. This paper extant a literature review pastures of cloud computing with a focus on data loss in cloud storage. Thus the segments of the records which are misplaced could be identified and that portion is restored.

Keywords--Cloud computing, Data loss, challenges, Cloud Storage.

I. INTRODUCTION

Cloud computing has become the ideal method for delivering information and online functionality. Some cloud services emphasis on providing customers with spacious range of services and functionalities. These services and functionalities include e-shopping, research, social media networking, entertainment, protecting important documents. Some other cloud services emphasis on small business, large enterprises and governments. Some cloud services provides cloud storage to users at no charge. While other services charges some subscription based fee. Charges based cloud services are referred to as private clouds and are owned and controlled by organization, which provides a network security for critical data and resources sharing. Resources can be accessed only with in an organization, others cannot access those cloud resources. Public cloud can be used by millions of users at no cost. It provides billion of resource to the users. Companies which offer public cloud are Facebook, Google, amazon, yahoo etc. Hybrid clouds combine various public and private cloud resources into a solution. Private, public and hybrid are the deployment models used. There are various security issues related to cloud computing. These issues are categorized into two kinds: security issues faced by cloud providers and security issues faced by customers. The responsibility of the provider is to ensure that their infrastructure is secure and that their client's data and applications are protected while the user must ensure that the provider has taken the proper security measures to protect their information, and the user must use strong passwords and authentication. This paper discuss about the data loss in cloud storage systems, and also the methods used in finding the missing part of file and how it can be replaced in cloud storage system. Data loss is one of the major issue in cloud storage systems. Due to this data loss some times customer or user cannot download large files completely .due to the data loss bandwidth also increases.

II. STORAGE ARCHITECTURE OF CLOUD COMPUTING

Cloud storage is an evolving service model for distant backup and data coordination. There are two kinds of cloud storage namely single cloud storage and multi cloud storage. Single cloud storage is the first storage model. After that multi cloud storage model is developed to overcome single cloud storage model. Single cloud storage model raises about cloud outage(for example Gmail back soon for everyone),vendor lock-ins.in vendor lock-ins it is costly to switch cloud providers. As suggested by[2],a solution to single cloud storage is multi cloud storage. In multi cloud storage ,it places a proxy between users and multiple clouds.it contains stripe data along multiple clouds. The advantage of multi cloud storage is that if any one clouds repair then other cloud will maintain the data. Since data is replicated across many clouds.

III. CHALLENGES IN CLOUD COMPUTING STORAGE

Data loss is one of the major problem in cloud computing storage and there are many security issues in cloud computing storage.as suggested in [3],integrity of data is one of the main security issue in cloud storage.to protect outsourced data in cloud storage, integrity checking, fault tolerance and recovery procedures are used.to provide security to cloud storage regenerating codes are added in stripe data across multiple servers to provide fault tolerance[3].to avoid integrity issue, remotely checking of integrity in regenerating coded data against corruptions is made[3].

IV. DATASTORAGE IN CLOUDS

Data storage options or tiers has ultra-fast SSDs, fast medium- and high-capacity HDDs[5]. Storage management features include data protection, high availability (HA) and disaster recovery (DR) as well as footprint reduction (DFR) for space reduction, such as compress of data, re-duplication and large provisioning, which enables huge information to be stored for longer time at cheaper costs[5]. Software tools are used to create services and solutions, it consist of APIs, middleware, databases, applications, hypervisors which is used to build virtual machines (VMs) and virtual desktop infrastructures (VDI), being with cloud stack ware[5]. Some of the cloud stack ware are Open Stack, and associated management tools. Examples of VMs and VDI hypervisors consist of Citrix/Xen, KVM, Microsoft Hyper-V, Oracle and VMware ESX. In all the three cases, data storage is configured into storage systems, storage appliances and compute servers[5]. Public clouds are services accessible at free charge or for a fee, supplying different functionality, such as Amazon Web Services (AWS), Google Docs or Seagate or data backup software[5]. Public clouds are managed by their respective owners, whose customers pick to use their own services. Private clouds are owned and maintained by organizations and as same as to legacy IT services delivery. However, it is to be pointed that private clouds built using publicly available components, and existing offsite in different cloud provider locations, hybrid clouds offer services as a combination of public and private clouds.

A .Uploading a file in cloud storage

To upload a file F, we first divide it into k(n-k) equalize native chunks, denoted by$(F_i)_i$=1,2,……,k(n-k).we then encode these k(n-k) native chunks into n(n-k) code chunks, denoted by $(p_i)_i$=1,2,…..,n(n-k).each p_i is formed by a linear combination of the k(n-k) native chunks[2].

B. Downloading a file in cloud storage

To download a file, first download the corresponding metadata object that contains the ECVs. Then, select any k of the n storage nodes, and download then (n-k)code chunks from the k nodes. The ECVs of the k(n-k)code chunks can form a k(n-k)*k(n-k)square matrix. If the MDS property is maintained, then by definition, the inverse of the square matrix must exist. Thus, multiply the inverse of the square matrix is multiplied with the code chunks to obtain the original k(n-k)native chunks. The idea is that FMSR codes are treated as standard Reed-Solomon codes[2][26].

C TRANSIENT FAILURES IN CLOUD STORAGE

TABLE 1EXAMPLES OF TRANSIENT FAILURES IN SOME OTHER CLOUD SERVICES [2]

Cloud Service	Failure Reason	Duration	Date
Google Gmail	Software Bug	4 days	Feb 27-Mar2,2011
Google Search	Programming error	40 mins	Jan 31,209
Amazon S3	Gossip protocol blowup	6-8 hours	July 20,208
Microsoft Azure	Malfunction in Windows Azure	22 hours	Mar 13-14,2008
Verizon Terremark	Network component	24 hours	Oct 27-28,2013

In this section, the importance of repair in cloud storage, especially in cloud failures that make stored data permanently unrecoverable is discussed. Two types of failures namely transient failure and permanent failure are considered here.

Transient failure. A transient failure is expected to be short-term, such that the "failed" cloud will return to normal after some time and no outsourced data are lost. Table 1 shows several real-life examples for the occurrences of transient failures in today's clouds, where the durations of such failures range from several minutes to several days.It is highlight that even though Amazon claims that its service is designed for providing 99.99 percent availability [6], there are arising concerns about this claim and the reliability of other cloud providers after Amazon's outage in April 2011 [12].Thus it is expected that transient failures are common, but they will eventually be recovered. If multiple-cloud storage is deployed with enough redundancy, then data can be retrieved from the other surviving clouds during the failure period.

D Data loss and corruption.

There are real-life cases where a cloud may accidentally lose data [12]. From the literature[24],it is seen that in the case of Ma.gnolia [24], half a terabyte of data, including its backups, are lost and are unrecoverable.

E Data center outages in disasters.

AFCOM [25] found that many data centers are ill prepared for disasters. For example, 50 percent of the respondents have no plans to repair damages after a disaster. It was reported [25],that the earthquake and tsunami in northeastern Japan in March 11, 2011 knocked out several data centers there.

F Permanent failure in cloud storage

Unlike transient failures, where the cloud is assumed to be able to return to normal, permanent failures will make the hosted data in the failed cloud no longer accessible, so we must repair and reconstruct the lost data in a different cloud or storage site to maintain the required degree of fault tolerance. In our definition of repair, we mean to retrieve data only from the other surviving clouds, and reconstruct the data in a new cloud or another storage site.

V. DATA LOSSES IN STORAGE SYSTEM OF CLOUD

This literature review mainly focus on data loss in cloud storage system. In cloud storage system, instead of placing a data file in multiple clouds, the literature focus to place data file in multiple clouds across multiple regions [2].The data file is initially splitted into multiple parts and stored in multiple clouds across multiple regions. The data file may be stored based on hash function. If any one part is missed in data file, based on hash function value, by using structured file algorithm which part of data file is missed can be found and by using multithreading algorithm the misplaced data file is replaced. The multi-threading algorithm is placed on multiple regions of cloud. Using this algorithm the data file which is missed can be easily replaced [3].

In [2],it is implemented NCCloud as a proxy that bridges user applications and multiple clouds. Its design consist of three layers. In file system layer, NCCloud act as a mounted drive, which can be easily interconnected with general user applications. The coding layer deals with the encoding and decoding functions. The storage layer deals with read/ write requests with different clouds.

Each file is associated with a metadata object, which is replicated at each repository. The metadata object holds the file details and the coding information (e.g., encoding coefficients for FMSR codes).

Python and C are used to implement NCCloud [2]. The file system layer is built on FUSE [21]. The coding layer implements both RAID-6 and FMSR codes. Our RAID-6 code implementation is based on the Reed-Solomon code [52] for baseline evaluation. In [4]it is used to implement the RAID-6 codes, and the optimizations made in zfec to implement FMSR codes for fair comparison.

FMSR codes generate multiple chunks to be stored on the same repository. To save the request cost overhead, multiple chunks destined for the same repository are aggregated before upload. Thus, FMSR codes keep only one (aggregated) chunk per file object on each cloud, as in RAID-6 codes. To retrieve a specific chunk, we calculate it's offset within the combined chunk and issue a range GET request.

We make NCCloud deployable in one or multiple machines. In the latter case, we use ZooKeeper [13] to implement a distributed file-based shared lock to avoid simultaneous updates on the same file. We conduct preliminary evaluations in a LAN environment and observe that the overhead due to ZooKeeper is minimal. Here, we focus on deploying NCCloud on a single machine, and we mount NCCloud as a local file system

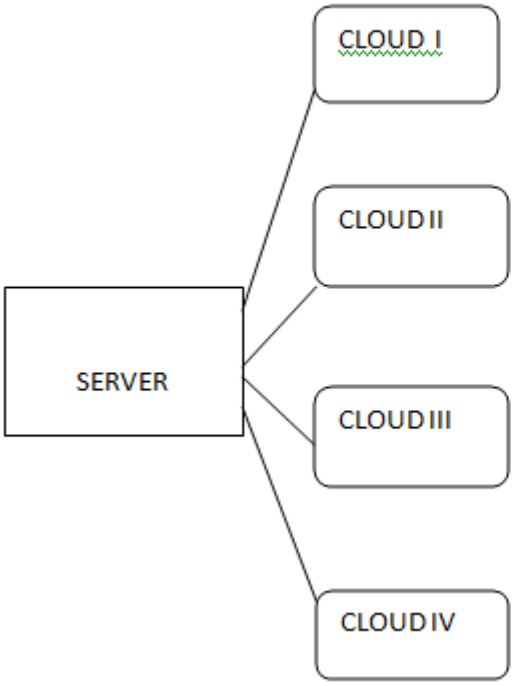

FIG I NORMAL OPERATION OF CLOUD STORAGE

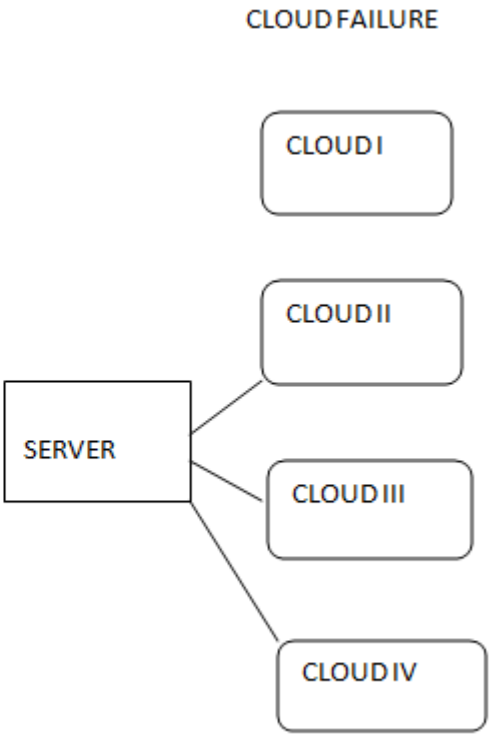

FIG II REPAIR OPERATION OF CLOUD STORAGE

Fig I represents the normal operation of cloud storage, i.e. whenever a file is uploaded, the server will store the copy of file in different cloud storage. FIG II represents the repair operation of cloud storage i.e. whenever one cloud affects failure condition, other cloud storages are used to restore or retrieve the original file content. During repair, the data can be regenerated from the new cloud. Fig II represents the distributed, multiple-cloud storage setting, in this setting large amount of data can be striped over multiple cloud vendors.in this setting ,server act as an interface between clouds and clients[2].whenever a cloud attempts permanent failure the server activates repair operation of cloud storage[4].during repair operation, there is no direct interactions between the clouds. The server reads the important data parts from other clouds and regenerates new data pieces and writes these regenerated new data pieces into new cloud [2]

VI CONCLUSION

Using multi thread algorithm and structured file algorithm the data part which is lost in a data file can be found and replaced.storing multiple parts of data in multiple regions of multiple clouds is the literature proposed concept.both multi thread algorithm and structured file algorithm are placed in multiple regions of cloud.by using this algorithm bandwidth increases,low download error, and downloading time becomes faster. This algorithm also provides reliability,fault tolerance.It is cost effective in case of permanent failures.it is implemented with the help of FMSR codes practical version. By using of FMSR codes degree of data redundancy can be increased. It is a proxy based multiple cloud storage system ,it also address the problem of cloud storage reliability.

Acknowledgment

Jeya shree R received Bachelor of Engineering degree in Computer Science and Engineering from Maharaja prithvi Engineering College , Avinashi, Coimbatore, Tamilnadu, India affiliated to Anna University Chennai.Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

Dr S.Uma is Professor and Head of PG Department of Computer Science and Engineering at Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India. She received her B.E., degree in Computer Science and Engineering in First Class with Distinction from PSG College of technology in 1991 and the M.S., degree from Anna University, Chennai, Tamilnadu, India. She received her Ph.D., in Computer Science and Engineering Anna University, Chennai, Tamilnadu, India with High Commendation. She has nearly 24 years of academic experience. She has organized many National Level events like seminars, workshops and conferences. She has

published many research papers in National and International Conferences and Journals. She is a potential reviewer of International Journals and life member of ISTE professional body. Her research interests are pattern recognition and analysis of non linear time series data.

Kousalya S received Bachelor of Engineering degree in Computer Science and Engineering from Easa College of Engineering and Technology, Coimbatore, India affiliated to Anna university Chennai. Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

REFERENCES

[1] Henry C.H.Chen and Patrick P.C. Lee,"Enabling Data Integrity Protection in Regenerating-Coding Based Cloud Stotage:Theory and Implementation,IEEE Transactions on Parallel and Distributed Systelms,vo 25,no.2,Feb 2014".

[2] Henry C.H.Chen ,Patrick P.C. Lee and Yang Tang" NCCloud:A Network-Coding-Based Storage System in a Cloud-of-Clouds,IEEE Transactions on computers,vol.63,no.1,Jan 2014".

[3] YungfengZhu,PatrickP.C.Lee,YinlongXu,YuchongHu,andLipingXiang,"On the Speedup of Recovery in Large-Scale Erasure-Coded Storage Systems,IEEE transactions on Parallel and Distributed Systems,vol.25,no.7,July 2014".

[4] CLEVERSAFE,CleversafeDispersedstorage,2008.[Online].Available:http://www.cleversafe.org/downloads

[5] H.Abu-Libdeh,L.Princehouse and H.weather spoon,"RACS:A Case for Cloud Storage Diversity,"proc.ACM First ACM Symp.Cloud Computing,2010

[6] Amazon Web Services,"AWS CaseStudy:Backupify",http://aws.amazon.com

[7] H.Blodget,"Amazon's Cloud Crash Disaster Permanently Destroyed Many Customer's Data,"http://www.businessinsider.com/apr 2014

[8] A.G.Dimakis, k.Ramachandran,Y.Wu and C.Suh,"A Survey on Network Codes for Distributed Storage,"proc.IEEE,vol.99,no.3,mar 2011

[9] Cloud computing storage and its issues[online],http://www.google.com.

[10] Storage Architecture of cloud computing[online],http://www.google.com.

[11] H. Blodget, "Amazon's Cloud Crash Disaster Permanently Destroyed Many Customers' Data," http://www.businessinsider. Com/amazon-lost-data-2011-4/, Apr. 2011.

[12] S. Jiekak, A.-M.Kermarrec, N.L. Scouarnec, G. Straub, and A.V. Kempen, "Regenerating Codes: A System Perspective," ACM SIGOPS Operating Systems Rev., vol. 47, no. 2, pp. 23-32, 2013.

[13] A. Kermarrec, N.L. Scouarnec, and G. Straub, "Repairing Multiple Failures with Coordinated and Adaptive Regenerating Codes,"Proc. Int'l Symp.Network Coding (NetCod '11), June 2011.

[14] O. Khan, R. Burns, J.S. Plank, W. Pierce, and C. Huang, "Rethinking Erasure Codes for Cloud File Systems: Minimizing I/O for Recovery and Degraded Reads," Proc. 10th USENIX Conf. File and Storage Technologies (FAST '12), 2012.

[15] R. Wauters,, "Online Backup Company Carbonite Loses Custo-mers' Data, Blames and Sues Suppliers," http://techcrunch.com/ 2009/03/23/online-backup-company-carbonite-loses-customers-data-blames-and-sues-suppliers/, Mar. 2009.

[16] A.G. Dimakis, P.B. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran, "Network Coding for Distributed Storage Sys-tems," IEEE Trans. Information Theory, vol. 56, no. 9, pp. 4539-4551, Sept. 2010.

[17] A.G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A Survey on Network Codes for Distributed Storage," Proc. IEEE, vol. 99, no. 3, 476-489, Mar. 2011.

[18] A. Duminuco and E. Biersack, "A Practical Study of Regenerating Codes for Peer-to-Peer Backup Systems," Proc. IEEE Int'l Conf. Distributed Computing Systems (ICDCS '09), 2009.

[19] B. Escoto and K. Loafman, "Duplicity," http://duplicity. nongnu.org/, 2013.

[20] D. Ford, F. Labelle, F.I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan, "Availability in Globally Distributed Storage Systems," Proc. Ninth USENIX Symp. Operating Systems Design and Implementation (OSDI '10), 2010.

[21] FUSE, "Introduction," http://fuse.sourceforge.net/, 2013.

[22] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google File System," Proc. 19th ACM Symp. Operating Systems Principles (SOSP '03), 2003.

[23] C. Gkantsidis and P. Rodriguez, "Network Coding for Large Scale Content Distribution," Proc. IEEE INFOCOM, 2005.

[24] B. Treynor, "Gmail Back Soon for Everyone," http://gmailblog. blogspot.com/2011/02/gmail-back-soon-for-everyone.html, 2013.

[25] K.M. Greenan, E.L. Miller, and T.J.E. Schwarz, "Optimizing Galois Field Arithmetic for Diverse Processor Architectures and Applica-tions," Proc. IEEE Int'l Symp. Modeling, Analysis and Simulation of Computers and Telelcomm. Systems (MASCOTS '08), 2008.

[26] J.S. Plank,J.Luo,C.D.Schuman,l.Xu, and Z.wilcox"A performance evaluation and examination of open source erasure codinig libraries for storage.