A proposed Solution: Data Availability and Error Correction in Cloud Computing

Anil Gupta *anil_sg@yahoo.com*
Maharaja Ranjit Singh College,
Hemkunt Campus, Khandwa Road,
Indore-452001, MP, India

Parag Pande *parag.pande@yahoo.com*
Shri Satya Sai Instituite of Science and Technology,
Sehore M.P, India

Aaftab Qureshi *aaftab_toc@yahoo.com*
Shri Satya Sai Instituite of Science and Technology,
Sehore, M.P, India

Vaibhav Sharma *vaibhav.sharma_09@yahoo.com*
Maharaja Ranjit Singh College,
Hemkunt Campus, Khandwa Road,
Indore-452001, MP, India

Abstract

Cloud Computing is the hottest technology in the market these days, used to make storage of huge amounts of data and information easier for organizations. Maintaining servers to store all the information is quite expensive for individual and organizations. Cloud computing allows to store and maintain data on remote servers that are managed by Cloud Service Providers (CSP) like Yahoo and Google. This data can then be accessed through out the globe. But as more and more information of individuals and companies is placed in the cloud, concerns are beginning to grow about just how safe an environment it is. In this paper we discussed security issues and requirements in the Cloud and possible solutions of some the problems. We develop an architecture model for cloud computing to solve the data availability and error correction problem.

Keywords: Cloud Computing, Security Issues, Cloud Security, Cloud Architecture.

1. INTRODUCTION

One of the identifying characters of cloud computing is that computing is delivered via the Internet as services. Computing and IT resources are encapsulated as services, hiding all the details of implementation, deployment, maintenance and administration [1]. Computing will be shifted from on-premise systems to remote systems and users are connected to their data via the Internet. Individual organizations will lose their control of their data to some extent, as the data is stored over the Internet and is likely leased from cloud operators. With cloud computing, deployment of IT systems and data storage is changed from on-premises user-owned IT infrastructures to off-premises third-party IT infrastructures. Having the whole IT systems and data on infrastructures with limited controls creates an obstacle for migrating traditional IT systems and data into clouds, as users have the following security concerns:

Limited control over the data may incur security issues.

As the data is on the single cloud, data availability becomes a great challenge.

Having the whole IT system and data on a single cloud may give the cloud operator excessive power for controlling and modifying users' data.

In this paper we try to specify security issues and requirements in the Cloud and possible solutions of some the problems. We develop an architecture model for cloud computing to solve the data availability and error correction problem.

Our paper is organized is as follows: Section 2 identifies the security issues and requirement that users must be aware when adopting cloud computing. Section 3 surveys the related work. Section 4 summaries the RAID models to address the security concerns. Section 5 explains our proposed architecture for cloud computing with features provided by the models. Section 6 concludes the paper.

2. SECURITY ISSUES AND REQUIREMENT

Security concerns [2,3] have been raised due to the new computing model introduced by cloud computing, which is characterized by off-premises computing, lost control of IT infrastructure, service-oriented computing, and virtualization, and so on. Security concerns from users can be briefly summarized as follows:

- **System failure and Data availability:** When keeping data at remote systems owned by others, data owners may suffer from system failures of the service provider,
  Requirement: If the Cloud goes out of operation, data will become unavailable as the data depends on a single service provider.

- **Data error:** Client data should be error free on the cloud. As the data is stored on the cloud whereas the client is at other side. If the correct storage strategy is not used data might not be stored correctly on the storage server of the Cloud.
  Requirement: Very essential requirement for the client on the cloud.

- **Data Migratibility:** Users that adopt cloud computing may subject to the risk that their data cannot be migrated to other clouds.
  Requirement: Without the capability of migrating data to other clouds, users may be forced to stay with a cloud if they have considerable dependence on the data.

- **Data confidentiality and integrity:** Data generated by cloud computing services are kept in the clouds. Keeping data in the clouds means users may lose control of their data and rely on cloud operators to enforce access control [25, 19].
  Requirement: They may not be able to prevent unauthorized disclosure or malicious modification of their data.

- **Long-term viability:** Ideally, cloud computing provider will never go broke or get acquired and swallowed up by a larger company. But you must be sure your data will remain available even after such an event.
  Requirement: How you would get your data back and if it would be in a format that you could import into a replacement application in such an event.

- **Data location:** When you use the cloud, you probably won't know exactly where your data is hosted. In fact, you might not even know what country it will be stored in.
  Requirement: Ask providers if they will commit to storing and processing data in specific jurisdictions, and whether they will make a contractual commitment to obey local privacy requirements on behalf of their customers.

- **Data segregation:** Make sure that encryption is available at all stages, and that these encryption schemes were designed and tested by experienced professionals.
  Requirement: For data security and privacy.

- **Data Recovery:** Even if you don't know where your data is, a cloud provider should tell you what will happen to your data and service in case of a disaster. Any offering that does not replicate the data and application infrastructure across multiple sites is vulnerable to a total failure.
  Requirement: otherwise clients will losses their all data. Provider has "the ability to do a complete restoration, and how long it will take."

- **Data security:** Security will need to move to the data level so that enterprises can be sure their data is protected wherever it goes. For example, with data-level security, the enterprise can specify that this data is not allowed to go outside of the European Union. It can also force encryption of certain types of data, and permit only specified users to access the data. It can provide compliance with the Payment Card Industry Data Security Standard (PCI DSS).
  Requirement: For data security and privacy.

- **Data privacy:** The data privacy is also one of the key concerns for Cloud computing. A privacy steering committee should also be created to help make decisions related to data privacy.
  Requirement: This will ensure that your organization is prepared to meet the data privacy demands of its customers and regulators.

These concerns have been identified in several earlier works [12, 19]. Armbrust et al. [19] considered these concerns as the top most obstacles to growth of cloud computing.

3. RELATED WORK

Extensive research efforts have been put into cloud computing and its related technologies, resulting in several well acknowledged cloud computing theories and technologies, including MapReduce [4] and its implementation Apache Hadoop [5], Microsoft Dryad [6], Condor DAGman [7], Eucalyptus [26], Nimbus [8], Reservoir [27], and CARMEN [9].

Various security related issues and concerns in cloud computing have been identified and are studied, including data privacy [10, 11, 12], data protection [13], access control [14, 15, 12], availability [16], authentication [17], scalability [18].

Armbrust et al. [19] identified ten obstacles to growth of cloud computing. The top three obstacles are actually very close to the concerns identified in Section 2.

Research in security patterns has established a structural way and a proven practice for secure system designs and implementations. They provide guidelines as well as knowledge that are proven and standardized [20, 21, 22, 23].

Domain security is a method developed by Qinetiq to develop architectural models for applications based on security requirements [24]. The architectures generated by the Domain Security method focus on the software engineering aspect of systems to implement, instead of security protocols, cryptographic operations and so on.

4. RAID MODELS

RAID 1: RAID 1 creates an exact copy (or mirror) of a set of data on two or more disks. As shown in figure 1. This is useful when performance read or data availability (reliability) is more important than data storage capacity. Such an array can only be as big as the smallest member disk. A classic RAID 1 mirrored pair contains two disks, which increases availability (reliability) over a single disk. Since each member contains a complete copy of the data, and can be addressed independently, ordinary wear-and-tear reliability is raised by the power of the number of self-contained copies.

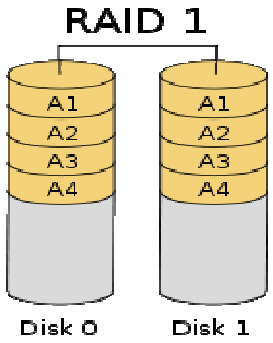

FIGURE 1: RAID 1 Model

RAID 5: RAID 5 uses block-level striping with parity data distributed across all member disks. As shown in figure 2. RAID 5 has achieved popularity because of its low cost of redundancy. This can be seen by comparing the number of drives needed to achieve a given capacity. For an array of n drives, with Smin being the size of the smallest disk in the array, In RAID 5 storage capacity is $S_{\min} \times (n-1)$.

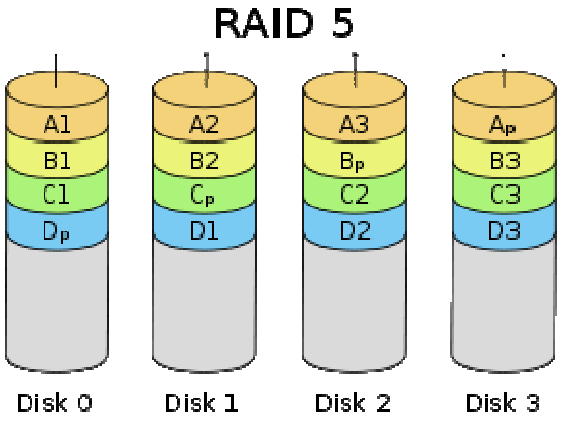

FIGURE 2: RAID 5 Model

Parity Calculation

A concurrent series of blocks (one on each of the disks in an array) is collectively called a stripe. If another block or some portion of that block is written on that same stripe, the parity block or some portion of that block is recalculated and rewritten. For any write operation it requires:

- Read the old data block
- Read the old parity block
- Compare the old data block with the write request. For each bit that has flipped (changed from 0 to 1, or from 1 to 0) in the data block, flip the corresponding bit in the parity block
- Write the new data block
- Write the new parity block

The disk used for the parity block is staggered from one stripe to the next; hence the term distributed parity blocks. RAID 5 write operations are expensive in terms of disk operations and traffic between the disks and the controller.

The parity blocks are not read on data reads, since this would add unnecessary overhead and would diminish performance. The parity blocks are read, however, when a read of blocks in the stripe fails due

to failure of any one of the disks, and the parity block in the stripe are used to reconstruct the errant sector. The CRC error is thus hidden from the main computer. Likewise, should a disk fail in the array, the parity blocks from the surviving disks are combined mathematically with the data blocks from the surviving disks to reconstruct the data from the failed drive on-the-fly.

This is sometimes called Interim Data Recovery Mode. The computer knows only that a disk drive has failed, so that the operating system can notify the administrator that a drive needs replacement; applications running on the computer are unaware of the failure. Reading and writing to the drive array continues seamlessly, though with some performance degradation.

5. PROPOSED MODEL

We are proposing a model for cloud, based on RAID architecture. In our proposed model we are using RAID 1 and RAID 5 level.This configuration can sustain the failure of all disks in either of the arrays, plus up to one additional disk from the other array before suffering data loss, i.e., by using this architecture at the data storage servers in the Cloud, we can handle the problem of Fault tolerance, data availability and Data recovery.

We combine both RAID 1 and RAID 5 for our proposed architecture. These combine architecture known as RAID 51 architecture. A **RAID 1** creates an exact copy (or **mirror**) of a set of data on two or more disks. This is useful for data availability (reliability). A RAID **5** uses block level striping with parity data distributed across all member disks. This is very important for data correction.

RAID 51: RAID 51 architecture is an array that consists of two RAID 5's that are mirrors of each other. In this configuration reads and writes are balanced across both RAID 5s. In this architecture, there are two set of RAID 5 model that are mirror of each other. As well as they don't have any idea that their mirror image is also exist. Mirroring provides guarantee of data availability and since they are not aware about the mirroring, if any one try to change the data can easily be traceable. Similarly, the RAID 1 has no idea that its underlying disks are RAID 5's. As we are using RAID 5 under the RAID 1 architecture, we get the data correction facility also. Because A RAID 5 uses block level striping with parity data distributed across all member disks. As shown in figure 3.

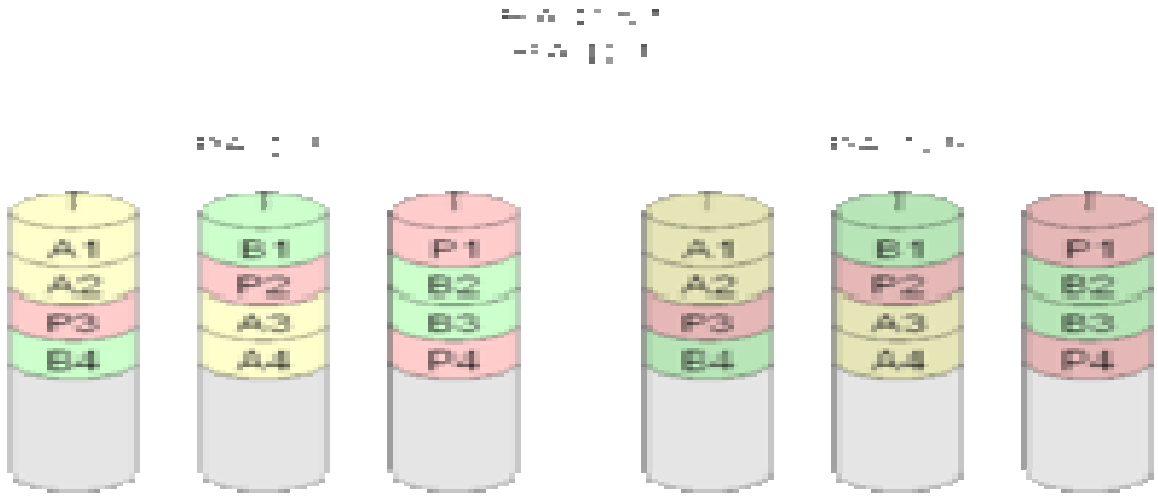

FIGURE 3: RAID 5+1 Model

Though RAID solves our problems but it suffer from poor performance when faced many write operations which are smaller than the capacity of a single stripe. This is because parity must be updated on each write, requiring read-modify-write sequences for both the data block and the parity block.

We can use RAID 6 in place of RAID 5 also. It extends RAID 5 by adding an additional parity block; thus it uses block-level striping with two parity blocks distributed across all member disks. RAID 6 does not have a performance penalty for read operations, but it does have a performance penalty on write operations because of the overhead associated with parity calculations. Performance varies greatly depending on how RAID 6 is implemented in the storage architecture.

Parity Calculation

Two different syndromes need to be computed in order to allow the loss of any two drives. One of them, P can be the simple XOR of the data across the stripes, as with RAID 5. A second, independent syndrome is more complicated and requires the assistance of field theory.

To deal with this, the Galois field GF (m) is introduced with m = 2k, where $GF(m) \cong F_2[x]/(p(x))$for a suitable irreducible polynomial p(x) of degree k. A chunk of data can be written as dk − 1dk − 2...d0 in base 2 where each di is either 0 or 1. This is chosen to correspond with the element dk − 1xk − 1 + dk − 2xk − 2 + ... + d1x + d0 in the Galois field. Let $D_0, ..., D_{n-1} \in GF(m)$correspond to the stripes of data across hard drives encoded as field elements in this manner (in practice they would probably be broken into byte-sized chunks). If g is some generator of the field and $\oplus$denotes addition in the field while concatenation denotes multiplication, then $\mathbf{P}$and $\mathbf{Q}$may be computed as follows (n denotes the number of data disks):

$$\mathbf{P} = \bigoplus_i D_i = \mathbf{D_0} \oplus \mathbf{D_1} \oplus \mathbf{D_2} \oplus \ldots \oplus \mathbf{D}_{n-1}$$

$$\mathbf{Q} = \bigoplus_i g^i D_i = g^0\mathbf{D_0} \oplus g^1\mathbf{D_1} \oplus g^2\mathbf{D_2} \oplus \ldots \oplus g^{n-1}\mathbf{D}_{n-1}$$

where $\oplus$ is a bitwise XOR operator and gi is the action of a linear feedback shift register on a chunk of data. Thus, in the above formula, the calculation of P is just the XOR of each stripe. This is because addition in any characteristic two finite fields reduces to the XOR operation. The computation of Q is the XOR of a shifted version of each stripe.

Mathematically, the generator is an element of the field such that gi is different for each nonnegative i satisfying i < n.

If one data drive is lost, the data can be recomputed from P just like with RAID 5. If two data drives are lost or the drive containing P is lost the data can be recovered from P and Q using a more complex process. Working out the details is not hard with field theory. Suppose that Di and Dj are the lost values with $i \neq j$. Using the other values of D, constants A and B may be found so that $D_i \oplus D_j = A$and $g^i D_i \oplus g^j D_j = B$. Multiplying both sides of the latter equation by gn − i and adding to the former equation yields $(g^{n-i+j} \oplus 1)D_j = g^{n-i}B \oplus A$and thus a solution for Dj which may be used to compute Di.

The computation of Q is CPU intensive compared to the simplicity of P. Thus, a RAID 6 implemented in software will have a more significant effect on system performance.

6. CONCLUSION

Cloud Computing is the cost, time and performance effective. Some basic Security issues are the key concern in the Cloud Computing use and in the implementation for the Client as well as for Vendors.

Security concern of the Cloud infrastructure relies on trusted computing and cryptography. Organizational data must be protected in a manner consistent with policies. No standard contract exists that the cover the security related issues. Having a list of common outsourcing provisions, such as privacy and security standards, regulatory and compliance issues, service level requirements and penalties, change management processes, continuity of service provisions, and termination rights, provides a useful starting point. The migration to a cloud computing environment is in many ways an exercise in risk management.

A RAID 1 model can be an effective protection against physical disk failure; it does not provide protection against data corruption due to viruses, accidental file changes or deletions, or any other data-specific changes. By design, any such changes will be instantly mirrored to every drive in the array segment. A virus, for example, that damages data on one drive in a RAID 1 array will damage the same data on all other drives in the array at the same time. For this reason system using RAID 1 to protect against physical drive failure should also have a traditional data backup process in place to allow data restoration to previous points in time. It would seem self-evident that any system critical enough to require disk redundancy also needs the protection of reliable data backups.

The risks must be carefully balanced against the available safeguards and expected benefits, with the understanding that accountability for security concern remains with the organization. Too many controls can be inefficient and ineffective, if the benefits outweigh the costs and associated risks. An appropriate balance between the strength of controls and the relative risk associated with particular programs and operations must be ensured.

7. REFERENCES

[1] Anil Gupta, Aaftab Qureshi, Parag Pande,"Cloud Computing Characteristics and Service Models: our own interpretation".

[2] Kresimir Popovic, et al., "Cloud "Computing issues and challanges" MIPRO 2010 May 24-28 Opatija, Croatia, pp 344-349.

[3] Gansen Zhao, et al., "Deployment Models: Towards Eliminating Security Concerns from Cloud Computing" IEEE 2010, pp 189-195.

[4] J. Dean and S. Ghemawat. "Mapreduce: simplified data processing on large clusters". *Commun. ACM*, 51(1):107–113, 2008.

[5] Apache Hadoop, 2009. http://hadoop.apache.org/.

[6] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly."Dryad: distributed data-parallel programs from sequential building blocks". In EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007, pages 59–72, New York, NY, USA, 2007. ACM.

[7] Condor DAGman, 2009. http://www.cs.wisc.edu/condor/dagman/. [18] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. "The eucalyptus open-source cloud-computing system". In Proceedings of Cloud Computing and Its Applications, October 2008.

[8] Nimbus. "Introduction to nimbus", 2009. http://workspace.globus.org/clouds/nimbus.html. [3] S. Beco, A. Maraschini, and F. Pacini. "Cloud computing and RESERVOIR project". NUOVO CIMENTO DELLA SOCIETA ITALIANA DI FISICA C-COLLOQUIA ON PHYSICS, 32(2), Mar-Apr 2009.

[9] CARMEN, 2009. http://www.carmen.org.uk/.

[10] Å. A. Nyre and M. G. Jaatun. “Privacy in a semantic cloud: What’s trust got to do with it?”. In The First International Conference on Cloud Computing, pages 107–118, 2009.

[11] S. Pearson, Y. Shen, and M. Mowbray. “A privacy manager for cloud computing”. In The First international Conference on Cloud Computing, pages 90–106, 2009.

[12] L. Kaufman. “Data security in the world of cloud computing”. IEEE SECURITY & PRIVACY, 7(4), July- August 2009.

[13] S. Creese, P. Hopkins, S. Pearson, and Y. Shen. “Data protection-aware design for cloud services”. In The First International Conference on Cloud Computing, pages 119–130, 2009.

[14] L. Hu, S. Ying, X. Jia, and K. Zhao. “Towards an approach of semantic access control for cloud computing”. In The First International Conference on Cloud Computing, pages 145–156, 2009.

[15] D. Chen, X. Huang, and X. Ren. “Access control of cloud service based on ucon”. In The First International Conference on Cloud Computing, pages 559–564, 2009.

[16] T. Uemura, T. Dohi, and N. Kaio. “Availability analysis of a scalable intrusion tolerant architecture with two detection modes”. In The First International Conference on Cloud Computing, pages 178–189, 2009.

[17] L. Yan, C. Rong, and G. Zhao. “Strengthen cloud computing security with federal identity management using hierarchical identity-based cryptography”. In The First International Conference on Cloud Computing, pages 167–177, 2009.

[18] G. Zhao, J. Liu, Y. Tang, W. Sun, F. Zhang, X. ping Ye, and N. Tang. “Cloud computing: A statistics aspect of users”. In The First International Conference on Cloud Computing, pages 347–358. Springer, 2009.

[19] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. “Above the clouds: A Berkeley view of cloud computing”. Technical Report UCB/EECS- 2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[20] M. Schumacher, E. Fernandez, D. Hybertson, and F. Buschmann. SECURITY PATTERNS: INTEGRATING SECURITY AND SYSTEMS ENGINEERING. John Wiley & Sons, 2005.

[21] T. Heyman, K. Yskout, R. Scandariato, and W. Joosen. “An analysis of the security patterns andscape”. In SESS ’07: Proceedings of the Third International Workshop on Software Engineering for Secure Systems, page 3, Washington, DC, USA, 2007. IEEE Computer Society.

[22] E. B. Fernandez, J. Wu, M. M. Larrondo-Petrie, and Y. Shao. On building secure scada systems using security patterns. In CSIIRW ’09: Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research, pages 1–4, New York, NY, USA, 2009. ACM.

[23] B. Blakley and C. Heath. SECURITY DESIGN PATTERNS, 2004. The Open Group Security Forum.

[24] K. J. Hughes. “Domain Based Security: enabling security at the level of applications and business processes”, 2002. www.qinetiq.com.

[25] A. Singh, M. Srivatsa, and L. Liu. “Search-as-a-Service: Outsourced Search over Outsourced Storage”. *ACM TRANSACTIONS ON THE WEB*, 3(4), September 2009.

[26] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. “The eucalyptus open-source cloud-computing system”. In Proceedings of Cloud Computing and Its Applications, October 2008.

[27] S. Beco, A. Maraschini, and F. Pacini. “Cloud computing and RESERVOIR project”. NUOVO CIMENTO DELLA SOCIETA ITALIANA DI FISICA C-COLLOQUIA ON PHYSICS, 32(2), Mar-Apr 2009.

[28] Wikipedia” www.wikipedia.org”.