An Adaptive Methodology for Integrity Checking in Cloud Storage

I. Jambulingam, T V Ananthan, P S Rajakumar

Abstract: Majority of the organization uses cloud for storage purpose in order to reduce the cost as well as maintenance. Due to increasing threat from internal and external sources, there would be possibility of corruption in the cloud storage files. Thus the storage must be monitored periodically for integrity checking. Since most of the Data Owners have limited resources thus the responsibility of integrity checking goes to the Third Party Auditors (TPA). Usually the static way of deciding to use particular hash tree methodology to store the cloud storage meta-data, which is mainly used for integrity checking throughout is inappropriate for two main reasons, firstly, due to more fluctuated loss or corrupted cloud data, secondly, based on the variations in the number of files in the cloud users*, directory; Initially the static approach would be good but it may not be optimal solution at the later period. therefore, in this paper, we have proposed Adaptive Integrity Checking method (AIC), which would lead a way for adaptive dynamic hash tree methodology for holding the cloud storage meta-data; which would drastically increases the performance of integrity checking in terms of both time and space complexity besides the benefits obtained in the EDHT-n version and HEDHT methodologies of handling the cloud storage integrity checking.

Keywords: Third Party Auditor (TPA), Adaptive Integrity Checking (AIC), Hybrid Enhanced Dynamic Hash Tree (HEDHT), Enhanced Dynamic Hash Tree (EDHT), Meta-Data, Cloud Service Provider (CSP), Microservice, API.

I. INTRODUCTION

Storage data is the asset to the client, to reduce cost; original file is being stored in cloud and deleted in the end-user. Cloud storage is a model of networked enterprise storage where data is stored in virtualized pools of storage which are generally hosted by third parties. Cloud storage provides customers with benefits, ranging from cost saving and simplified convenience to mobility opportunities and scalable service. These great features attract more and more customers to utilize and storage their data to the cloud storage: according to the analysis report, the volume of data in cloud is expected to achieve 40 trillion gigabytes in 2020. Even though cloud storage system has been widely adopted, it fails to accommodate some important emerging needs such as the abilities of auditing of cloud files by cloud clients. Thus integrity should be foremost requirement, any corruption had happened in CSP, it should be detected and corrected either during the file request given by the users or during auditing stage and made immediate recovery in order to maintain the integrity in the form of consistency across all of its data backup copies [14] by executing the recovery routines. Most importantly, the user may not aware about these processes are running in background, and thus Business Continuity Planning would be ensured always. Subsequently the trustworthiness towards the CSP would be improved. EDHT-N and HEDHT are the proposed Hash Tree Methodologies as advanced to MHT [2],[22] as shown in Figure 1 and Figure 2, where meta-data of the cloud file will be stored in static way, now we are proposed Adaptive Integrity Checking for the same purpose in a dynamic manner to speed up the audit and recovery process effectively compared to the earlier proposed paper.

II. PROPOSED TECHNIQUE

Adaptive Integrity Checking (AIC)

Audit would be performed by TPA [5],[7],[10] on regular frequency with cloud storage to determine the faulty file in cloud, subsequently, TPA would record those error details in its database to ascertain the correct integrity checking methodology dynamically based on the Adaptive Integrity Checking (AIC) method and the same to be adopted. Initially, we say best methodology because it requires less number of hash computations for tree formation as well as recovery process compared to static EDHT-n/HEDHT methodologies.

TPA [12] would capture the error rates of each and every user directories available in cloud on a particular frequencies or whenever user access the files, hence based on these various error rate statistics details, we could pick the most occurring error rate.
percentage or sometimes if the error rate percentage are not repeating then we should take an average of all those recorded value and finally decide which hash tree methodology to be adopted.

We also determine appropriate hash tree methodologies to store the cloud meta-data dynamically based on the variation in the number of files in the cloud users’ directories. Thus it improves the performance of the auditing. AIC batch processing [11] would furthermore improve the speed of the auditing as well as recovery process.

III. ADAPTIVE INTEGRITY CHECKING (AIC) IN THIRD PARTY AUDITOR (TPA) USING API AND MICROSERVICE

Integrity Checking APIs are developed in a RESTful style as shown in the Table 1. These 10 APIs will have CRUD capabilities GET (get a single item or a collection), POST (add an item to a collection), PATCH (edit an item that already exists in a collection), DELETE (delete an item in a collection).

Table 1: APIs used for audit and error detection of corrupted cloud storage to ease the recovery process.

Also it were developed in a microservice - a software development technique - a variant of the service-oriented architecture (SOA) architectural style that structures an application as a collection of loosely coupled services, so each services are fine-grained and the protocols are lightweight.

As per Figure 3, User [3],[19],[21],[23] would request thru Cloud client for enquiry/delete/update operation of cloud file to the specific CSP thru TPA. API is a contract that provides guidance for a consumer (Cloud Client) to use the underlying service available in TPA, CSP, MDS, Recovery system[17]. API is usually a portion of a microservice, allowing for interaction with other microservices.

The broken out sections of the business logic, each encompassing a microservice. Thus complexity of the application is reduced, as the different services have well-defined interactions with each other.

Microservices architecture has become a common approach for any cloud integrity checking to achieve agility and the Continuous Delivery of applications to meet the growing demand of the cloud users.

Cloud Developers and users today are concerned with speedy development cycles, software scale, performance, and flexibility. Microservices offer all of these.

Table 1: List of APIs for Cloud Integrity Checking

| # | REST API        | Provider | Consumer       | Request Type |
|---|----------------|----------|----------------|--------------|
| 1 | cloud-file     | TPA      | Cloud Client   | PATCH, GET, DELETE |
| 2 | root-digest    | CSP      | TPA            | GET          |
| 3a| recover-file   | Recovery System | TPA  | GET          |
| 3b| file-retrieve  | CSP      | TPA            | GET          |
| 4 | create-modified-file-digest | MDS  | Cloud client  | POST         |
| 5 | file-push      | CSP      | Cloud client   | POST         |
| 6 | digest-push    | TPA      | Cloud client   | POST         |
| 7 | create-new-file-digest | MDS  | Cloud client  | POST         |
| 8 | tpa-audit      | CSP      | TPA            | GET          |
| 9 | construct-edit-construct-hedht | CSP  | TPA            | POST         |

Cloud-file: User would request thru Cloud client [18],[21] for enquiry/delete/update/create operation of cloud file to the specific CSP thru TPA.

Root-digest: Only in case of enquiry and update operation, proof of ownership [1],[4],[6],[15],[16],[8],[9] will be determined, hence TPA would fetch the challenge corresponds to the user cloud file and send to CSP, which in turn send the root of EDHT-n/HEDHT for the received challenge. If root send by CSP matches with the root available in TPA then file reside in cloud is in intact.

Recover-file: Root of EDHT-n/HEDHT send by CSP does not matches with the root available in TPA then file reside in cloud is not intact. In case of file corrupt, invoke the recovery process.

File-retrieve: In case of file intact, invoke the actual file for view or modify. In case of modify, the file should be locked in serialiable mode to avoid dirty read, phantom read, non-repeatable read problems and sent to TPA.

Create-modified-file-digest: Fetched cloud file would be...
modified by the user/group users whose got privileges and subsequently sent to MDS.

**File-push:** In case of file update/create, the modified/created file will be pushed to CSP along with meta-data received from MDS to reconstruct the EDHT-n/HEDHT in case of existing, else it would create new EDHT-n/HEDHT for the newly created Cloud Directory file.

**Digest-push:** Cloud client send the meta-data received from MDS [24] to TPA to reconstruct EDHT-n/HEDHT in case of modified/created file or construct new EDHT-n/HEDHT in case of newly created cloud directory file. Thus CSP and TPA would always be kept in consistency the EDHT-n/HEDHT.

**Tpa-audit:** Audit would be performed by TPA on regular frequency with CSP to determine the faulty file in cloud, hence based on these values, TPA would record the errors in TPA database to ascertain the correct integrity checking methodology dynamically.

**Create-new-file-digested:** User can add new file to the cloud in particular directory by requesting the cloud client which subsequently request MDS to generate meta-data.

**Construct-edhtn or construct-hedht:** TPA send the same meta-data and methodology proposed by AICA to CSP directly instead of sending to cloud client, CSP uses these details to reconstruct the suggested EDHT-n/HEDHT.

**Aica-execute:** TPA would execute Adaptive Integrity Checking Algorithm (AICA) to arrive which optimal hash tree methodology to use for storing the cloud storage meta-data (either to use new or to retain the old methodology); by inputting the parameters obtained i.e., error percentage rate identified during TPA audit using tpa-audit API, which is noted as per Table 2 and variations in number of files in the cloud user directories identified during file-push to subsequently construct the appropriate suggested EDHT-n/HEDHT as per Table 3. Once the new hash tree methodologies is identified, it should be informed to CSP by invoking the construct-edhtn or construct-hedht API in order to synchronize between TPA and CSP of cloud storage meta-data to maintain consistency, which would ease the cloud storage auditing process efficiently as shown in Figure 4.
IV. RESEARCH METHODOLOGY

After a series of attempts and summaries, we put forward our solution as follow:

| Table 4 Notations used in AIC Algorithm |
|----------------------------------------|
| find($\ell_{hashtree}$) | To find height of hash tree of particular methodology |
| $\ell_{hashtree}$ | Height of hash tree |
| $N_s$ | Number Of Siblings required in a particular methodology |
| VN_S | Various Number of Siblings i.e., EDHT-3 is 3; EDHT-4 is 4 and so on |
| $N_f$ | Number of files |
| $N_k$ | Number of files currently |
| $N_{htc}$ | Number of hash tree Defect calculation to correct it |
| $N_{htcc}$ | Number of hash tree computation for each methodology |
| d_p | Defect percentage |
| min_r | Minimum range |
| max_r | Maximum range |
| $f_{aic}$ | Final Adaptive Integrity Checking Method |
| $c_{aic}$ | Current Best Integrity Checking Method |
| temp_r | Temporary range |
| $f_d$ | Defected files |
| $\ell_{htvs}$ | Height of hash tree for various number of siblings |
| $T_{hcv}$ | Total average hash tree calculation for various number of siblings |
| $T_{hcv}$ | Total Worst hash tree calculation for various number of siblings |
| $T_{hcv}$ | Total Best hash tree calculation for various number of siblings |
| $f_{mhcc}$ | To find minimum hash tree calculation from the collection |
| $h_{cph}$ | Hash calculation at particular height |
| $f_{i}$ | Initial Number of files |

Algorithm 1 find($\ell_{hashtree}$)

Input: $N_s$, $N_f$

Output: $\ell_{hashtree}$

$h_{cph} \leftarrow 1$

if $N_f = 1$ then
  return 2;
else
  while(true)
    $\ell_{hashtree} \leftarrow \ell_{hashtree} + 1$;
    $h_{cph} \leftarrow h_{cph} \times \ell_{hashtree}$;
    if $h_{cph} = N_f$ then
      break;
      $\ell_{hashtree} \leftarrow \ell_{hashtree} + 1$;
    end if
  return $\ell_{hashtree}$;
end if

Algorithm 1 is used to find out a height for Hybrid EDHT and EDHT-n versions, basically it takes the number of files as an input as well as the number of sibling to represent the version number only in case of EDHT-n methodology and return the height.

Algorithm 2 $N_{htcc}$

Input: $N_s$, $\ell_{hashtree}$

Output: $N_{htcc}$

if ($N_s > 0$) then
  $N_{htcc} \leftarrow (\ell_{hashtree} - 1) \times N_s$;
else
  for $i = 2$; $i <= \ell_{hashtree}$; $i++$
    $N_{htcc} \leftarrow N_{htcc} + i$;
  end for
else
  return $N_{htcc}$;
end if

Algorithm 2 is used to calculate the hash computation required for the hybrid EDHT and EDHT-n version to identify one defected file and thus it will be multiplied with number of defected files to compute total hash computation to find out all the faulty files.

Algorithm 3 hashTreeComputation

Input: $N_s$, $\ell_{hashtree}$

Output: $N_{htc}$

$N_{htc} \leftarrow 0$

if ($N_s > 0$) then
  $N_{htc} \leftarrow (\ell_{hashtree} - 1) \times N_s$;
else
  for $i = 2$; $i <= \ell_{hashtree}$; $i++$
    $N_{htc} \leftarrow N_{htc} + i$;
  end for
end if

Algorithm 3 is used to compute total hash computation required for hybrid EDHT and EDHT-n version for the given height of it and number of siblings only in case of EDHT-n version.

Algorithm 4 adaptiveIntegrityChecking

Input: $d_p$, $N_f$

Output: $f_{aic}$/ $c_{aic}$, $N_{htc}$/ $N_{htcc}$

for $N_{fc} \leftarrow i$; $N_{fc} <= N_f$; $N_{fc}++$
  $f_d \leftarrow N_{fc} / 100 \times d_p$;
  $\ell_{htvs} \leftarrow \text{find}(\ell_{hashtree})(\text{VN}_S, N_{fc})$;
  $h_{cph} \leftarrow h_{cph} \times \ell_{htvs}$;
  if $h_{cph} = N_f$ then
    $N_{htc} \leftarrow N_{htc} + h_{cph}$;
  end if
  $N_{htcc} \leftarrow N_{htcc} \times N_{htc}$;
end for

Algorithm 4 is used to check the integrity of the file and return the status of the file.

Algorithm 5 $T_{hcv}$

Input: $d_p$, $N_f$

Output: $T_{hcv}$

if ($N_s > 0$) then
  $T_{hcv} \leftarrow f_{mn}(N_f)$;
else
  $T_{hcv} \leftarrow f_{mn}(N_f)$;
end if

Algorithm 5 is used to compute total hash computation required for hybrid EDHT and EDHT-n version for the given height of it and number of siblings only in case of EDHT-n version.

Algorithm 6 $f_{mn}$

Input: $d_p$, $N_f$

Output: $f_{mn}$

if ($N_s > 0$) then
  $f_{mn} \leftarrow N_f / 100 \times d_p$;
else
  $f_{mn} \leftarrow N_f / 100 \times d_p$;
end if

Algorithm 6 is used to compute the minimum number of files required for the integrity checking.
temp ← maxr + 1;
f_aic ← c_aic;
else
maxr ← Nfc;
f_aic ← c_aic;
minr ← temp;
end if
end for
print f_aic" "+minr+"-" + maxr;

Algorithm 4 is used to determine the optimal hash tree methodology in dynamic way in order to store the cloud file meta-data, which is a primary aim for Adaptive Integrity Checking. We have used all the above algorithms to identify the cloud file range to fit into a specific Best, Worst and Average case possible hash tree methodology as well as to dynamically change the cloud file to store its meta-data in hash tree based on the error pattern, which were recorded periodically by the TPA system, in order to do the speedy recovery by reducing the hash computation process once we identify the file reside in cloud is faulty, subsequently it reduce space as well as time complexity.

V. RESULTS AND DISCUSSION

From Graph 1 to 6, it is clearly denotes the best, worst and average Hash Tree Methodologies has been generated dynamically to store cloud file meta-data for auditing purpose based on the repeated value of error percentage as well as average value of the error percentage noted in Table 3 for various number of cloud file storage, thus Adaptive Integrity Checking (AIC) would efficiently recover the file lost as well as auditing process with less number of hash computation and memory space compared to static data structure for storing the cloud storage meta-data.
VI. CONCLUSION

Cloud Storage is an Infrastructure as a Service (IaaS) which is a cost effective to an organizations with an integrity challenges, thus storage intactness to be checked periodically. In this paper we have done the research work to find the dynamic hash tree methodology to store the cloud file meta-data to perform the audit process for the cloud storage as well as speedy recovery process for the lost or corrupted file stored in the CSP to maintain integrity thru the efficient proposed EDHT-n and HEDHT data structure rather static way. Also, AIC batch processing has implemented to speed-up both the processes.

ACKNOWLEDGMENT

In this paper, we proposed Adaptive Integrity Checking based on these methodologies cloud storage integrity is achieved optimally by batch auditing and subsequently batch recovery in case, if file not intact in the cloud storage which has the remarkable future in Cloud Integrity. I solemnly thank my Guide and Computer Science and Engineering Department Professors and University for preparing and implementing this paper.

REFERENCES

1. A. Juels and J. S. Kaliski, “PORs: Proofs of retrievability for large files,” in Proc. 14th ACM Conf. Comput. Commun. Security, Alexandria, USA, 2007, pp. 584–597.
2. Chang Liu, Rajiv Ranjan, Chi Yang, Xuyun Zhang, Lizhe Wang, Senior Member, IEEE, and Jinjun Chen, Senior Member, IEEE, “MuR-DPA: Top-Down Levelled Multi-Replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud”, IEEE Trans. Computers, vol. 64, no. 9, Sep. 2015.
3. Wenting Shen, Jing Qin, Jia Yu, Rong Hao, and Jiankun Hu, Senior Member, IEEE, “Enabling Identity-Based Integrity Auditing and Data Sharing with Sensitive Information Hiding for Secure Cloud Storage”, IEEE Transactions on Information Forensics and Security, vol. 14, no. 2, Feb. 2019.
4. Yuan Zhang, Student Member, IEEE, Chunxiang Xu, Member, IEEE, Xiaohui Liang, Member, IEEE, Hongwei Li, Member, IEEE, Yi Mu, Senior Member, IEEE, and Xiaojun Zhang, “Entity Public Verification of Data Integrity for Cloud Storage Systems from Indistinguishability Obfuscation”, IEEE Transactions on Information Forensics and Security, vol. 12, no. 3, Mar. 2017.
5. Dan Gonzales, Member, IEEE, Jeremy Kaplan, Evan Saltzman, Zev Winkelman, Dulani Woods, “Cloud-Trust - a Security Assessment Model for Infrastructure as a Service (IaaS) Clouds”, IEEE Transactions on Cloud Computing, vol. 5, no 3, July-Sept. 1 2017.
6. G. Atteniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, “Provable data possession at untrusted storages,” in Proc. 14th ACM Conf. Comput. Commun. Secur., 2007, pp. 598–609.
7. Yujie Wang, Qianhong Wu, Member, IEEE, Bo Qin, Wenchang Shi, Robert H. Deng, Fellow, IEEE, Jiankun Hu, “Identity-Based Data Outsourcing with Comprehensive Auditing in Clouds”, IEEE Transactions on Information Forensics and Security, vol. 12, no. 4, Apr. 2017.
8. G. Atteniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, “Remote data checking using provable data possession,” ACM Trans. Inform. Syst. Secur., vol. 14, no. 1, pp. 1–34.
9. G. Atteniese, R. Di Pietro, L. V. Mancini, and G. Tsudik, “Scalable and efficient provable data possession,” in Proc. 4th Int. Conf. Secur. Privacy Commun. Netow., 2008, pp. 1–10.
10. Kaijing Xue, Pelin Hong, 2014, ‘A Dynamic Secure Group Sharing Framework in Public Cloud Computing’, IEEE Transactions on Cloud Computing, vol. 2, no. 4, p. 459-470.
11. Jinxi Wei, Ru Zhang, Junyi Liu, Jing Li, Xin Xin Niu, Yuanhang Yao, 2017, Dynamic data integrity auditing for secure outsourcing in the cloud’, Concurrency and Computation: Practice and Experience, vol. 29, no. 12, p. e4096.
12. C. Erway, A. Kupcu, C. Papamantou, and R. Tamassia, “Privacy-Preserving Public Auditing for Data,” in Proc. 16th ACM Conf. Comput. Commun. Secur., 2009, pp. 213–222.