International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Hybrid Cloud Approach for Secure Authorized Deduplication

Jagadish [1], Dr.Suvarna Nandyal[2]

[1]M.Tech.(CSE), Department of Computer Science & Engineering,
Poojya Doddappa Appa College of Engineering, Gulbarga, Karnataka, India

[2]Professor and HOD, Department of Computer Science & Engineering,
Poojya Doddappa Appa College of Engineering, Gulbarga, Karnataka, India

Abstract: ***Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this work makes the first attempt to formally address the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. The work also presents several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, the work implement a prototype of proposed authorized duplicate check scheme and conduct tested experiments using the prototype. The work shows that the proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.***

Keywords: Data deduplication, Convergent encryption, Confidentiality, Hybrid cloud, Authorized Duplicate check.

1. Introduction

Cloud computing enables new business models and cost effective resource usage. Instead of maintaining their own data center, companies can concentrate on their core business and purchase resources when it will needed. Especially when combining publicly accessible clouds with a privately maintained virtual infrastructure in a hybrid cloud, the hybrid cloud technology can open up new opportunities for businesses. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage.

Deduplication can take place at either the file level or the block level for file level deduplication, it eliminates duplicate copies of the same file. Traditional encryption, while providing data confidentiality is incompatible with data deduplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher texts making deduplication impossible. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a Convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same cipher text. A Hybrid Cloud is a combined form of private clouds and public clouds in which some critical data resides in the enterprise's private cloud while other data is stored in and accessible from a public cloud. Hybrid clouds seek to deliver the advantages of scalability, reliability, rapid deployment and potential cost savings of public clouds with the security and increased control and management of private clouds..

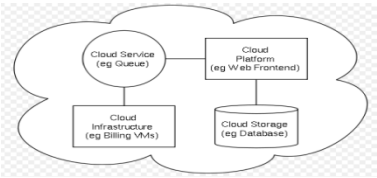

Figure 1.1: Architecture of Cloud Computing

The critical challenge of cloud storage or cloud computing is the management of the continuously increasing volume of data. Data deduplication or Single Instancing essentially refers to the elimination of redundant data. However, indexing of all data is still retained should that data ever be required. In general the data deduplication eliminates the duplicate copies of repeating data.

This paper is organized as follows, section 1 discusses the introduction, and section 2 describes related work. Section 3 details the system design and implementation. Section 4, presents the performance evaluations of our system design. Finally, section 5 presents some concluding remark.

2. Related Work

"A secure cloud backup system with assured deletion and version control. A. Rahumed", H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui [1],has presented Cloud storage is

an emerging service model that enables individuals and enterprises to outsource the storage of data backups to remote cloud providers at a low cost. Hence results shows that FadeVersion only adds minimal performance overhead over a traditional cloud backup service that does not support assured deletion. **"A reverse deduplication storage system optimized for reads to latest backups", C. Ng and P. Lee. Revdedup** [2] had present RevDedup, a de-duplication system designed for VM disk image backup in virtualization environments. RevDedup has several design goals: high storage efficiency, low memory usage, high backup performance, and high restore performance for latest backups. They extensively evaluate our RevDedup prototype using different workloads and validate our design goals. **"Role-based access controls", D. Ferraiolo and R. Kuhn** [3],has described the Mandatory Access Controls (MAC) are appropriate for multilevel secure military applications, Discretionary Access Controls (DAC) are often perceived as meeting the security processing needs of industry and civilian government. **"Secure deduplication with efficient and reliable convergent key management", J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou** [4], had proposed Dekey, an efficient and reliable convergent key management scheme for secure de-duplication. They implement Dekey using the Ramp secret sharing scheme and demonstrate that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular upload/download operations." **Reclaiming space from duplicate files in a server less distributed file system", J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.** [5],has presented the Farsite distributed file system provides availability by replicating each file onto multiple desktop computers. Measurement of over 500 desktop file systems shows that nearly half of all consumed space is occupied by duplicate files. The mechanism includes 1) convergent encryption, which enables duplicate files to coalesced into the space of a single file, even if the files are encrypted with different users' keys, and 2) SALAD, a Self Arranging, Lossy, Associative Database for aggregating file content and location information in a decentralized, scalable, fault-tolerant manner..**"A secure data deduplication scheme for cloud storage", J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl** [6],has provided the private users outsource their data to cloud storage providers, recent data breach incidents make end-toend encryption an increasingly prominent requirement data deduplication can be effective for popular data, whilst semantically secure encryption protects unpopular content. "**Weak leakage-resilient client-side deduplication of encrypted data in cloud storage", J. Xu, E.-C. Chang, and J. Zhou** [7],has described the secure client-side deduplication scheme, with the following advantages: our scheme protects data confidentiality (and *some* partial information) against both outside adversaries and honest-but-curious cloud storage server, while Halevi *et al.* trusts cloud storage server in data confidentiality. **"Secure and constant cost public cloud storage auditing with deduplication", J. Yuan and S. Yu**[8] has proposed, Data integrity and storage efficiency are two important requirements for cloud storage. The author proposed scheme is also characterized by constant realtime communication and computational cost on the user side. **"Privacy aware data intensive computing on hybrid clouds", K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan** [9] has proposed, the emergence of cost-effective cloud services offers organizations great opportunity to reduce their cost and increase productivity.The system, called Sedic, leverages the special features of Map Reduce to automatically partition a computing job according to the security levels of the data it works. **"Gq and schnorr identification schemes Proofs of security against impersonation under active and concurrent attacks", M. Bellare and A. Palacio**[10] has provided, the proof for GQ based on the assumed security of RSA under one more inversion, an extension of the usual onewayness assumption that was introduced. Both results extend to establish security against impersonation under concurrent attack**.**

3. Methodology

The basic objective of this work is the problem of privacy preserving deduplication in cloud computing and a proposed System focus on these aspects:

1) Differential Authorization**:** Each authorized user is able to get his/her individual token of his file to perform duplicate check based on his privileges.
2) Authorized Duplicate Check: Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

3.1 Proposed System

In Proposed system, Convergent encryption has been used to enforce data confidentiality. Data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP(storage cloud service provider). Security analysis demonstrates that that system is secure in terms of the definitions specified in the proposed security model.

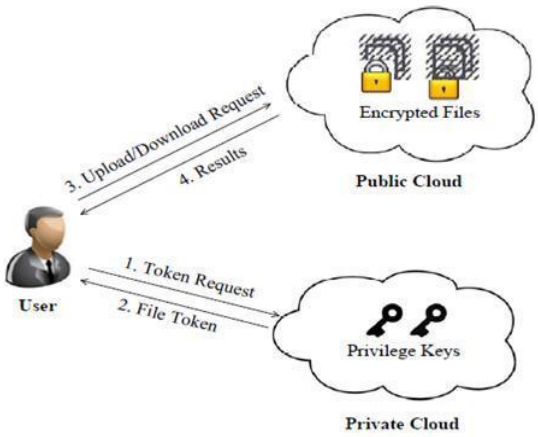

Figure 1: Architecture for Authorized Deduplication

This work describes a company by where the employee details such as name, password, email id, contact number and

designation is registered by admin or owner of the company based on his userid and password employees of the company able to perform operations such as file upload download and duplicate checks on the files based on his privileges.

There are three entities define in hybrid cloud architecture of authorized deduplication.

- **Data Users:** A user is an entity that wants to outsource data storage to the S-CSP(storage cloud service provider) and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.
- **Private Cloud:** This is new entity for facilitating users secure use of cloud services. The private keys for privileges are managed by private cloud, which provides the file token to users. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud.
- **S-CSP(storage cloud service provider):**This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power.

3.2 SHA1 Algorithm Description

In the proposed system convergent key for each file is generated by using secure hashing algorithm-1 the steps of this algorithm is given below

Step1: Padding

- Pad the message with a single one followed by zeroes until the final block has 448 bits.
- Append the size of the original message as an unsigned 64 bit integer.

Step2: Initialize the 5 hash blocks (h0,h1,h2,h3,h4) to the specific constants defined in the SHA1 standard.
Step3:Hash (for each 512bit Block)
○ Allocate an 80 word array for the message schedule
■ Set the first 16 words to be the 512bit block split into 16 words.
■ The rest of the words are generated using the following algorithm
step4: word[i3] XOR word[i8] XOR word[i14] XOR word[i16] then rotated 1 bit to the left.
○ Loop 80 times doing the following.
■ Calculate SHAfunction() and the constant K (these are based on the current round number.
■ e=d
■ d=c
■ c=b (rotated left 30)
■ b=a
■ a = a (rotated left 5) + SHAfunction() + e + k + word[i]
○ Add a,b,c,d and e to the hash output.
step5: Output the concatenation (h0,h1,h2,h3,h4) which is the message digest.

3.3: Flow Chart

The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users and this interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

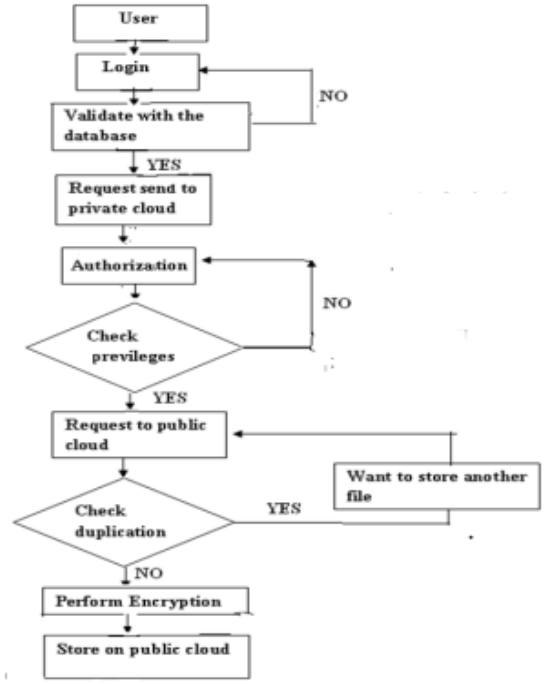

Figure 2: Flow Diagram of the proposed work.

In deduplication system, hybrid cloud architecture is introduced to solve the problem of unauthorized deduplication of file. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server. The user needs to send a request to the private cloud server to get a file token. The user needs to get the file token from the private cloud server to perform the duplicate check for some file. The user either uploads this file or prove their ownership based on the results of duplicate check. If it is passed, the private cloud server will find the corresponding privileges of the user from its stored table list and send to the user then user can upload his files. The same way user can download his file from storage cloud.

4. Results and Discussion

We conduct test based evaluation on our prototype. Our evaluation focuses on comparing the overhead induced by authorization steps, including file token generation and share token generation, against the convergent encryption and file upload steps. We evaluate the over- head by varying different factors.

Figure 3: Registration screen

In fig 3, shows the initial registration of a screen. The admin can add different employee informations. Thus, the Admin registering an employment as a director.

Figure 4: Successful registration

In fig4, shows a successful registration of an Employee of valid information.

Figure 5: Selection of team leader

In fig 5, after the getting a valid information from an employer. The admin selects a team leader.

Figure 6: Successful registration of a team leader.

Figure 7: Login as Director.

Figure8: Choosing a file.

In fig 8, shows every user can upload the files onto the cloud and also they give the access permissions to upload and download a file into cloud.

Figure 9: Access permission to team leader.

In fig 9, shows the access permission can be given to the various priorities like team leader, engineers etc. Later the file has upload into the Amazon cloud, and later the file fetches the required information from the hybrid cloud which contains i.e. employees name and all etc., later they upload the file. Hence the file gets stored and encrypted form an image is been generated.

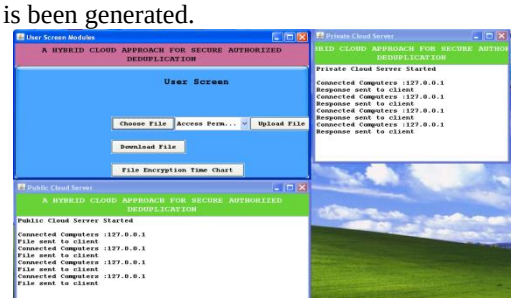

Figure 10: Set of private and public cloud server.

In the back end registered employees can be displayed and the token generated by private cloud for the files .If the same file is given to other user same token is generated by the private cloud and a tag is generated for the duplicate file. Unique files having no tags and it is represented as none. In this project work the time required to encrypt and to store the files in the amazon cloud is calculated and it is shown in the file encryption chart by taking the file name along x-axis and encryption time in milliseconds along y-axis.

If three files of different sizes such as 427kb,672kb and 2.15mb are uploaded to the cloud the files are stored in encrypted form in the amazon cloud and the time required to encrypt these files is based on network speed and it is 484ms,203ms and 453ms respectively for these files and the time is noted in the notepad and this can be shown in figure 11.

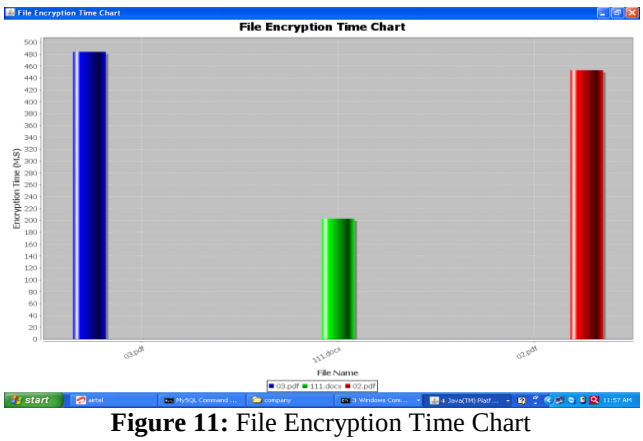

Figure 11: File Encryption Time Chart

5. Conclusion and Future Work

In this Project, the notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. In this project we perform several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. As a proof of concept in this project we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments on our prototype. From this project we show that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

Futures work: It excludes the security problems that may arise in the practical deployment of the present model. Also, it increases the national security. It saves the memory by deduplicating the data and thus provides us with sufficient memory. It provides authorization to the private firms and protects the confidentiality of the important data

References

[1] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In *Proc. of APSYS*,Apr 2013.
[2] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In *IEEE Transactions on Parallel and Distributed Systems*, 2013.
[3] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure datadeduplication scheme for cloud storage. In *Technical Report*, 2013.
[4] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In *ASIACCS*,pages 195–206, 2013.
[5] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. *IACR Cryptology ePrint Archive*, 2013:149, 2013.
[6] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacy aware data intensive computing on hybrid clouds. In *Proceedings of the 18th ACM conference on Computer and communications security*,CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.
[7] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In *USENIX Security Symposium*, 2013.
[8] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In *EUROCRYPT*, pages 296–312, 2013.
[9] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In *Proc. of StorageSS*, 2008.
[10] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In *Proc. of USENIX LISA*, 2010.
[11] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, *ACM Symposium on Information, Computer and Communications Security*, pages 81–82. ACM, 2012.
[12] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In *Workshop on Cryptography and Security in Clouds (WCSC2011)*, 2011.
[13] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 491–500. ACM, 2011.
[14] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 441–446. ACM, 2012.
[15] Z. Wilcox-O'Hearn and B. Warner. Tahoe: the least-authority filesystem. In *Proc. of ACM StorageSS*, 2008.