A Survey on: Secure Data Deduplication on Hybrid Cloud Storage Architecture

Aparna Ajit Patil
Dr. D.Y.Patil College of Engineering,Pune
Computer Department,Savitribai Phule Pune University, India.

Dhanashree Kulkarni
Assistant Prof. Dr D.Y Patil COE,
Computer Department,Savitribai Phule Pune University, India

ABSTRACT
Data deduplication is one of the most important Data compression techniques used for to removing the duplicate copies of repeating data and it is widely used in the cloud storage for the purpose of reduce the storage space and save bandwidth. To keep the confidentiality of sensitive data while supporting the deduplication, to encrypt the data before outsourcing convergent encryption technique has been proposed .To better protect data security, this project makes the first attempt to formally address the problem of authorized data deduplication .Different from the traditional deduplication system, differential benefits of the user are further considered the duplicate check besides the data itself. Hybrid cloud architecture contains several new deduplication constructions supporting authorized duplicate check. The proposed security models contain the demonstration of security analysis scheme. As a proof of concept, contains the implementation framework of proposed authorized duplicate check scheme and conduct testbed experiments using these prototype. In proposed system contain authorized duplicate check scheme incurs minimal overhead compared to normal operations.

General Terms
Convergent Encryption, Baseline Algorithm, Dekey Algorithm.

Keywords
Deduplication, authorized duplicate check, confidentiality, hybrid cloud, Proof of ownership.

1. INTRODUCTION
Cloud computing provides unlimited virtualized recourse to user as services across the whole internet while hiding the platform and implementing details. Cloud storage service is the management of evergreen increasing mass of data. To make data management scalable in cloud computing, deduplication has been a conventional technique. Data compression technique is used for eliminating the duplicate copies of repeated data in cloud storage to reduce the data duplication. This technique is used to improve storage utilization and also be applied to network data transfers to reduce the number of bytes that must be sent. Keeping multiple data copies with the similar content, deduplication eliminates redundant data by keeping only one physical copy and refer other redundant data to that copy. Data deduplication occurs file level as well as block level. The duplicate copies of identical file eliminate by file level deduplication .For the block level duplication which eliminates duplicates blocks of data that occur in non-identical files. Although data deduplication takes a lot of benefits, security as well as privacy concerns arise as users' sensitive data are capable to both insider and outsider attacks. In the traditional encryption providing data confidentiality, is contradictory with data deduplication. Traditional encryption requires different users to encrypt their data with own keys.

For making the feasible deduplication and maintain the data confidentiality used convergent encryption technique. It encrypts decrypts a data copy with a convergent key, the content of the data copy obtained by computing the cryptographic hash value of. After the data encryption and key generation process users retain the keys and send the ciphertext to the cloud. Since the encryption operation is determinative and is derived from the data content, similar data copies will generate the same convergent key and hence the same ciphertext. A secure proof of ownership protocol is used to prevent the unauthorized access and also provide the proof to user regarding the duplicate is found of the same file.

2. METHODS USED IN SECURE DEDUPLICATION
Following are the secure primitive used in the secure deduplication

2.1 Symmetric Encryption
Symmetric encryption uses a common secret key *k* to encrypt and decrypt information. A symmetric encryption scheme made up of three primary functions.

- KeyGen $_{SE}$ $(1^{\lambda})\rightarrow$ k is the key generation algorithm that generates k using security parameter 1^{λ};
- Enc $_{SE}$ (k, M)$\rightarrow$ C is the symmetric encryption algorithm that takes the secret k, and message M and then outputs the ciphertext C, and
- Dec $_{SE}$ (k, C) $\rightarrow$ *M* is the symmetric decryption algorithm that takes the secret *k* and ciphertext *C* and then outputs the original message *M*.

2.2 Convergent Encryption
Convergent encryption [5], provides data confidentiality in deduplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derive *tag* for the data copy, such that to detect duplicates tag will be used Here, we assume that the tag holds the property of correctness , i.e., if two data copies are the same, the tags of the data also same. The user first sends the tag to the server side to check if the identical copy has been already stored for detect duplicates.[4].

2.3 Proof of Ownership
The notion of proof of ownership (PoW) [11] enables users to prove their ownership of data copies to the storage server. Specifically, Proof of ownership is implemented as an interactive algorithm run by a user and a storage server.

2.4 Identification Protocol

The identification of protocol having two phases as follows:

1. Proof: The user can demonstrate his identity to a verifier by performing some identification proof related to his identity.

2. Verify: The verifier occurs verification with input of public information.

3. LITERATURE SURVEY

Following are the different methods which are used in secure data deduplication in cloud storage.

3.1 DupLESS Server-Aided Encryption for Deduplicated Storage

DupLess: Server aided encryption for deduplicated storage for cloud storage service provider like Mozy, Dropbox, and others perform deduplication to save space by only storing one copy of each file uploaded .Message lock encryption is used to resolve the problem of clients encrypt their file however the saving are lock. Dupless is used to provide secure deduplicated storage as well as storage resisting brute-force attacks. Clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol in dupless server. It allow clients to store encrypted data with an existing service, have the service occurs deduplication on their on the part , and yet achieves strong confidentiality guarantees. It show that encryption for deduplicated storage can successfully reach desired performance and space savings close to that of using the storage service with plaintext data [2].

Characteristic:

1. More Security .

2. Easily-deployed solution for encryption that supports deduplication

3.User Friendly: Use command-line client that supports both Dropbox and Google Drive.

4. Resolve the problem of message lock Encryption.

3.2 Proofs of Ownership in Remote Storage Systems

It stores only the single copy of the duplicate data. Client-side deduplication tries to identify deduplication chance already at the client and save the bandwidth of uploading copies of existing files to the server[11].To overcome the attacks Shai Halevi1, Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg proposes the Proof of ownership which lets a client efficiently prove to a server that that the client keep a file, rather than just some short information about it present solutions based on Merkle trees and specific encodings, and analyse their security.[9]

Characteristic:

1. To identify the attacks that exploit client-side deduplication..
2. Proofs of ownership provide the rigorous security.
3. Rigorous efficiency requirements of Peta-byte scale storage systems.

3.3 A Secure Deduplication with Efficient and Reliable Convergent Key Management

Data deduplication is a used for removing duplicate copies of data, and has been widely applied in cloud storage to reduce not only storage space but also upload banddwidth. Promising as it is, an appearing challenge is to accomplish secure deduplication in cloud storage. Although convergent encryption has been extensively acquired for secure deduplication, a uncertain issue of making convergent encryption practical is to efficiently and reliably manage a huge number of convergent keys.

Techniques:

1. Key management
2. Convergent Encryption[4]

3.4 Twin Clouds: An Architecture for Secure Cloud Computing

S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider proposed architecture for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. In come towards, the user communicates with a trusted cloud. Which encrypts as well as verifies the data stored and operations occurred in the untrusted cloud .It divide the computations such that the trusted cloud is used for security-critical operations in the less time-critical setup phase, whereas queries to the outsourced data are processed in parallel by the fast cloud on encrypted data [10].

3.5 Private Data Deduplication Protocols in Cloud Storage

Most important issue in the cloud storage is utilization of the storage capacity. In this paper, there are two categories of data deduplication strategy, and extend the fault-tolerant digital signature scheme proposed by Zhang on examining redundancy of blocks to achieve the data deduplication. The proposed scheme in this paper not only reduces the cloud storage capacity, but also improves the speed of data deduplication. Furthermore, the signature is computed for every uploaded file for verifying the integrity of files.[8]

Table 1: Comparison Of Different Methods Of Data Deduplication In Cloud Storage

Author	Method	Feature	Result
M. Bellare, S. Keelveedhi, and T. Ristenpart	DupLESS Server-Aided Encryption for Deduplicated Storage	• Space Saving • Resolve the cross user deduplication • Security: Provide strong security against External attacks. • High Performance	Simple storage Interface.

S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg	Proofs of Ownership in Remote Storage Systems	• Time Saving • Rigorous security • Identify attacks • Saving bandwidth	Performance measurements indicate that the scheme incurs only a small overhead compared to naive client-side deduplication.
J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou	A Secure deduplication with efficient and reliable convergent key management	• Reduce storage space & bandwidth • Efficient • Reliable key Management • Provide confidentiaalty	Convergent key share across multiple server .
S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider	Twin clouds: An architecture for secure cloud computing	• Secure computation • Store large amount of data • Low latency • Secure execution environment	Client uses the trusted Cloud as a proxy that provides a clearly defined interface to manage the outsourced data, programs, and queries.
W. K. Ng, Y. Wen, and H. Zhu	Private data deduplication Protocols in cloud storage	• Improve speed of data duplication • Fault tolerant • Reduce cloud storage capacity	Enhance the efficiency of data.

Following result is observed in Table1

- DupLESS Server-Aided Encryption for Deduplicated Storage is used for the simple storage interface and also provides the strong security against the external attacks like brute force attack. It provides high performance as well as resolves the cross user duplication.
- Proof of ownership presents the Performance measurements indicate that the scheme incurs only a small overhead compared to naive client-side deduplication. It identifies attacks and saving bandwidth.
- A Secure deduplication with efficient and reliable convergent key management for reduces the storage space and bandwidth. Convergent key share across multiple server.
- Twin clouds: An architecture for secure cloud computing contain Client which uses the trusted Cloud as a proxy that provides a clearly defined interface to manage the outsourced data, programs, and queries. It having low latency and also provide the secure execution environment.
- Private data deduplication Protocols in cloud storage Enhance the efficiency of data as well as Improve speed of data duplication.

4. CONCLUSION

In this survey article proposed the secure deduplication with the Help of token generation and Secure upload download it can assure the user about high data security and also avoid data deduplication. Security analysis determine that given schemes are secure in terms of insider as well as outsider attacks specified in the proposed security model. As a proof of concept, it executed a prototype of proposed authorized duplicate check scheme and conduct testbed experiments on given prototype. In this paper we have to provide the different techniques to reduce the deduplication in cloud storage and maintain the security

In future by using Cloud Service Provider (CSP) have significant resources to govern distributed cloud storage servers and to manage its database servers. It also provides virtual infrastructure to host application services. These services can be used by the client to manage his data stored in the cloud servers. The CSP provides a web interface for the client to store data into a set of cloud servers, which are running in a cooperated and distributed manner. In addition, the web interface is used by the users to retrieve, modify and restore data from the cloud, depending on their access rights. Moreover, the CSP relies on database servers to map client identities to their stored data identifiers and group identifiers.

5. REFERENCES

[1] OpenSSL Project. http://www.openssl.org/.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Server aided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[3] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication .IACR Cryptology ePrint Archive, 2013.

[4] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In *EUROCRYPT*, pages 296– 312, 2013.

[6] 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.

[7] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[8] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S Ossowski and P. 2012.

[9] 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 Security2012.

[10] 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 (WCSC 2011), 2011.

[11] 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.

[12] 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.

[13] A. Rahumed , H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui. A secure cloud backup system with assured deletion and version control. In 3rd International Workshop on Security in Cloud Computing, 2011.

[14] M. Bellare, C. Namprempre , and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 2009.

[15] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177,2002