RESEARCH ARTICLE OPEN ACCESS

Hybrid Cloud Approach for Reliable Authorized Deduplication

Shaik Shaju [1], V. V. Krishna Reddy [2] , D. V. Subbaiah[3]
M.Tech [1], Assistant professor [2], Associate Professor [3]
H.O.D of Computer Science [3]
Department of Computer Science and Engineering
Priyadarshini College of Engineering and Technology
Nellore
Andhra Pradesh – India

ABSTRACT
Data deduplication is certainly one of important data compression techniques for eliminate duplicate data of repeating information, and has been widespread in cloud storage devices to reduce the number of storage space and also save bandwidth. On the other hand, there have been recently wide privacy considerations as data could come in contact with those third party servers also to unauthorized parties. To declare the user control over access to their own data, it is a capable method to encrypt the data before outsourcing. To higher protect data safety, this paper makes the first effort to properly address the condition of authorized information deduplication. Different through traditional deduplication methods, the differential privileges of users are usually further considered throughout duplicate check in addition to the data itself. We also existing several new deduplication improvements supporting authorized duplicate sign on a hybrid cloud architecture. In this paper we are using network storage system called as venti which is used to reduce the duplication and also providing data security by using cryptography hash function.
Keywords:- Hybrid cloud, Confidentiality, Deduplication, Authorized duplication check.

I. INTRODUCTION

Now a days, there are many cloud storage systems that are widely used for safe file store, convenient file access and file synchronization. For example, Skydrive, Dropbox, Google drive and Ndrive is well-known cloud storage service. Recently, in a cloud storage system, data deduplication is actively used for reducing storage capacity and network bandwidth. In cloud storages, very few vendors only provide data deduplication technology. For example, Dropbox adapts VLC (Variable-Length Chunking) for processing data deduplication, so Dropbox can reduce network bandwidth when data transfers between client and server.

To make information managing scalable in cloud computing, data deduplication [1] has become a 2010 well-known technique and contains attracted a growing number of attention recently. Data deduplication is often a specific data compression method for eliminating duplicate duplicates of repeating files in storage. The technique is utilized to improve storage space utilization and may also be applied to system data transfers to reduce the amount of bytes that need to be sent. Instead of maintaining multiple data copies with the similar content, deduplication eliminates obsolete data by keeping just one physical copy and also referring other duplicate data to that copy. Data deduplication usually takes place at sometimes the file level or the block level. For filelevel data deduplication, it eliminates duplicate copies from the similar file. Data deduplication can also take place in the block level, which in turn eliminates duplicate hindrances of data that occur in non-identical documents. Although data deduplication brings a great deal of benefits, security and privacy concerns arise while users' sensitive data are susceptible to both inside and also outsider attacks. Classic encryption, while offering data confidentiality, is incompatible with data deduplication. Specifically, standard encryption requires diverse users to encrypt their data using own keys. Therefore, identical data illegal copies of different users will lead to different ciphertexts, creating deduplication impossible. Convergent encryption [2] continues to be proposed to put in force data confidentiality though making deduplication feasible. It encrypts/decrypts any data copy that has a convergent key, and that is obtained by calculating the cryptographic hash value from the content of your data copy. After important generation and data encryption, users support the keys and send out the ciphertext on the cloud. Since the encryption functioning is deterministic and hails from the data content material, identical data illegal copies will generate identical convergent key and as such the same ciphertext. To stop

unauthorized access, a safe proof of ownership protocol [3] can also be needed to provide proof that anyone indeed owns identical file when a duplicate is located. After the resistant, subsequent users using the same file is going to be provided a pointer on the server without wanting to upload the same file. A user can download the encrypted file using the pointer from this server, which can solely be decrypted because of the corresponding data owners using convergent keys. Therefore, convergent encryption permits the cloud to execute deduplication on the ciphertexts as well as the proof of property prevents the unauthorized user to access the file.

Here we are using Venti which is a network storage system that permanently stores data blocks. The venti which is used to reduce the deplication. A 160-bit SHA-1 hash of the data (called score by Ventiventi) acts because the address of the information. This enforces a write-once coverage since no different data block is available with the very same address: the addresses connected with multiple writes on the same data are generally identical, so duplicate data is easily identified as well as the data block is stored only once or twice. Data blocks cannot be removed, making it suitable for permanent or backup storage.

II. LITERATURE SURVEY

A. Secure Deduplication

Using the advent of cloud computing, secure information deduplication has fascinated much attention recently from research group. In [4] proposed a deduplication system within the cloud storage to scale back the storage size from the tags for integrity check. To boost the security of deduplication and protect the info confidentiality, In [5] showed the best way to protect the info confidentiality by transforming the predicatable information into unpredicatable information. In their method, another third gathering called key server is introduced to create the file label for duplicate verify. In [6] introduced a novel encryption scheme providing you with differential security pertaining to popular data as well as unpopular data. For popular data which can be not particularly sensitive, the conventional encryption is conducted. Another two-layered encryption structure with stronger safety while supporting deduplication is proposed for unpopular information. In this approach, they achieved better tradeoff between your efficiency and security from the outsourced data. Throughout [7] addressed the particular key management issue throughout block-level deduplication by means of distributing these tips across multiple machines after encrypting the particular files.

B. Proof of ownership

In [3] proposed the notion of "proofs of ownership" regarding deduplication systems, in a way that a client can efficiently encourage the cloud storage devices server that he/she are the owners of a folder without uploading the folder itself. Several Proof of ownership constructions based on the Merkle-Hash Tree usually are proposed [3] to enable client-side deduplication, including the bounded leakage setting. In [14] proposed another efficient proof of ownership structure by choosing the projection of the file onto many randomly selected bit-positions since the file proof. Note that most the above schemes usually do not consider data privateness. Recently, in [13] extended Proof of ownership regarding encrypted files, but they just don't address how to attenuate the key managing overhead.

C. Twin Clouds Architecture

Recently, in [16] provided an architecture comprising twin clouds with regard to secure outsourcing connected with data and arbitrary computations a great untrusted commodity cloud. In [12] moreover presented the hybrid cloud strategy to support privacy-aware information-intensive computing. In our perform, we consider to deal with the approved deduplication problem above data in public areas cloud. The security model of our systems resembles those related perform, where the personal cloud is assume in truth but curious.

III. HYBRID ARCHITECTURE FOR SECURE DEDUPLICATION

With a high level, our location on interest is definitely an enterprise network, consisting of a set of associated clients who'll use the S-CSP and store data together with deduplication technique. With this setting, deduplication is usually frequently used throughout these settings pertaining to information backup and failure recovery applications while greatly reducing safe-keeping. Such systems are widespread and they are often more ideal to user record backup and synchronization applications than richer hard drive abstractions. There are 3 entities defined within our system, that can be, users, private cloud and S-CSP in public areas cloud. The S-CSP does deduplication by checking when the contents of two files are the similar and stores only single of them. The access to a file can be defined based on some privileges. The exact definition of an privilege varies all over applications. Each privilege is represented available as a short communication called token. Each file is regarding some file tokens, which denote your tag with specific privileges. A user compute

and send duplicate-check tokens on the public cloud pertaining to authorized duplicate examine. Users have having access to the private foriegn server, a semitrusted third party which will help in performing deduplicable encryption simply by generating file tokens to the requesting users. We will reveal further the role from the private cloud server underneath. Users are additionally provisioned with per-user encryption important factors and credentials (e.x., user certificates). With this paper, we will only consider the file level deduplication pertaining to simplicity. In another word, we refer any data copy to become a whole file and file-level deduplication which eliminates the hard drive of any unnecessary files. Actually, block-level data deduplication is usually easily reduce from file-level data deduplication, which is comparable to [7]. Specifically, to be able to upload a record, a user primary performs the file-level replicate check. If the file is really a replica, then all its blocks must be duplicate as well; otherwise, the user additional performs the block-level replicate check and identifies the initial blocks to possibly be uploaded. Each facts copy is of a token for your duplicate check.

• ***S-CSP:*** This really is an entity providing you with a information storage service in public cloud. The S-CSP supplies the data outsourcing program and stores data on the part of the users. To reduce the storage price, the S-CSP reduces the storage associated with redundant data by means of deduplication and maintains only unique information. We assume that S-CSP is always online and features abundant storage capacity and computation electric power.

• ***Data User:*** A user can be an entity that really want to outsource data storage on the S-CSP and access the results later. In a storage system assisting deduplication, the user solely uploads single data but does not upload any replica data in order to save the upload bandwidth, which may be owned by the similar user or diverse users. In the actual authorized data deduplication program, each user is issued a few rights in the setup of the structure. Each record is secured using the convergent encryption critical and privilege keys to understand the authorized deduplication using differential privileges.

• ***Private Cloud:*** In comparison with the traditional deduplication structure in cloud calculating, this is a new entity introduced regarding facilitating user's secure using cloud service. Specially, since the calculating resources at info user/owner side are restricted and also the public cloud isn't fully trusted in practice, private cloud has the capacity to provide data user/owner with an execution environment and infrastructure working just as one interface between user and also the public cloud. The private keys to the privileges are managed because of the private cloud, who answers the actual file token requests from your users. The interface which is available from the private fog up allows user to be able to submit files and queries to become securely stored and computed respectively. Notice that that is a novel architecture regarding data deduplication throughout cloud computing, which consists of a twin clouds (i.e., the public cloud and also the private cloud).

Basically, this hybrid fog up setting has attracted a lot more attention recently. For instance, an enterprise might make use of a public cloud program, such as Amazon online S3, for aged data, but carry on and maintain in-house storage pertaining to operational customer info. Alternatively, the trusted private cloud is really a cluster of virtualized cryptographic co-processors, which are offered like a service by a third party and provide the essential hardware based security capabilities to implement a new remote execution environment trusted because of the users.

IV. A NEW APPROACH TO ARCHIVAL STORAGE

In this paper we describes the network storage called Venti. Each and every hash of block's contents act as block identifier. In this the duplicate copy and consumption of storage is reduced. Constructing the varieties of storage application building block is called Venti and this is also describes the design of an archival storage. The main goal of Venti is to provide the write–once archival repository and this is shared by the multi client machines and applications. The Venti is often a block-level network storage system intended for archival data. For storing a collected futes and directories as a single object the vac application is used with vac the contents of the selected data's are stored as a tree of blocks on a venti server. The Vic writes each files as a separate collection of venti block. So the writes–once model and duplicate copies of a block will makes venti a useful storage application.

V. CRYPTOGRAPHIC HASH FUNCTION

We all design and implement a new system which might protect the safety for predicatable communication. The main concept of our technique is how the novel encryption important generation algorithm. Regarding simplicity, we use the hash functions to define the particular tag generation functions and convergent keys in this particular section. In standard convergent encryption, to aid duplicate

check, the key is derived from the file F by employing some cryptographic hash operate $k_F = H(F)$. In order to avoid the deterministic important generation, the encryption important k_F for file F in this system will be generated with the private important cloud server using privilege key k_p. The encryption key can be viewed as the form associated with $k_{F,p} = H_0(H(F), k_p) \oplus H_2(F)$, exactly where H_0, H and H_2 are typical cryptographic hash functions. The file F is encrypted using another key k, while k will probably be encrypted with $k_{F,p}$. In this means, both the non-public cloud server along with S-CSP cannot decrypt the particular ciphertext.

VI. ASSESSMENT

Our evaluation concentrates on comparing the over head induced by acceptance steps. We evaluate the overhead by varying different facets, including. A) Number of Stored Files B) Dedpliucation Ratio C) File size. For every single step, we record the start and end time of computer and therefore find the breakdown of the whole time spent. We present the common time taken in each data from the figures.

A. Number of Stored Files

To evaluate the effect of quantity of stored files from the system, we upload 10000 10MB unique files towards system and history the breakdown for any file upload. Coming from Figure 1, every single step remains constant down the time. Token checking is finished with a hash table as well as a linear search can be executed regarding collision.

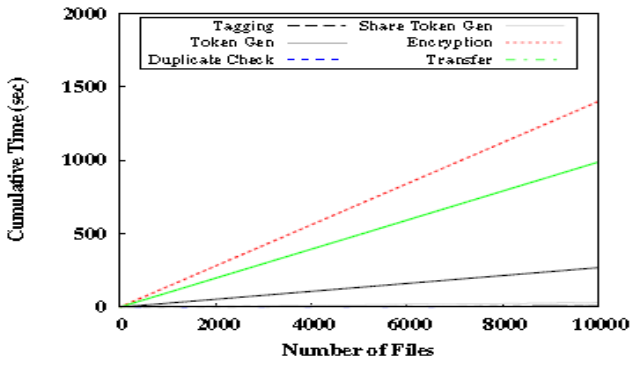

Fig. 1. Time Breakdown for Different Number of Stored Files

B. Deduplication Ratio

To evaluate the effect from the deduplication ratio, all of us prepare two special data sets, all which consists connected with 50 100MB documents. We first upload the primary set as a first upload. For the next upload, we choose portion of 50 documents, according to the particular given deduplication relation, from the preliminary set as identical files and remaining files in the second set because unique files. The common time of uploading the next set is displayed in Figure 2. Because uploading and encryption would be skipped in the case of duplicate files, some time spent on both of which decreases with increasing deduplication ratio. The time spent on identical check also decreases as the searching would always be ended when duplicate is available. Total time used on uploading the file with deduplication relation at 100% should be only 33. 5% having unique files.

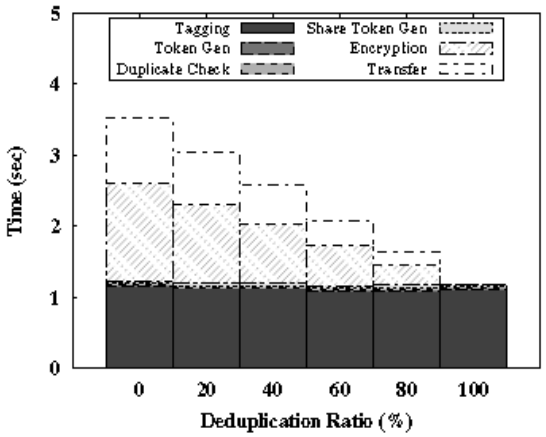

Fig. 2. Time Breakdown for Different Deduplication Ratio

C. *File Size*

To judge the effect of file size to the time used on different steps, we all upload 100 distinct files (i.e., without any deduplication opportunity) of particular file size and record the time break down. With all the distinctive files enables us to judge the worst-case scenario where we've got to upload almost all file data. The standard time of the steps from examination sets of different file size are plotted along with Figure 3. Any time spent on marking, encryption, upload increases linearly using the file size, since these operations involve the complete file data together with incur file I/O using the whole file.

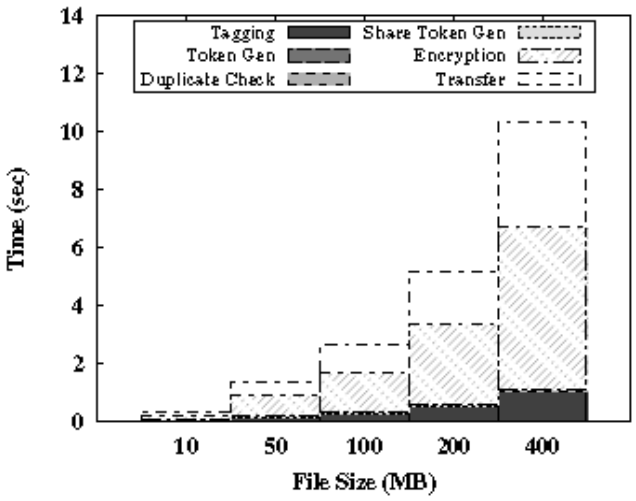

Fig. 3. Time Breakdown for Different File Size

VII. CONCLUSION

In this paper, the investigation is founded on the notion of authorized data deduplication was proposed to guard the data protection by including differential rights of users inside duplicate check. We also mentioned several new deduplication buildings that supporting authorized duplicate register hybrid cloud structure, in which the particular duplicate-check tokens of files are generated with the private cloud server having private keys. Security analysis demonstrates our schemes are secure with regard to insider and outsider attacks specified inside proposed security model. We also planned to further improve the deduplication tactic by assigning the particular integrity check to be able to highly authorized individuals.

ACKNOWLEDGEMENT

This work was supported by **V.V.Krishna Reddy**, M.Tech. Assistant Professor, Department of Computer Science and Engineering and **D.V. Subbaiah**, M.Tech. Associate Professor, H.O.D of Computer Science and Engineering by their valuable guidance, constant encouragement, and keen interest.

REFERENCES

[1] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In *Proc. USENIX FAST*, Jan 2002.

[2] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In *ICDCS*, pages 617–624, 2002.

[3] 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.

[4] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. *IACR Cryptology ePrint Archive*, 2013:149, 2013.

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In *USENIX Security Symposium*, 2013.

[6] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In *Technical Report*, 2013.

[7] 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.

[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] Z. Wilcox-O'Hearn and B. Warner. Tahoe: the least-authority filesystem. In *Proc. of ACM StorageSS*, 2008.

[11] 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.

[12] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacyaware 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] 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.

[14] 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.

[15] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In *Proc. of USENIX LISA*, 2010.

[16] 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.