ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Hybrid Cloud Approach for Secure Authorized Deduplication

Ponnaluru Surendra
P.G. Scholar (M. Tech),
Department of CSE,
Srinivasa Institute of Technology & Sciences,
Ukkayapalli, Kadapa, Andhra Pradesh.

K.Rajasekhar Reddy
Assistant Professor,
Department of CSE,
Srinivasa Institute of Technology & Sciences,
Ukkayapalli, Kadapa, Andhra Pradesh.

ABSTRACT:

Data de duplication 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 de duplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data de duplication. Different from traditional de duplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new deduplication constructions supporting authorized duplicate check in a 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, we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

Index Terms:

Deduplication, authorized duplicate check, confidentiality, hybrid cloud.

1. INTRODUCTION:

Cloud computing provides many “virtualized” resources to users as services across the entire Internet, while hiding platform and implementation details. Nowadays cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. GMAIL is one of the best examples of cloud storage which is used by most of us regularly. One of the major issues of cloud storage services is the management of the ever-increasing volume of data.

To make data management scalable in cloud computing, deduplication [5] has been a well-known technique which is being used by most of the users. Data Deduplication is one of the specialised data compression techniques which is used to eliminate duplicate copies of data. Deduplication can take place at file level or either block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files. Although there are several advantages of data deduplication security and privacy concerns arise as users’ sensitive data are susceptible to both insider and outsider attacks.

Encryption techniques which were used traditionally were not compatible with data deduplication while providing data confidentiality. Traditional encryption requires different users to encrypt their data with their own keys by which identical data copies of different users will lead to different cipher texts, making deduplication impossible. Convergent encryption [4] 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. Whenever the key is generated users retain the keys and send the cipher text to the cloud. In order to prevent unauthorized access, a secure proof of ownership protocol [2] is also needed to provide the proof that the user indeed owns the same file when a duplicate is found.

Hence convergent encryption allows the cloud to perform deduplication on the ciphertexts and the proof of ownership prevents the unauthorized user to access the file. Traditional deduplication systems based on convergent encryption, although providing confidentiality to some extent; do not support the duplicate check with differential privileges. Contradiction occurs when we try to realize both deduplication and differential authorization duplicate check at the same time.

II. LITERATURE SURVEY :

In archival storage systems, there is a huge amount of duplicate data or redundant data, which occupy significant extra equipments and power consumptions, largely lowering down resources utilization (such as the network bandwidth and storage) and imposing extra burden on management as the scale increases. So data de-duplication, the goal of which is to minimize the duplicate data in the inter level, has been receiving broad attention both in academic and industry in recent years. In this paper, semantic data de-duplication (SDD) is proposed, which makes use of the semantic information in the I/O path (such as file type, file format, application hints and system metadata) of the archival files to direct the dividing a file into semantic chunks (SC). While the main goal of SDD is to maximally reduce the inter file level duplications, directly storing variable SCes into disks will result in a lot of fragments and involve a high percentage of random disk accesses, which is very inefficient. So an efficient data storage scheme is also designed and implemented: SCes are further packaged into fixed sized Objects, which are actually the storage units in the storage devices, so as to speed up the I/O performance as well as ease the data management. Primary experiments have demonstrated that SDD can further reduce the storage space compared with current methods .. With the advent of cloud computing, secure data deduplication has attracted much attention recently from research community.

Yuan et al. proposed a deduplication system in the cloud storage to reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality, Bellare et al. showed how to protect the data confidentiality by transforming the predicatable message into unpredicatable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check. Stanek et al. presented a novel encryption scheme that provides the essential security for popular data and unpopular data. For popular data that are not particularly sensitive, the traditional conventional encryption is performed. Another two-layered encryption scheme with stronger security while supporting deduplication is proposed for unpopular data. In this way, they achieved better trade between the efficiency and security of the out-sourced data. Liet al. addressed the key management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files.

EXISTING SYSTEM:

» Data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges.
» Such architecture is practical and has attracted much attention from researchers.
» The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

DISADVANTAGES OF EXISTING SYSTEM:

» Traditional encryption, while providing data confidentiality, is incompatible with data deduplication.
» Identical data copies of different users will lead to different ciphertexts, making deduplication impossible.

III. OVERVIEWOF THE HYBRID CLOUD CONCEPTS HYBRID CLOUD :

A hybrid cloud is a cloud computing environment in which an organization provides and manages some resources in-house and has others provided externally .For example, an organization might use a public cloud service, such as Amazon Simple Storage Service(Amazon S3) for archived data but continue to maintain in house storage for operational customer data The concept of a hybrid cloud is meant to bridge the gap between high control, high cost “private cloud” and highly callable , flexible , low cost “public cloud”. “Private Cloud” is normally used to describe a VMware deployment in which the hardware and software of the environment is used and managed by a single entity.

The concept of a “Public cloud” usually involves some form of elastic/subscription based resource pools in a hosting provider datacenter that utilizes multi-tenancy. The term public cloud doesn’t mean less security, but instead refers to multi-tenancy. The concept revolves heavily around connectivity and data portability. The use cases are numerous: resource burst-ability for seasonal demand, development and testing on a uniform platform without consuming local resources, disaster recovery, and of course excess capacity to make better use of or free up local consumption. VMware has a key tool for “hybrid cloud” use called “vCloud connector”.

It is afree plugin that allows the management of public and private clouds within the vSphere client. The tool offers users the ability to manage the console view, power status, and more from a "workloads" tab, and offers the ability to copy virtual machine templates to and from a remote public cloud offering.

IV. HYBRID CLOUD FOR SECURE DEDUPLICATION :

At a high level, our setting of interest is an enterprise network, consisting of a group of affiliated clients (for example, employees of a company) who will use the S-CSP and store data with deduplication technique. In this setting, deduplication can be frequently used in these settings for data backup and disaster recovery applications while greatly reducing storage space. Such systems are widespread and are often more suitable to user file backup and synchronization applications than richer storage abstractions. There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud . The S-CSP performs deduplication by checking if the contents of two files are the same and stores only one of them. The access right to a file is defined based on a set of privileges.

The exact definition of a privilege varies across applications. For example, we may define a rolebased privilege according to job positions (e.g., Director, Project Lead, and Engineer), or we may define a time-based privilege that specifies a valid time period (e.g., 2014-01-01 to 2014-01-31) within which a file can be accessed. A user, say Alice, may be assigned two privileges "Director" and "access right valid on 2014- 01-01", so that she can access any file whose access role is "Director" and accessible time period covers 2014-01- 01.

Each privilege is represented in the form of a short message called token. Each file is associated with some file tokens, which denote the tag with specified.A user computes and sends duplicate-check tokens to the public cloud for authorized duplicate check. Users have access to the private cloud server, a semitrusted third party which will aid in performing deduplicable encryption by generating file tokens for the requesting users. We will explain further the role of the private cloud server below. Users are also provisioned with per-user encryption keys and credentials

V. PROPOSED SYSTEM:

In this paper, we enhance our system in security. Specifically, we present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that our system is secure in terms of the definitions specified in the proposed security model.

SYSTEM ARCHITECTURE:

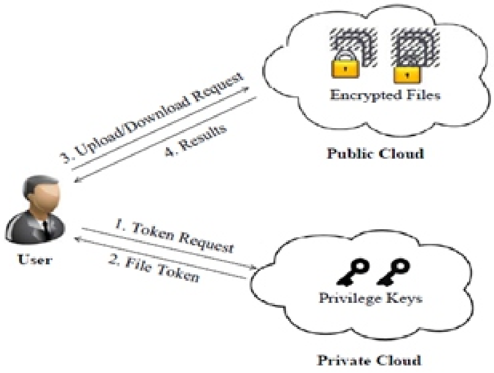


Figure 1 Architecture for Authorized de duplication.

In this paper, we will only consider the file level deduplication for simplicity. In another word, we refer a data copy to be a whole file and file-level deduplication which eliminates the storage of any redundant files. Actually, block-level deduplication can be easily deduced from file-level deduplication, Specifically, to upload a file, a user first performs the file-level duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.

• S-CSP. 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.

• Data Users. A user is an entity that wants to outsource data storage to the S-CSP 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. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

• Private Cloud. Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service. 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. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

Notice that this is a novel architecture for data deduplication in cloud computing, which consists of a twin clouds (i.e., the public cloud and the private cloud). Actually, this hybrid cloud setting has attracted more and more attention recently. For example, an enterprise might use a public cloud service, such as Amazon S3, for archived data, but continue to maintain in-house storage for operational customer data. Alternatively, the trusted private cloud could be a cluster of virtualized cryptographic co-processors, which are offered as a service by a third party and provide the necessary hardware based security features to implement a remote execution environment trusted by the users.

ADVANTAGES OF PROPOSED SYSTEM:

» The user is only allowed to perform the duplicate check for files marked with the corresponding privileges.
» We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys.
» Reduce the storage size of the tags for integrity check. To enhance the security of de duplication and protect the data confidentiality.

SECURE DEDUPLICATION SYSTEMS:

MAIN IDEA. o support authorized deduplication, the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, we call it file token instead. To support authorized access, a secret key kp will be bounded with a privilege p to generate a file token. Let ϕ′ F;p = TagGen(F, kp) denote the token of F that is only allowed to access by user with privilege p. In another word, the token ϕ′ F;p could only be computed by the users with privilege p. As a result, if a file has been uploaded by a user with a duplicate token ϕ′ F;p, then a duplicate check sent from another user will be successful if and only if he also has the file F and privilege p. Such a token generation function could be easily implemented as H(F, kp), where H(_) denotes a cryptographic hash function.

4.1 A First Attempt:

\
Before introducing our construction of differential deduplication, we present a straightforward attempt with the technique of token generation TagGen(F, kp) above to design such a deduplication system. The main idea of this basic construction is to issue corresponding privilege keys to each user, who will compute the file tokens and perform the duplicate check based on the privilege keys and files. In more details, suppose that there are N users in the system and the privileges in the universe is defined as P = fp1, . . . , psg. For each privilege p in P, a private key kp will be selected. For a user U with a set of privileges PU, he will be assigned the set of keys fkpi gpiPU .File Uploading:Suppose that a data owner U with privilege set PU wants to upload and share a file F with users who have the privilege set PF = fpjg. The user computes and sends S-CSP the file token ϕ′ F;p = TagGen(F, kp) for all p 2 PF .

• If a duplicate is found by the S-CSP, the user proceeds proof of ownership of this file with the S-CSP. If the proof is passed, the user will be assigned a pointer, which allows him to access the file.

• Otherwise, if no duplicate is found, the user computes the encrypted file CF = EncCE(kF , F) with the convergent key kF = KeyGenCE(F) and uploads (CF , fϕ′ F;p g) to the cloud server. The convergent key kF is stored by the user locally.

File Retrieving:Suppose a user wants to download a file F. It first sends a request and the file name to the S-CSP. Upon receiving the request and file name, the S-CSP will check whether the user is eligible to download F. If failed, the S-CSP sends back an abort signal to the user to indicate the download failure. Otherwise, the S-CSP returns the corresponding ciphertext CF .

Upon receiving the encrypted data from the S-CSP, the user uses the key kF stored locally to recover the original €file F. Problems. Such a construction of authorized deduplication has several serious security problems, which are listed below.

First, each user will be issued private keys Fkpi gpiPU for their corresponding privileges, denoted by PU in our above construction. These private keys fkpi gpiPU can be applied by the user to generate file token for duplicate check. However, during file uploading, the user needs to compute file tokens for sharing with other users with privileges PF .

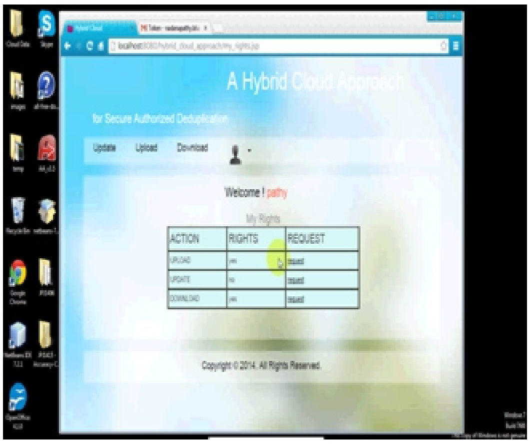
To compute these file tokens, the user needs to know the private keys for PF , which means PF could only be chosen from PU. Such a restriction makes the authorized deduplication system unable to be widely used and limited.

• Second, the above deduplication system cannot prevent the privilege private key sharing among users. The users will be issued the same private key for the same privilege in the construction. As a result, the users may collude and generate privilege private keys for a new privilege set P that does not belong to any of the colluded user.

For example, a user with privilege set PU1 may collude with another user with privilege set PU2 to get a privilege set P =PU1 [PU2. • The construction is inherently subject to brute-force attacks that can recover files falling into a known set.

That is, the deduplication system cannot protect the security of predictable files. One of critical reasons is that the traditional convergent encryption system can only protect the semantic security of unpredictable files.

VI. RESULT:


A Hybrid Cloud Approach

for Secure Authorized Deduplication

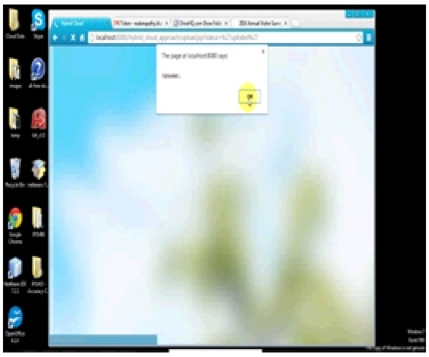
Welcome ! pathy

My Rights

ACTION	RIGHTS	REQUEST
UPLOAD	yes	issued
UPDATE	no	issued
DOWNLOAD	yes	issued

Copyright © 2014. All Rights Reserved.

Figure 2


The page at localhost:8080 says:

Uploaded..

Figure 3.

VII. CONCLUSION:

The notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented 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 serve with private keys.

Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transefer.

VI. FUTURE SCOPE :

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 provide us with sufficient memory. It provides authorization to the private firms and protect the confidentiality of the important data.

REFERENCES :

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

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

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

[4] 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.

[5] 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.

[6] Bugiel, S., N¨urnberger, S., Sadeghi, A.-R., Schneider, T.: Twin Clouds: An architecture for secure cloud computing (Extended Abstract). In: Workshop on Cryptography and Security in Clouds (WCSC 2011), March 15-16 (2011)

[7] Chung, K.-M., Kalai, Y., Vadhan, S.: Improved delegation of computation using fully homomorphic encryption. In: Rabin, T. (ed.) CRYPTO 2010. LNCS, vol. 6223, pp. 483–501. Springer, Heidelberg (2010)

[8] Cloud Security Alliance. Top threats to cloud computing, v. 1.0 (2010)