A Hybrid Cloud Appeal for Improving Storage Utilization for Secure Authorized Deduplication

B. Sowmya, N.Vaishnavi, B. Nayomi, Ms.P.S.Latha Kalyampudi Assistant Professor, Dept.of IT,

BVRIT HYDERABAD College of Engineering for Women

Abstract - *Storing the same data in the cloud raises compliance challenge such as duplication. To overcome this challenge deduplication concept is used. Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. In most organizations, the storage systems contain duplicate copies of many pieces of data. For example, the same file may be saved in several different places by different users, or two or more files that aren't identical may still include much of the same data. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. Companies frequently use deduplication in backup and disaster recovery applications, but it can be used to free up space in primary storage as well. To avoid this duplication of data and to maintain the confidentiality in the cloud we using the concept of Hybrid cloud. 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 paper makes the first attempt to formally address the problem of authorized data deduplication.*

1. NTRODUCTION

Cloud computing has been envisioned as the next generation architecture of IT enterprise. It is defined as a type of emerging computing technology that relies on sharing computer resources over the network. 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. Nowadays 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 the data shared by different 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 on cloud. To make the data management scalable in cloud computing, deduplication [2] has been a well-known technique recently use. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes. Instead of keeping multiple data copies with same content, deduplication eliminates the redundant data by keeping only one physical copy and referring other redundant data to that copy. Data deduplication brings a lot of benefits, though security and privacy concerns arise as users sensitive data are susceptible to both insider and outsider attacks.

To avoid this duplication of data and to maintain the confidentiality in the cloud we are using the concept of Hybrid cloud[1]. It is a combination of public and private cloud.

Hybrid cloud storage combines the advantages of scalability, reliability, rapid deployment and potential cost savings of public cloud storage with the security and full control of private cloud storage. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts and decrypts the data copy with a convergent key and which is obtained by computing the cryptographic hash value of the content of the data copy. After the key generation and data encryption, users retain the keys and send ciphertext to the cloud .Since the encryption operation is deterministic and it is derived from the data content, the same convergent key is generated by identical copies and hence the same ciphertext.

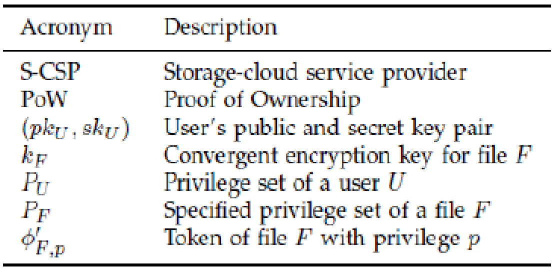
To prevent unauthorized access, a secure proof of ownership protocol is also needed to provides the proof that the user indeed owns the same file when the file duplicate is found and After the proof, subsequent users with the same file provide a pointer from the server without needing to upload the same file. A user can able to download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, Convergent encryption will allow the cloud to perform deduplication on the ciphertexts and the proof of ownership prevents the unauthorized user to access the file.

However, the previous deduplication systems cannot support to differential authorization and duplicate check, which is very important in many applications. In an authorized deduplication system each user is issued a set of privileges during system initialization. Each file uploaded to the cloud is bounded by a set of privileges which specify which kind of users is allowed to perform the duplicate check and access right of the files from the cloud storage.

1.1 CONTRIBUTIONS

In this paper, aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, we consider hybrid cloud which is a combination of pubic cloud and private cloud and Hybrid clouds seek to deliver the advantages of scalability, reliability, control and management of private clouds, rapid deployment and potential cost savings of public clouds with the security and increased where unlike existing system private cloud is involved as a proxy which allow data users to securely perform the duplicate check. Such an architecture is practical and has attracted much attention from researchers. The data owners not only outsource their data storage by utilizing public cloud while the data operation is managed by the private cloud. A new deduplication system supporting differential duplicate check is proposed under this hybrid cloud architecture where the cloud service provider resides in the public cloud. The user is only allowed to perform the duplicate check for the files marked with the corresponding privileges. We present an advanced scheme to support stronger security by encrypting the file with the different privileges. Furthermore, any unauthorized users cannot decrypt the cipher text even they collude with the cloud service provider CSP.

Table1. Notations Used in This Paper



Acronym	Description
S-CSP	Storage-cloud service provider
PoW	Proof of Ownership
(pk_U, sk_U)	User's public and secret key pair
k_F	Convergent encryption key for file F
P_U	Privilege set of a user U
P_F	Specified privilege set of a file F
$\phi'_{F,p}$	Token of file F with privilege p

2. SECURE PRIMITIVES

In this section, we first define the notations used in this paper, review some secure primitives used in our secure deduplication.

Symmetric encryption: Symmetric encryption uses a com- mon secret key k to encrypt and decrypt information. A symmetric encryption scheme consists of three primitive functions:

- **KeyGen$_{SE}$ (1^k)** is the key generation algorithm that generates k using security parameter 1 ;

- **Enc$_{SE}$ (k, M) - C** is the symmetric encryption algorithm that takes the secret k and message M and then outputs the ciphertext C;

- **Dec$_{SE}$ (k., C)** - M is the symmetric decryption algorithm that takes the secret k and ciphertext C and then outputs the original message M.

Convergent encryption: Convergent encryption [4],[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 derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property holds [4], i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived, and the tag cannot be used to deduce the convergent key and compromise data confidentiality.

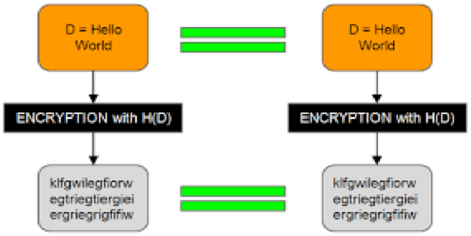

D = Hello World
D = Hello World
ENCRYPTION with H(D)
ENCRYPTION with H(D)
klfgwilegfiorw egtriegtiergiei ergriegrigfifiw
klfgwilegfiorw egtriegtiergiei ergriegrigfifiw

Fig 2.1 Convergent Encryption

Both the encrypted data copy and its corresponding tag will be stored in the server side. Convergent encryption scheme can be defined with four primitives.

- **KeyGenCE (M)-> K** is the key generation algorithm that maps a data copy M to a convergent key K_F;
- **EncCE (K,M)-> C** is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertext C;
- **DecCE (K,C)-> M** is the decryption algorithm that takes both the ciphertext and the convergent key as inputs and then outputs the original data copy.
- **TagGen (M)-> T(M)** is the tag generation algorithm that maps the original data copy M and outputs a tag T(M).

Proof of Ownership: The notion of proof of ownership (PoW) [3] enables users to prove their ownership of data copies to the storage server. Specifically, PoW is implemented as an interactive algorithm (denoted by PoW) run by a user and a verifier (i.e., storage server). The verifier derives a short value (M) from a data copy M. To prove the ownership of the data copy M, the user needs to send′ to the verifier such that $\emptyset' = \emptyset(M)$. The formal security definition for PoW roughly follows the threat model in a content distribution network, where an attacker does not know the entire file, but has accomplices who have the file. The accomplices follow the "bounded retrieval model", such that they can help the attacker obtain the file, subject to the constraint that they must send fewer bits than the initial min-entropy of the file to the attacker [3].

Identification protocol: An identification protocol P can be described with two phases: Proof and Verify. In the stage of Proof, a user U can demonstrate his identity to a verifier by performing some identification proof related to his identity. The input of the user is his private key sk_U that is sensitive information such as private key of a public key in his certificate or credit card number, etc. that he would not like to share with the other users. The verifier performs the verification with input of public information pk_U related to sk_U. At the conclusion of the protocol, the verifier outputs either accept or reject to denote whether the proof is passed or not. There are many efficient identification protocols in literature, including certificate-based, identity-based identification etc. [6], [7].

3 SYSTEM MODEL

3.1 HYBRID CLOUD ARCHITECTURE

Hybrid Cloud: It is a mixture of two or more clouds (private, community or public) that remain different entities but their exists a togetherness in between them which offering the benefits of multiple deployment models .in the Hybrid cloud it has the ability to manage dedicated services via cloud services and to connect the collocation. hybrid cloud services is cloud computing services that includes the combination of private, public and community cloud services, from different service providers. A hybrid cloud service crosses isolation and provider boundaries so that it can't be simply put in one category of private, public, or community cloud service. It allows one to extend either the capacity or the capability of a cloud service, by aggregation, integration or customization with another cloud service. Varied use cases for hybrid cloud composition exist. For example, an organization may store sensitive client data in house on a private cloud application, but interconnect that application to a business intelligence application provided on a public cloud as a software service. is example of hybrid cloud extends the capabilities of the enterprise to deliver a specific business service through the addition of externally available public cloud services. Another example of hybrid cloud is one where IT organizations use public cloud computing resources to meet temporary capacity needs that cannot be met by the private cloud.

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.

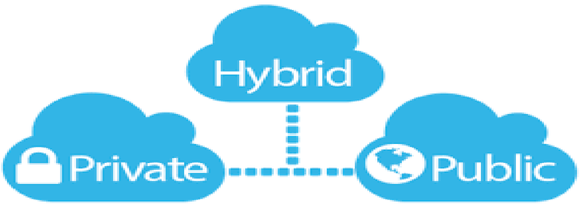


Figure. 3.1 Hybrid Cloud Appeal

3.2 DATA DE-DUPLICATION

In the current digital world, data is of prime importance for individuals as well as for organizations. As the amount of data being generated increases exponentially with time, duplicate data contents being stored cannot be tolerated. Thus, employing storage optimization techniques is an essential requirement to large storage areas like cloud storage. storage. Deduplication is a one such storage optimization technique that avoids storing duplicate copies of data.

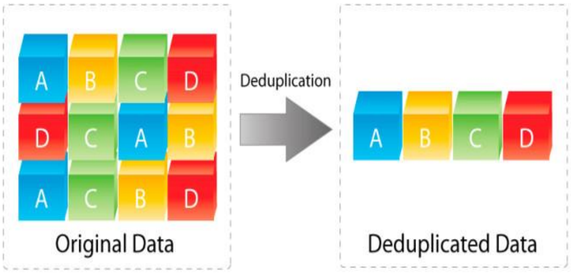


Figure 3.2 Data Deduplication

Data deduplication or Single Instancing technique is used for eliminating the duplicate copies of same data and reduce storage overhead. Deduplication is basically a compression technique for removing redundant data. Data deduplication techniques ensure that only one unique instance of data is retained on storage media, such as disk, flash or tape. It takes place on file level and block level. In file level approach duplicate files are eliminate, and in block level approach duplicate blocks of data that occur in non-identical files. But block level deduplication frees up more space than the file level deduplication. Deduplication reduces the storage needs by up to 90-95% for backup application, 68% in standard file system. Important issues in data deduplication that security and privacy to protect the data from insider or outsider attack. Deduplication is widely is used various applications like backup, metadata management, primary storage, etc. for storage optimization [10].

4. SYSTEM ARCHITECTURE

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[4]. 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.

Public Cloud: Public cloud entity is used for the storage purpose. User upload the files in public cloud. Public cloud is similar as S-CSP. When the user want to download the files from public cloud, it will be ask the key which is generated or stored in private cloud. When the users key is matching with files key at that time user can download the file, without key user cannot access the file. Only authorized user can access the file. In public cloud all files are stored in encrypted format. If by chance any unauthorized person hack the file, but without the secret or convergent key he/she doesn't access original file. On public cloud there are lots of files are stored each user access its respective file if its token matches with S-CSP server token.

First if a user wants to upload the files on the public cloud then user first encrypt that file with the convergent key and then sends it to the public cloud at the same time user also generates the key for that file and sends that key to the private cloud for the purpose of security. In the public cloud we use one algorithm for deduplication. Which is used to avoid the duplicate copies of files which is entered in the public cloud. Hence it also minimizes the bandwidth that means it requires the less storage space for storing the files on the public cloud. In the public cloud any person that means the unauthorized person can also access or store the data so concluding that in the public cloud the security is not provided. In general for providing more security user can use the private cloud instead of using the public cloud.

User generates the key at the time of uploading file and store it to the private cloud. When user wants to downloads the file that he/she upload, The user sends the request to the public cloud. Public cloud provides the list of files that are uploads the many user of the public cloud because there is no security is provided in the public cloud. When user selects one of the file from the list of files then private cloud sends a message like enter the key!. User has to enter the key that he generated for that file. When user enter the key the private cloud checks the key for that file and if the key is correct that means user is valid then private cloud give access to that user to download that file successfully then user downloads the file from the public cloud and decrypt that file by using the same convergent key which is used at the time of encrypt that file in this way user can make a use of the architecture. Users have access to the private cloud server, a semi trusted third party which will aid in performing duplicable encryption by generating file tokens for the requesting users.

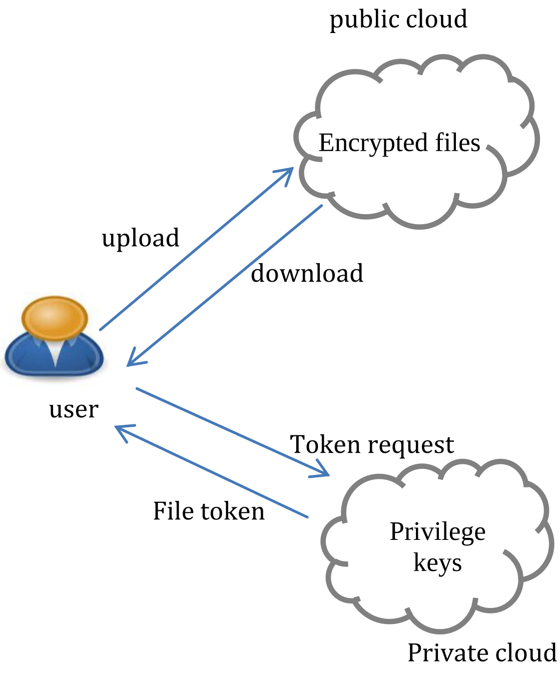


Figure. 4.1 System Architecture

Users are also provisioned with per-user encryption keys and credentials (e.g., user certificates). 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.

5. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system. We implement a prototype of the proposed authorized deduplication system, in which we model three entities. A Client program is used to model the data users to carry out the file upload process. A Private Server program is used to model the private cloud which manages the private keys and handles the file token computation. A Storage Server program is used to model the S-CSP which stores and duplicate files.

Our implementation of the Client provides the following function calls to support token generation and deduplication along the file upload process.

FileTag (File) - It computes SHA-1 hash of the File as File Tag;

Token Req (Tag, User ID) - It requests the Private cloud for File Token generation with the File Tag and User ID;

DupCheckReq (Token) - It requests the public cloud for Duplicate Check of the File by sending the file token received from private cloud;

ShareTokenReq (Tag, {Priv.}) - It requests the Private Server to generate the Share File Token with the File Tag and Target Sharing Privilege Set;

FileEncrypt (File) - It encrypts the File with Convergent Encryption using 256-bit AES algorithm in cipher block chaining (CBC) mode, where the convergent key is from SHA-256 Hashing of the file;

TokenGen (Tag, User ID) - It loads the associated privilege keys of the user and generate the token with HMAC-SHA-1 algorithm.

TokenReq (Tag, User ID) - It requests the Private cloud for File Token generation with the File Tag and User ID;

6. OPERATIONS PERFORMED ON HYBRID CLOUD

- File Uploading : When user want to upload the file to the public cloud then user first encrypt the file which is to be upload by make a use of the symmetric key, and send it to the Public cloud. At the same time user generates the key for that file and sends it to the private cloud. in this way user can upload the file in to the public cloud.

For Uploading a File:

```
BEGIN
 Step1- Read file and compute the hash value
   Step2- Cloud server verifies for duplication with
   the generated hash value.
   Step3- Sends duplication result whether the file
   already exists or not
   Step4- If the file not exist Display message “file
   does not exist”
   Step5- Then uploads the file by encrypting the
   file with the convergent key and then upload.
   Step6- If the file is already exist  then Display
   the   message “file already exist”.
END
```

- File Downloading: When user wants to download the file that he/she has upload on the public cloud. User makes a request to the public cloud. then public cloud provide a list of files that many users are upload on it. Among that user select one of the file form the list of files and enter the download option.at that time private cloud sends a message that enter the key for the file generated by the User, then user enters the key for the file that he/she is generated, then private cloud checks the key for that file and if the key is correct that means the user is valid, only then and then the user can download the file from the public cloud otherwise user can’t download the file. When user download the file from the public cloud it is in the encrypted format then user decrypt that file by using the same symmetric key.

For Downloading a File:

```
BEGIN
  Step1- Read the file
    Step2- Cloud server checks for duplication
    Step3- Sends duplication response whether the
     file already  exists or not
     Step4- If the file exist  Display “file exist” then
    decrypt the file and downloads the file from
    the cloud server.
    Step5- If a file does not exist Display message
     “file does not exist”
END
```

7. DESIGN GOALS

In this paper, we address the problem of privacy preserving deduplication in cloud computing and propose a new deduplication system supporting for:

- Differential Authorization: Each authorized user is able to access its individual token of his file to perform duplicate check based on authority. Under this assumption, any user cannot generate a token for duplicate check out of his access or without the aid from the private cloud server.

- Authorized Duplicate Check: Authorized user is able to access his/her own token from private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate. The security requirements considered in this paper lie in two folds, including the security of file token and security of data files. For the security of file token, two aspects are defined as unforgeability and indistinguishability of file token. The details are given below.

- Unforgeability of file token/duplicate-check token: User make registration in private cloud for generating file token. Using respective file token he/she upload or download files on public cloud. The users are not allowed to collude with the public cloud server to break the unforgeability of file tokens. In our system, the S-CSP is honest but curious and will honestly perform the duplicate check upon receiving the duplicate request from users. The duplicate check token of users should be issued from the private cloud server in our scheme.

- Indistinguishability of file token/duplicate-check token: It requires that any user without querying the private cloud server for some file token, he cannot get any useful information from the token, which includes the file information and key information.

- Data Confidentiality: Unauthorized users without appropriate token, including the S-CSP and the private cloud server, should be prevented from access to the underlying plaintext stored at S-CSP. In another word, the goal of the adversary is to retrieve and recover the files that do not belong to them. In our system, compared to the previous definition of data confidentiality based on convergent encryption, a higher level confidentiality is defined and achieved.

8. CONCLUSION

Hybrid clouds offer a greater flexibility to businesses while offering choice in terms of keeping control and security. Hybrid clouds are usually deployed by the organizations willing to push part of their workloads to public clouds either for cloud bursting purposes or for projects requiring faster implementation Because hybrid clouds vary based on company needs and structure of implementation. In proposed system authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check system presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed system is secure in terms of insider and outsider attacks specified in the proposed security model. The proposed authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou,"A Hybrid Cloud Approach for Secure Authorized Deduplication", *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS,* VOL. 26, NO. 5, MAY 2015.

[2] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," *in Proc. 1st USENIX Conf. File Storage Technol.,* Jan. 2002,p. 7.

[3] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," *in Proc. ACM Conf Comput. Commun. Security*, 2011, pp. 491–500.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," *in Proc. 32nd Annu. Int. Conf. Theory Appl*. Cryptographic Techn., 2013, pp. 296–312.

[5] 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 Proc. Int. Conf. Distrib. Comput. Syst.,* 2002, pp. 617–624.

[6] M. Bellare, C. Namprempre, and G. Neven, "Security proofs for identity-based identification and signature schemes," J. Cryptol., vol. 22, no. 1, pp. 1–61, 2009.

[7] M. Bellare and A. Palacio, "Gq and schnorr identification Proofs of security against impersonation under active an concurrent attacks," in *Proc. 22nd Annu. Int. Cryptol. Conf. Adv. Cryptol.,*2002, pp. 162–177.

[8] 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.

[9] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C.Lee, and Wenjing Lou (2014) 'A Hybrid Cloud Approach for Secure Authorized Deduplication' in Proc.IEEE Trans. Parallel and Distrib. Syst.,

[10] Dutch T Meyer and William J Bolosky."A study of practical Deduplication". ACM Transactions on Storage (TOS), 7(4):14, 2012.