ISSN: 2321-7782 (Online)
Volume 2, Issue 1, January 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

New Challenges for Security against Deduplication in Cloud Computing

Deepika Singh[1]
Computer Science and Engineering
Government Engineering College, Modasa (GTU)
Modasa - India

Preetika Singh[2]
Computer Science and Engineering
Sri Sai Institute Of Engineering and Technology, PTU
Pathankot - India

Abstract: ***cloud computing provides an on demand, self service, rapid, elastic and ubiquitous access to various computing resources. Deduplication is the process of reducing redundancy of data storage services provided by the cloud computing by keeping a single copy of any file in spite of 'n' number of users of the file. Users are provided with the link to the file and only a single copy is stored at the server.***

However, as Deduplication emerges as an answer to the increased demand of storage services in the cloud infrastructures, it introduces the vulnerability of side channel attacks due to cross user deduplication.

This paper reviews and analyses various attacks related to deduplication and the techniques proposed to overcome those attacks. It has been observed that in spite of various solutions provided, deduplication still suffers from the vulnerability of one or the other side channel attack. Hence in this paper we propose a solution which helps to reduce the vulnerability of several side channel attacks while keeping the essence of deduplication alive. Also it helps to establish trust between the cloud storage provider and user.

Keywords: cloud computing, deduplication, storage services, side channel attack, security, trust.

I. INTRODUCTION

Cloud computing, in addition to other services provides various infrastructures as service. Storage-as-a-service is one of the most important and widely used infrastructures provided by cloud computing technology. With the increasing demand of computers and other computer based services, the demand for data storage is also increasing day by day. In this scenario cloud computing offers best solutions for rapid, elastic, reliable, and measured storage 7...

The increasing demand of cloud storage requirement has led to the process of deduplication. The term data deduplication refers to techniques that store only a single copy of redundant data, and provide links to that copy instead of storing other actual copies of this data 1.. The deduplication process is used to protect the cloud server from storing redundant data. If two users want to upload the same file, only a single file will be uploaded on the cloud server and the users will be provided with a link that will fetch the whole file for them whenever they want to retrieve it. Suppose user1 on cloud stores a file A. He will request to upload the file and the file will be successfully uploaded now when a user, user2 will upload the same file, the cloud will deduplicate the file by providing user2 the link of file A, which is already present on the cloud. Thus 'n' number of users can be allowed to access same file with a single copy stored on cloud.

The deduplication can be performed on the cloud server. If the whole file is first transferred to the cloud server before any deduplication, this is server-based approach for deduplication. This process saves the storage space in above mentioned way but the network bandwidth for sending the redundant data is wasted. Thus client side deduplication is used to save network bandwidth as well as storage space. However the deduplication process has the capability to save both storage space and

network bandwidth but this process give rise to a security problem in cloud computing, the side channel attack. The cross virtual machine users can use several attacks to find the confidential data related to each other as well as the administrator. Thus cross user deduplication leads to the vulnerability of side channel attacks in cloud computing.

This paper reviews various attack models and solutions provided for the deduplication security. Section I of this paper gives a brief introduction to the paper. Section II reviews several attack models that can be used to misuse the deduplication technology. In section III various solutions proposed have been reviewed and analysed while the last section provides our proposed solution to enhance the security of deduplication and establish a trust between the cloud provider and the cloud users while keeping the essence of deduplication.

II. ATTACK MODELS

Several attack models have been discovered, which can lead to the exploitation of deduplication towards an insecure storage method. However looking at the excellent possibilities of enhancing storage efficiency, several solutions have been proposed. This section describes various attack models. The first attack can be used to predict an already known file possessed by the user. The second attack is related to creating a secret channel for extracting information while the third attack is related to distribution of any file among various users of cloud storage.

A. Attack Model I: Predicting files

This attack can be used to predict whether a particular file is possessed by a specific user 1.. Furthermore this attack can be more efficiently used to predict a file if the file contains data with limited possibilities for example yes or no in case of a medical test report. Suppose the attacker wants to find out whether user1 possesses a file, File A. He will upload a copy of file A if the file gets uploaded this will indicate that the file is not possessed by the user1. Whereas in the other case the attacker will be able to find out if the file is possessed by user1. Also, in order to hide his identity as an attacker he will terminate the connection as soon as the file uploading starts.

B. Attack Model II: Creating a secret channel

If the attacker manages to install any malicious software on the machine of user1, this software can be used to establish a secret channel between the user1 and the attacker 1.. There are several ways of creating this type of channel one of them is to bypass the firewall and communicate with its control server. Consider this example; suppose user1 is using the system with malicious software installed, the software will generate two files in two different conditions. When user1 will backup his files on control server this file will be stored on the server. Now, attacker can easily use the attack described in previous section to find which file was stored by the software. This attack can use any number of conditional files, and thus can be very harmful as well as undetectable.

C. Attack Model III: The Content Distribution Attack

The content distribution attack can be used to distribute a specific file to various users without providing the identity of the distributor. The type of file can be a bootlegged video or a file containing a virus etc.

The users in deduplication are enabled to use a file if they are included in the access control list of the file. The access to a file is gained through a hash, h(F) where F is the file discussed above and the h(F) is the hash value corresponding to the file F, calculated by the control server. The attacker can get h(F) by uploading the required file and distribute it to the several users in order to let them access the file.

III. Analysis of Solutions

The attacks described above present the various security threats while using cross user deduplication, However looking at the storage savings provided by deduplication it is impossible to neglect deduplication in cloud storage solutions. The deduplication can provide 90% savings for network bandwidth as well as the disk space 6...Therefore a number of solutions have been proposed by various researchers in this regard. These solutions have been discussed in detail in this section.

A. Encrypting files before uploading

The attacks described above come into existence because of cross user deduplication, if the sensitive files will be encrypted before uploading them to the cloud, this will result in avoiding deduplication and hence none of the attacks will be possible. The files can be encrypted by the private key of the user. This key should not be disclosed and should be kept personal by each and every user. However this leads to the problem of dictionary attack. The private key of any user is vulnerable to offline dictionary attacks. Also if two users by accident use same private key then deduplication will still occur. Another problem related with this option is regarding the generation of the personal key and the bookkeeping task for example that if a user forgets his personal key. The previous research on deduplication proposed a "convergent encryption" this type of encryption is capable of generating identical encrypted files from original files even when the users use different personal keys. This method keeps the essence of deduplication while providing a good level of security. 3.

B. Performing target based deduplication.

Target based deduplication performs deduplication at the server side rather than the client side. This method however provides the security at the risk of network and bandwidth savings.

C. Randomisation

Randomisation solution was proposed in 1.. This method assigns a number named as, a random threshold to every file. The file is not deduplicated until the number of users of the file reach to the threshold. After the number of users of file increases the threshold, the deduplication takes place. The biggest loophole of this method is that the threshold is chosen uniformly at random thus lacking reliability.

D. Proof of ownership

This method uses block level deduplication, this process divides the file into fixed or variable size blocks before deduplicating them. In this method, the user needs to provide the proof that he owns the particular file. Suppose user 1 uploads a file A. This file will be first divided into number of blocks B, each block will be stored corresponding to a particular index I and then stored at the server. Now if user 2 uploads the same file the server will ask the user to prove that he is the real owner of the file. For this the server will ask the user to send a particular block B(x) corresponding to a specific index I(x) where x is any random number corresponding to the number of blocks. If user 2 satisfies the challenge he is added to the access control list of the file and in other case the request is not granted. However this method reduces the chances of certain attacks like content distribution, it suffers from the problem of probing attack 1...

E. Gateway based deduplication

This method aims to provide transparency to the users and hence the attacker 4.. The gateway acts as a common terminal for all the networks providing residential internet access. The abstract view of this method is shown in figure 1 4.. All the upload and download requests of users of cloud storage service are handled by the gateway itself. As shown in the figure below there are five different modules of the gateway, SSP (storage service provider), gateway server, gateway client, bandwidth manager and the user client. When user client wishes to upload a file from user PC, it sends a request from its home network to the gateway server. Since the deduplication does not take place at the client side, the user is now free to proceed with his other tasks. The gateway server handles the file to the gateway client. The job of gateway client is to check with SSP weather the

given file is already present on SSP. If the file is already present, SSP server creates a link to the file, however in the other case the file is handled to the bandwidth manager which uploads the file to SSP. This method divides the whole setup into two parts one of which is easily accessible to the attacker and the other is not, and to provide best security by deduplication attacks, the deduplication takes place at the other end.

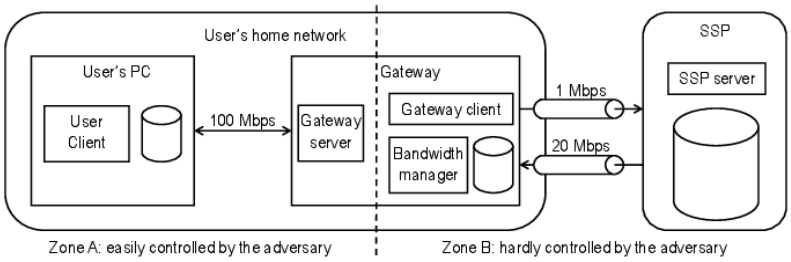

Figure 1: Gateway Based Deduplication 4.

This solution is very realistic with minimum vulnerability but the main aim of cloud computing, ubiquitous access is lost.

IV. FUTURE WORK

The solutions described above provide security in different ways but none of the solutions cover all the vulnerability. In addition to that not even a single method has been proposed which can establish the trust between the user and the cloud service provider. The client is never assured that his files are known only to him and to no one else including the SSP. Therefore we propose a solution whose main aim is to establish the trust between the user and SSP. This method involves injecting an application on client side LAN. The abstract view of the setup is shown in figure 1. The various modules and the working of the system are described below:

User is the person who uses the storage service provided. To use the services of SSP the user has to first register himself to the application the application will generate a user id corresponding to each user. M1 is the hash generator responsible for generating hash values of a file by certain algorithm like SHA or MD5. When the user uploads a file M1 calculates its hash value and passes the hash value, user id and username to M2. M2, the second module acts as the database containing hash value and user id corresponding to each file. M3 is the database containing the filename corresponding hash value and the user id of that file. M4, the Encryption –decryption Module encrypts the file before sending them to the cloud and decrypts the file in the reverse case. The cloud is the storage server which contains encrypted file with their respective hash values. When second user will upload a file already present on the cloud its hash value will be calculated and the database M2 will be checked for the presence of calculated hash value. As the hash value is already present in the database, the user will be attached to the list of database having hash value and user id. When this file needs to be retrieved by the user, the database M2 will be checked and the file from cloud will be downloaded with reference to the hash value. After this the file will be retrieved from the cloud, it will be decrypted and will be provided to the user.

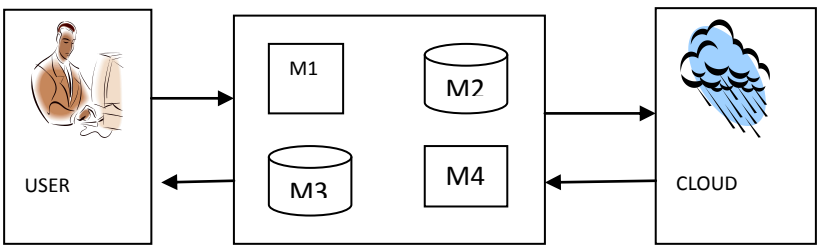

Figure 2: Proposed Application

V. CONCLUSION

This paper aims to present various risks induced in cloud storage services due to the use of deduplication. Taking into consideration the amount of network bandwidth and disk space saved by deduplication, various methods have been proposed against these risks. These methods with their pros and cons have also been presented in the paper. These solutions however suffer from one or the other loophole. In our work we propose a solution that not only removes the risk of all the three attacks described but also it helps to establish a trust between the cloud service provider and the user.

References

1. D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," Security Privacy, IEEE, vol. 8, no. 6, pp. 40 –47, nov.-dec. 2010.
2. http://en.wikipedia.org/wiki/Cloud_computing
3. http://searchcloudcomputing.techtarget.co/ Security Analysis of Cloud Computing.
4. Olivier Heen, Christoph Neumann, Luis Montalvo and Serge Defrance, "Improving the Resistance to Side-channel Attacks on Cloud Storage Services", 2013.
5. M. Dutch. Understanding data deduplication ratios. White paper, June 2008.
6. M. Dutch and L. Freeman, Understanding data de-duplication ratios, SNIA, February 2009, http://www.snia.org/forums /dmf/news/articles/SNIA DeDupe Ratio Feb09.pdf
7. National Institute of Science and Technology. "The NIST Definition of cloud computing, Luis M. Vaquero1, Luis Rodero-Merino1, Juan Caceres1, Maik Cloud Computing".p.7. Retrieved July 24 2011

AUTHOR(S) PROFILE

Deepika Singh received her bachelor degree, in Computer Technology from Rashtrasant Tukadoji Maharaj University, Nagpur in 2010. From 2010 to 2012 she actively worked in academic activities as a Lecturer in Gujarat Technological University. Currently she is pursuing Mtech final semester from Gujarat Technological University. Deduplication in clouds is her area of research.

Preetika Singh received her bachelor degree, in Computer Engineering from Rashtrasant Tukadoji Maharaj University, Nagpur in 2011. From 2011 to 2012 she actively worked in academic activities as a Lecturer in Gujarat Technological University. Currently she is pursuing her Mtech final semester from Punjab Technical University. Cloud security is her area of research.