ANALYSIS OF ATTACK TECHNIQUES ON CLOUD BASED DATA DEDUPLICATION TECHNIQUES

AKM Bahalul Haque

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ABSTRACT

Data in the cloud is increasing rapidly. This huge amount of data is stored in various data centers around the world. Data deduplication allows lossless compression by removing the duplicate data. So, these data centers are able to utilize the storage efficiently by removing the redundant data. Attacks in the cloud computing infrastructure are not new, but attacks based on the deduplication feature in the cloud computing is relatively new and has made its urge nowadays. Attacks on deduplication features in the cloud environment can happen in several ways and can give away sensitive information. Though, deduplication feature facilitates efficient storage usage and bandwidth utilization, there are some drawbacks of this feature. In this paper, data deduplication features are closely examined. The behavior of data deduplication depending on its various parameters are explained and analyzed in this paper.

KEYWORDS

Cloud Computing, Attacks, Data, Deduplication, Storage

1. INTRODUCTION

The amount of data is increasing day by day and is supposed to reach 163 zettabytes every year in the next eight years that is by 2025 [1]. Due to this increment of data, there is a high demand in data storage which may be a problem. So, there has always been a vice versa challenging issue between the increased data and storage. Thus, they need to utilize the storage using compression. Moreover, the user wants easy and on-demand access to their data. This is where the cloud has emerged and established its present importance and the future structure of cloud computing. Users can store data in the cloud and can access anytime anywhere with an internet connection. According to NIST the definition of cloud computing, "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics (On-demand self-service, Broad network access, Resource pooling, Rapid elasticity, Measured Service); three service models (Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), Cloud Infrastructure as a Service (IaaS)); and, four deployment models (Private cloud, Community cloud, Public cloud, Hybrid cloud)" [2]. All the data produced around the world are the mixture of digital and analogue data. Analog data is converted to digital data for processing. There remains a question, whether all the data produced and used by users are unique! All the data cannot be unique. Every portion of data around the world cannot be unique. It would need a vast amount of storage and a system to cool the system down. People use smartphones. Smartphones have different types of sensors, and continuously the data is backed up over the internet. So is everyone. A group of friends might click the same pictures and backup from each of their devices. This happens all the time while on the travel or at college. If the cloud providers use the

DOI: 10.5121/ijccsa.2019.9101

same data again and again to store, it will run out of storage. So, they take the deduplication technology at their disposal and go for the better and efficient storage facility. Every technological advance has its downside or security issues. There are several public cloud providers who provide services to users every day along with private cloud services.

2. SOME TERMINOLOGIES OF CLOUD COMPUTING

Software as a service (SAAS) [3] is one of the infrastructures which provide software services to the users who want to use specific services e.g. services based on big data, server less computing etc. Through these services users are able to run and use the software as per their need without having the hardware setup complexities and costs. Platform as a Service(PaaS), [4] provides platform facility for the clients to build and develop their own applications. These facilities increase their work efficiency and also gives a common platform for teamwork. Infrastructure as a Service (IAAS) [5] provides Virtualized Computing Resources all over the Internet. It hosts the infrastructure components and its traditional presence in an On-Premises data centre. The infrastructure includes server storage, networking hardware, virtualization, hypervisor layer etc. It provides users the facility to collaborate between the services and different elements. In this paper, deduplication scheme has been used and tested to observe better the deduplication scenario and analyse it to get some useful information for vulnerability issues in deduplication enabled cloud computing environment. Various characteristics of deduplication depending on the data type will be tested and a detailed analysis will be shown here in this paper.

3. DEDUPLICATION

Data deduplication is meant to delete the duplicate data. It is an important issue for many good reasons [6] [7]. Due to the rapid increment of data, there is in need of a significant amount of storage for storing this data. There is a vast amount of same data, which are being stored. For example within the same organization or group of people with the same interest rather uses and processes same data over time. In an educational institution or any organization, the users are processing lecture notes or assignments on the same cloud servers. If there could be any means to eradicate this duplicate data as well as ensuring availability and integrity, it would have been a great success in the case of storage management in the cloud. That is where the deduplication [8] comes in. It refers to removing the duplicate data from the storage. If the same data exists in storage only one copy will originally be stored, and others will be removed from the storage saving the disk space. For example, if there are 10 similar files and each of the file has 5 MB of data and deduplication method is implemented in this case, only 5 MB storage space will be stored. Various types of deduplication exist depending on different natures.

3.1 File Level Deduplication

File-level [9] deduplication as mentioned in the name is based on the file. Every file stored, is given with a unique identifier which is dependent on the file's attributes. All the information about that file (as a whole) is stored in reference with that identifier. So that, the same file having different names can be easily identified. New files uploaded in the server are compared with the previous one. If the same identifier exists, it simply adds a pointer to it for redirecting the file to its original location and removes the identical file. If there is any change in any file, the whole

Fig 1: File Based Deduplication

3.2 Block Level Deduplication

As it is seen in the file level deduplication, the process is done according to the file. The whole file index is saved. On the other hand, block-level deduplication is done by the blocks or chunks of the file.

That is the file is divided into blocks or chunks. Each block is equipped with an identifier which is unique. For this unique identifier to be produced, a unique hash value is given to them. [10]. Next time, while a file is updated in the storage, it is again divided into several pieces of blocks according to the size. There are different sizes of blocks starting from the power of 2 to as low as 128 KB. [11] Incrementing the block sizes may make the system performance degrade if the system does not have enough RAM. The hash values produced and stored are, used for the comparison purpose to identify the similarity among them. Blocks (also referred to as chunks) can be of different sizes.

- Smaller block size divides the file into more number of blocks.
- Larger block size divides the file into few number of blocks

It means that more blocks will be occupied in the storage and have to be compared. So, the computational time has to be higher. Though more blocks will be checked for checking identical blocks but the trade-off is computational power.

at210hgstb768shrin456s8	64 KB Data Block
po657hybdfo4927bftou4w	64 KB Data Block
mn569sgftbx579ksbif23tr	64 KB Data Block
4co5our37264boxcnsf386	64 KB Data Block
Yb3495ns294nsfdkeih375	64 KB Data Block
Xbdkwetwe859bsdhwi274b	64 KB Data Block
Sfw84h66d93fhegwixb375	64 KB Data Block
Xbdk385bjsdg32742bfigf84	64 KB Data Block
	New 64 KB block containing same Hash 4co5our37264boxcnsf386

Fig 2: Block Based Deduplication

3.3 Target-Based Deduplication

In this type of deduplication, the deduplication happens on the target side, that is, the decision of deduplication e.g. whether the deduplication should take place or how it will be done happens on the target backup cloud. This nature of de-duplicating the data is called target-based deduplication. When deduplication was introduced in the data backup industry first, the target based deduplication was introduced by the service providers. This method of deduplication was introduced as there was no need for extra software to be used both in the data centers and in the client site. Even if the providers wanted to move their data to any other offsite locations, it was also possible as the layer was implemented onto the layer for the backup server. On the other hand, the target cloud backup has to have hardware required for the deduplication procedure. Target deduplication is used by both intelligent disk targets and virtual tape libraries. Moreover, the storage provider needs to have the backup deduplication software for performing this action.

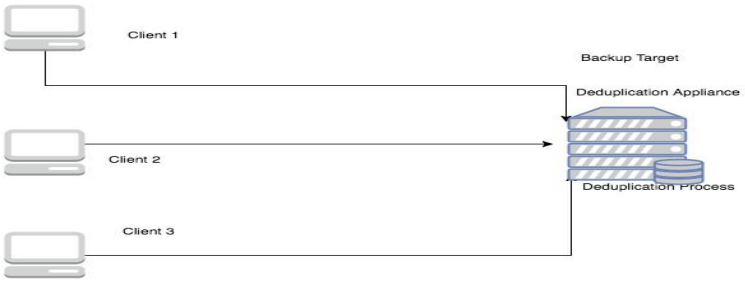

Fig 3: Target Based Deduplication

3.4 Source-Based Deduplication

In source based deduplication procedure, as the name suggests, the deduplication happens before the data is transferred for backup. The decision is taken on the client side. For this type of deduplication to happen, the service provider doesn't have to implement any new hardware as the single instance is transmitted from the source. A client software has to be set up on the client side capable of hash value production and determining which unique blocks have to be uploaded. The backup agent connects to the server and checks against the already stored blocks in it. This type of deduplication allows the client not to use an enormous amount of bandwidth and create traffic on the internet. So, there is a low bandwidth usage on the client side. While transferring large amount of data for backup, in this type of deduplication, the backup process tends to be slower as a huge number of blocks' hashes needs to be compared with those of already stored in the backup target [11] [12] [13].

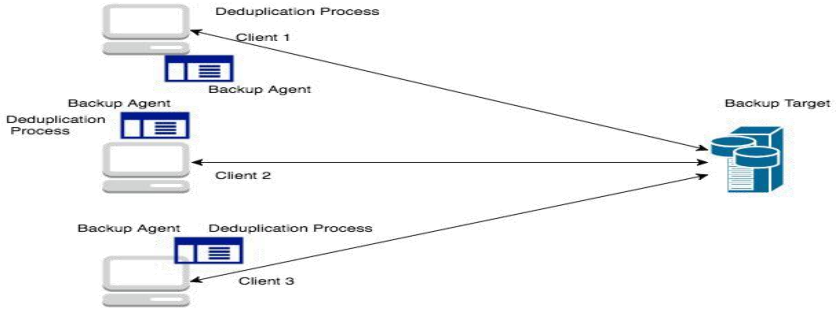

Fig 4: Source Based Deduplication

3.5 Deduplication Ratio

Deduplication ratio [14] [15] is defined by the ratio between the storage space that is logical to be used and the space that is actually or physically used. That means, it depicts how many similar data has been removed from the storage after it is de-duplicated. The deduplication ratio 4:1 means, one instance of four identical files have been stored. That is, four times physical data storage is protected.

There are several influential factors for deduplication ratio. The redundancy of data is very important if anyone wants to achieve deduplication. Without having redundancy or similarity, there will simply be no deduplication. The more the redundancy the more the deduplication ratio. Virtualbox backups tend to have higher deduplication ratios because they contain similar files

from the previous backup versions. Another factor is the frequent change of data. If the data changes frequently then the deduplication ratio is supposed to decrease as the possibility of finding a duplicate block gets lessened. Other factors include data retention period, data type. Data backup nature etc.

Data retention period helps the existing data to be compared with the new data being uploaded so that more identical data blocks will be found. Let us assume the deduplication ratio of a storage environment is 10:1. If the data stays longer in the storage system for a long span of time, the deduplication ratio might get to 20:1. The higher ratio occurs for a very simple reason. The deduplication system tends to find more similar blocks or files inside the storage. Though this feature depends on how much redundancy the data possess and how long it remains stored. Protecting the data and ensuring its integrity is one of the main important tasks of the cloud service provider. Back up is one of the techniques to ensure that. There are a lot of policies for data backup in a backup storage device. These policies define, how frequently the data backup procedure will run from the device to the cloud provider, which files has to be backed up etc. If there is a persistent backup of data, the deduplication will be higher. There are other backup methodologies like incremental or differential backups. Depending on the various backup technique, the deduplication ratio varies.

The data type is another factor in data deduplication. The data type impacts the data deduplication as some specific data types which normally has lower deduplication ratio. End users use a vast variety of data each, and every day, they also produce a variety of data every day. There are some files which are already pre-compressed like jpeg, MPEG or zip files. These types of files are likely to show less deduplication ratio as they are already compressed. Moreover, it is also seen that compression also gives an extra bump while data deduplication as the files is already compressed and contains less duplicate data so the file will take much less space while de-duplicating. The scope of deduplication can be local or Global. Local deduplication considers the file within a very limited space like in a single hard disk or a single drive and global deduplication means across the whole system. So, there will be a variety of files in the deduplication environment. This diversity in files will lead the deduplication ratio higher.

4. RELATED WORKS

Data deduplication refers to removing the duplicate data from files [16]. This technology uses files or blocks comparison. During the file based comparison, each file newly uploaded is compared with the other. And during the block-based comparison, each block is compared and checked against each other. The files are stored in different block sizes or chunk sizes. Now, it is not possible to check and compare the whole file against each other, which would take a significant amount of time during the backup process. So, the approach is to mathematically compute hash values for each file and blocks. In the case of file-based deduplication, file hashes are compared and in the case of block-based deduplication, the hashes of each block are compared. These hashes are stored in the random-access memory of the deduplication storage environment.

The chunks can be of fixed size and variable sizes. During the whole file chunking, the file is given a hash value. In this case, even if one byte inside the file is changed, the hash value of the file is changed. So, the file will be stored again. In case of block size or chunking approach, each chunk is given a hash value and their values are stored in RAM.
The hash values are compared against each other. So, if one portion of a file is changed, that block will be updated only rather updating the whole file. [17] [18] In case of block-based deduplication, if one byte of the file is changed and assuming, that one byte is not among the de-duplicated chunks, the variable chunk size method can be useful in this aspect. Depending on the timing of the deduplication, that is, when the deduplication will happen, defines another two types

of deduplication. Inline deduplication and post-process deduplication. In inline deduplication, the data is de-duplicated while it is written on the disk. The write request has to identify the duplicate blocks those exist in the server so that, the identical data can be discarded.

Another type of deduplication technique is a post-process deduplication technique. In this type of deduplication technique, the file storage is scanned. At first, all the files are stored in the storage. Later the whole storage is scanned for finding identical blocks. It is neither dependable on the write path not dependable on the write interception between the client and the server. As the cloud computing era is advancing, there are several security issues that arises. Some of the security issues of cloud computing in different infrastructures are explained here [19]. All of them are not the issues with data deduplication, but it provides a useful insight to dig deeper into the cloud security field. On the other hand, as the deduplication feature has increased, its security features and specifications have become a major concern also. Among them, Side channel attack is one of the major important issues which solely points towards the data deduplication vulnerabilities. These vulnerabilities include in a cross-user de-duplicated cloud environment. In this environment, data deduplication attack scenario like, collecting the file information and also owning the specific content of any specific file is mentioned [20].

After knowing, if the cloud environment has deduplication feature enabled, the authors of the paper [20] has been able to identify the vulnerabilities and perform the attack in cross-user source-based deduplication technique. The solution of these attacks has been proposed as encrypting the files before uploading and also target based deduplication scheme can be used. Encryption in data deduplication plays some adverse role as after encrypting the identical files, they produce different hash values which clearly is not very convenient while de-duplicating the data as de-duplicating data refers to comparing the hash value produced from the files. Another approach to solving this contradicting problem has been proposed. [21].

Data deduplication is a feature of storage utilization technique through lossless data compression. This feature uses the hash function to produce a unique signature which is used to compare between files. If two files have the same content they are likely to produce the identical hash functions. If the files are divided into blocks and then the hash function is applied it will produce the exact same hash value. But there is a problem of using hash functions. They are susceptible to hash collision, birthday attack, brute force attack, limited brute force attack [22]. Moreover, if the information about which hash methods e.g. MD5, SHA-256 is used in a specific deduplication environment, it is also vulnerable to attack as these techniques have their own vulnerabilities. Researchers have explained MD5 can be compromised by using the structure of the block inside it[23]. The recent compromise of Dropbox deduplication system named Dropship [24] has already raised many flags about the privacy concerns of the users. Dropship can exploit cross-user source-based data deduplication strategy. In another type of attack, in case of Dropbox, the id that is used as a connecting key between the client device and the client's user account can be used to retrieve all the files from the user's account. Vulnerability also lies in the transmission features of Dropbox. The transmission protocol uses HTTPS POST. The transmission path can be manipulated and the specific files can be retrieved from the account though it needs a few crucial information from the victim [25].

5. DATA DEDUPLICATION IN CLOUD STORAGE

Different cloud service providers have emerged in recent days and their popularity has increased depending on the range of services they provide. These services are mobile office, content delivery storage services etc. Some of the most popular services like Google Docs [26], DropBox [27], Amazon S3 [28] etc. provide on demand user access to the consumers. If the users' needs to access more services like storage, they need to provide more. This pay per usage policy makes the user think the providers to have unlimited access to the services [29].

Cloud service providers data can be sectioned as the Virtual Machine Images and User Data. A snapshot of the current virtual system is stored if the user has stopped using the system. The files processed by the users are also stored. Files processed by the users are allocated to the virtual machine as boot partition. During the user's runtime two things are used; CPU & Memory. The memory should be maintained even after the runtime is closed by the user. So storage has been always been a matter of concern for the service providers. While thinking about the deduplication mechanism and file types we know that virtual machine images tend to show more deduplication ratio. The files change from one machine to another is very few in number as the operating system across them is the same [30]. This useful piece of insight shows more efficiency while using the data deduplication the cloud storage environment more specifically separating the virtual machine image files from the user data. This is an efficient way to utilize storage by reduplicating the identical files. It has impact to other factors like efficient energy consumption, cooling needed for the storage environment etc. If the storage is efficient it inherently impacts energy consumption and cooling system. So, the more efficient the storage the more efficient the energy consumption and cooling system would be.

5.1 Blocks and Hash Values in Deduplication

In the file based deduplication hash value is used. The hash value is created from the stored file. While the data is updated or stored newly, this hash is compared to the new to see if there is change. Files containing identical hash will not be stored. If any difference exists, only the newly added portion is sent to storage. Hashes are considered to be unique. If one part of the file is updated, the hash will be changed, so that the service providers can update it. Hash is also used in block based deduplication. Here, every block of data is mathematically hashed. These hashes are appropriately stored and indexed. While a file is updated, the new blocks are also mathematically hashed and compared with the previous ones. So, there is a match in the hashes; they are not stored and thus de-duplicated. If there is no match, only the updated data is stored, and the index is updated.

In this case, the versioning system helps for data recovery, bandwidth utilization, time consuming, etc. as only the updated file is sent over the internet. Sometimes, the whole data is transmitted and hashed on the server side to do this operation, but the same algorithm is followed for data deduplication. Hashing technique and comparison is a bit different in sub-block delta versioning system. The block is done at the byte level. As the block is done in byte level, it is prawn to find more duplicate blocks or chunks.

For example, if in some open server systems, the disk is formatted into a much smaller byte, e.g., 512 bytes. So, the smaller the chunk, the much duplicate chunks to be found in the storage. On

another note, since, the blocks/chunks are so small more hashing is needed to be done mathematically and indexed. For this reason, the indexing is much more extensive, and it takes much time for hash comparison. The hash indexing will also increase with the deduplication. On the other hand, it saves more storage and gives better deduplication ratio overall.

6. ATTACKS ON CLOUD BASED DATA DEDUPLICATION ENVIRONMENT

Cloud computing has evolved from ubiquitous access of data. The users can store data and perform different types of operation in the cloud. But these features have facilitated the attackers to hamper the security infrastructures of the cloud. Various attacks on the cloud computing depending on their nature of services have been discovered by the attackers. These attacks are effective and makes the system and storage vulnerable to attackers. Though most of these attacks are generalized, but still, they have given an overview to the security experts in developing their

security measure as well as to the attackers to design and perform new type of attacks and make the system vulnerable.

As deduplication has evolved, there are also security issues in it. There might be attacks which lie inside the very inside of the feature of this technique. Deduplication removes redundancy from the storage, but every user has access to the same cloud providers' storage considering the users using the same service providers' service. So, if the followings are seen –

- deduplication exists in storage
- the cloud storage uses cross user deduplication
- the deduplication is source based

The possibility of existence of certain file in that storage is possible. So deduplication has to be detected at first. If deduplication can be detected, there are a few attack opportunities.

Attacks through hash manipulation include brute-force attack, birthday attack, limited brute-force attack. The brute-force attack using has manipulation is possible theoretically only. While performing the attack the attacker needs to look for the hash collision trying all kinds of hash which he will produce from a file. If a similar hash exists, that is hash collision happens, it is assumed that the same file is already uploaded in the server.

Privacy concern of Dropbox is one of the recent issues because of Dropship exploit. It has a few other vulnerabilities also given the conditions above. In case of Dropbox, while connecting with the server for a file, the user is authenticated once with the server. Later a Host ID is created and the client connects to Dropbox using that id only. The device does not authenticate any further using the username and password of the user. The id used to connect with the client and the Dropbox account. This is a randomly generated key of 128 bit, but the mechanism used for the key generation is unknown to the common people. If a malware infection, Trojan infection or any kind of social engineering can be used to get that 128-bit id, it can be used to retrieve all the files from the specific user's account. Here the primary concern is to get the valid host id.

The transmission characteristics of Dropbox raises another concern also. It uses the HTTPS POST. The client software uses "https://dlclientXX.dropbox.com/retrieve" to retrieve files from the Dropbox server. Here the "XX" is replaced by the host id and the chunk size(given that the chunk size is known). If the intended hash value of both (host id and chunk) can be replaced in the URL the file can be downloaded from the server. The problem with this attack is, it can be detected by Dropbox as the host id that will be used to retrieve the file was not used to download the file. So it is easily detectable by Dropbox.

Table1: Various Attack vectors

Attack Methods	Detectability	Result
Hash Compromise	None	Confidentiality is compromised
Host ID Compromise	Dropbox	Confidentiality is compromised
URL infection	Dropbox	All user files

A brute force attack is basically the only way to find cryptographic hash collisions.[31] Since data deduplication requires hashing of files or chunks it is often a subject of brute force attack. A brute force attack is an exhaustive attempt of attacks by an adversary to target the encrypted information and get the preimages which are practically infeasible since the message digest is

often very long. For an N bit hash value, the level of effort for a brute force operation is proportional to 2n. [22] So instead a limited brute force is in practice in which the adversary has knowledge of a set of preimages and a hash function. The attacker then computes all the possible hash values of the preimages to match the known hash value and attempts to establish a hash collision.[22].

Birthday attack corresponds to the mathematical Birthday paradox problem which is a type of brute force attack.[22] It applies the probabilistic theory to determine the collisions of hash functions that is for the input of different user files, the hash values produced is the same.[32] For a variable sized input, if the hash value produced by a hash function is identical, then there occurs a hash collision. Let us correspond the 365 days in a year to any number. Between 1 to 365 a uniform hashing function may produce as an output. The probability of two hashes colliding in a set of 23 values during data deduplication process is above 50%. The implication of birthday attack in the context of hash collision calculates the number of preimages required to get a probability of 50% hash collision rate.

Side channel attacks (Pinkas et al. 2010) are also possible if the cross user client side deduplication is used. Using the side channel attack identifying a specific file and the identifying the file content both are possible.

One type of attack is identifying a file. It means, identifying a specific file among a group of files. The attacker has to assume about a file that is possessed by the victim and do the rest of the work. Let's assume, A and B use the same cloud environment. A has detected that the cloud service uses a cross-user deduplication**.**

So, what A will do, he will try to upload the assumed file to the cloud and check if the deduplication occurs. In another case, A might have a group of files, and among them, there might be a file which is in possession of B. So, A tries to upload that assumed file and checks if the deduplication occurs. Here, the deduplication occurrence can be observed by measuring the network traffic or the upload status of the file, but network traffic is one of best ways to check if deduplication occurs in that server.

So, among all the files, the attacker has uploading the correct one will be showing a bandwidth usage change in the network traffic than the previous one. After checking this factor, if there is deduplication in the cloud, A can perfectly identify, if the file is already uploaded or in possession of B.

Figure 5: Side Channel Attack Type One

Another type of attack is to identify the contents of the files that is, to know what the victim has in possession. In this type of attack, the attacker A tries to know the file contents of the victim B. If the system uses a cross-user source-based deduplication, A will try to detect deduplication and use the assumed files. Here, let's assume A and B are students in the same university and the university uses the cloud network which supports deduplication feature. Now B has stored her result of a specific subject in her university cloud account.

A also had the same subject, and he knows how the result is published that is the format of the result. A is curious to know B's result. Now, A uses a brute-force method to identify it. A knows the subject's name, the corresponding supervisor, and other formats. So, A will create a few of the same result formats guessing B's result. B comes to university every day and accesses his files, and she runs a backup, of course, every day or every week assuming that auto backup is disabled. When the backup is done, the data is stored in the backup storage. The data is de-duplicated as the university server is used by all the students. The students store their personal and academic information on the server. So, the academic documents contain similarly formatted and some certain similar types of data. The deduplication occurs in case of those similar files and A has targeted one of those files.

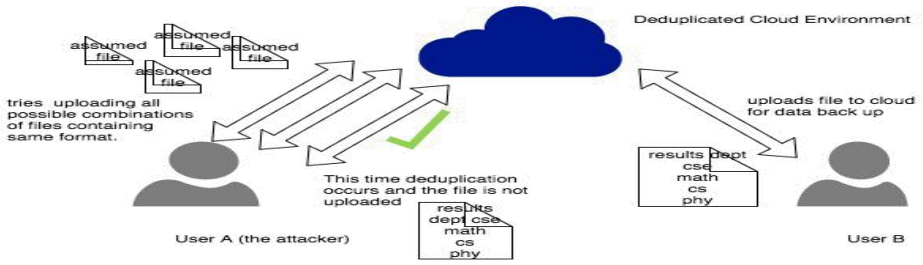

Figure 6: Side Channel Attack Type Two

Now A has multiple versions of the same files that he has produced for the attack. There are different assumed values. Now, if the values are close enough it will take less time to compare against the already backed up files. Each of those assumed values make a certain format of the files that B has. Now at this point, A will apply brute-force attack for identifying the file contents. Basically, what A has to do is to use all the files as possible combinations and try for backup all of them. Among the files, if the deduplication exists, that file will not be uploaded. This can be observed by checking the network traffic. The file which is not uploaded is the file that A is looking for.

7. DATA ANALYSIS IN DEDUPLICATION ENVIRONMENT

In the deduplication table (DDT), it is seen that, the number of blocks which are supposed to occupy in the storage and the actually occupied blocks are increasing. The deduplication ratio

largely depends on the block sizes and the size of the files. The ratio is greater when, smaller number of block sizes are used for the file system. Changing the block size to smaller number, increases the deduplication ratio. This happens because, the increased number of blocks are compared against each other. For a definite size of file, the number of blocks referenced increases with respect to the smaller block size. Among those referenced blocks the allocated blocks also increases, but it depends on, how much data redundancy those files have.

Every block is given a hash value which is stored in the RAM. As the number of blocks increases the indexing is also increased in an exponential manner. In real world scenario, the file size is much bigger and even a lot of smaller files are stored in the cloud server. Maintaining this large number of indexing in the memory is very complex and expensive because cloud storage providers have to store and maintain millions of indexing in the memory. Each entry (each in core) in the deduplication table takes up some space in memory which is 320 bytes. On the other hand, huge number of blocks are produced from gigabits of data in the cloud

storage. This (320 bytes * number of blocks) produces significant memory usage that leads to memory management problem. The random-access memory price is increased exponentially with increased index. So, the block based deduplication techniques might become a complex way to utilize the storage sizes considering the aspects stated above. If the block sizes are smaller more and more blocks are produced and the index gets very huge in number. [33]On the other hand, in case of file based deduplication technique, each file is given a unique identifier and stored in the index. If another file is uploaded it is also given another identifier which is checked against the previous one to find duplicates. It is a very simple way to remove duplicates but a lot faster and it also takes comparatively less memory management. The whole process gets easier to maintain and evaluate also.

As more blocks are being used in case of changing to smaller block size file system, it will take much time and is much more prone to deduplication. If any attacker wants to use client side cross user deduplication assuming files content and format, cloud providers using smaller block size file system will be more vulnerable to side channel attacks. So, increased deduplication ratio provides more chances of deduplicated files eventually making a favorable scenario for the attacker. The difference in deduplication ratio is seen in case of text file and image files. While deduplicating the image files (Figure 7), the deduplication ratios does not play an effective role with the change of the block size while on the other hand, in case of text file the deduplication ratio changes while changing the block size. In case of image files, the data is already compressed. Deduplication is a lossless compression of files. As the image file is already compressed it will not be deduplicated like the other files. For text files, the file is not complex and there tends to more similarity inside the files. So the files will be deduplicated more than the other files.

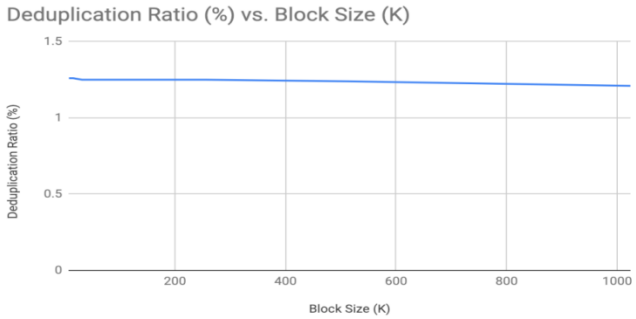

Figure 7: Image File Analysis

As more clearly storage environment which contains text files are prone to the side -channel attack in case of cross user source based deduplication.

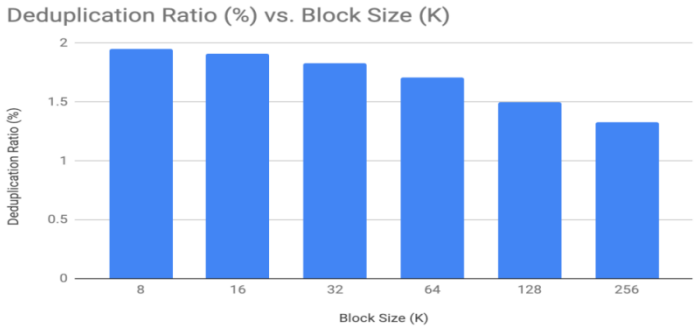

Figure 8: Text file analysis

In the file based deduplication hash value is used. The hash value is created from the stored file. While the data is updated or stored newly, this hash is compared to the new hash value to see if there has been any change. Files containing identical hash will not be stored. If any difference exists, only the newly added portion is sent to storage. Hashes are considered to be unique. If one part of the file is updated, the hash will be changed, so that the service providers can update it.

8. CONCLUSIONS

Data deduplication is a feature which provides efficient storage environment. Using this feature can be risky as it is vulnerable to side channel attack. Source based deduplication provides us with a lot of facilities like bandwidth utilization and server side data storage optimization. Separating the data file and image files can be an effective choice. Moreover, files having less deduplication ratio can be stored in separate file system so that even if the attack takes place not all files are hampered. Moreover, It will help to manage the DDT index. Proper security mechanism can be implemented so that the attack can be stopped immediately. In all respects, it is always a trade-off between performance and feature.

ACKNOWLEDGEMENTS

I would like to thank my family members and my honorable teachers for their inspiration.

REFERENCES

[1] A. Cave, “What Will We Do When The World’s Data Hits 163 Zettabytes In 2025?,” Forbes.[Online]. Available:https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-wedo-when-the-worlds-data-hits-163-zettabytes-in-2025/. [Accessed: 02-Sep-2017].

[2] I. T. L. Computer Security Division, “Cloud Computing | CSRC.” [Online]. Available: https://csrc.nist.gov/Projects/Cloud-Computing. [Accessed: 22-July-2017].

[3] “What is Software as a Service (SaaS)? - Definition from WhatIs.com,” Search Cloud Computing. [Online]. Available:http://searchcloudcomputing.techtarget.com/definition/Software-as-a-Service. [Accessed:22-July-2017].

[4] “What is Platform as a Service (PaaS)? - Definition from WhatIs.com,” Search Cloud Computing. [Online]. Available:http://searchcloudcomputing.techtarget.com/definition/Platform-as-a-Service-PaaS. [Accessed: 22--July-2017].

[5] “What is Infrastructure as a Service (IaaS)? - Definition from WhatIs.com, ”Search Cloud Computing. [Online]. Available:http://searchcloudcomputing.techtarget.com/definition/Infrastructure-as-a-Service-IaaS. [Accessed: 22-July-2017].

[6] “What is data deduplication? - Definition from WhatIs.com,” Search Storage. [Online]. Available: http://searchstorage.techtarget.com/definition/data-deduplication.[Accessed:5- Jul-2017].

[7] C.Poelker, “Data deduplication in the cloud explained, part one,” Computerworld, 20-Aug2013. [Online]. Available:https://www.computerworld.com/article/2474479/datacenter/data-deduplication-in-the-cloud-explained--part-one.html. [Accessed: 14-August2017].

[8] J.D. Burnham– 03.24.15,“Understanding Data Deduplication-- and Why It’s Critical for Moving Data to the Cloud,” Druva, 24-Mar-2015.

[9] C. Poelker, “Data deduplication in the cloud explained, part two: The deep dive,” Computerworld, 08-Oct-2013.[Online]. Available: https://www.computerworld.com/article/2475106/cloud-computing/data-deduplication-inthe-cloud-explained--part-two--the-deep-dive.html. [Accessed: 6-Jul-2017].

[10] G. Schmidt et al.,DS8800 Performance Monitoring and Tuning. IBMRedbooks, 2012.

[11] "Source vs Target Based Data Deduplication." [Online]. Available: http://www.storageswitzerland.com/Articles/Entries/2013/1/2_Source_vs_Target_Based_Data_Deduplication.html. [Accessed: 2-August-2017].

[12] "What is target deduplication? - Definition from WhatIs.com,"SearchDataBackup. [Online]. Available: http://searchdatabackup.techtarget.com/definition/targetdeduplication.[Accessed: 03-Dec-2017].

[13] "What is source deduplication? - Definition from WhatIs.com," SearchDataBackup. [Online]. Available: http://searchdatabackup.techtarget.com/definition/sourcededuplication. [Accessed: 02-Dec-2017].

[14] "What is data deduplication ratio? - Definition from WhatIs.com," SearchDataBackup. [Online]. Available: http://searchdatabackup.techtarget.com/definition/data-deduplicationratio.[Accessed: 5-Sep-2017].

[15] "How to estimate your data deduplication ratio," SearchDataBackup. [Online]. Available: http://searchdatabackup.techtarget.com/tip/How-to-estimate-your-data-deduplicationratio. [Accessed: 22-Oct-201

[16] N. Mandagere, P. Zhou, M. A. Smith, and S. Uttamchandani, "Demystifying Data Deduplication," in Proceedings of the ACM/IFIP/USENIX Middleware '08 Conference Companion, New York, NY, USA, 2008, pp. 12–17.

[17] A. Venish and K. Siva Sankar, Framework of data deduplication: A survey, vol. 8. 2015.

[18] J. Paulo and J. Pereira, "A Survey and Classification of Storage Deduplication Systems," ACM Comput. Surv., vol. 47, no. 1, p. 11:1–11:30, Jun. 2014.

[19] K. Hashizume, D. G. Rosado, E. Fernández-Medina, E. B. Fernandez, "An analysis of security issues for cloud computing", Journal of Internet Services and Applications, vol. 4, no.1, pp. 1-13, 2013.

[20] Harnik, D., Pinkas, B. and Shulman-Peleg, A. (2010) Side Channels in Cloud Services: Deduplication in Cloud Storage. IEEE Security and Privacy, 8, 40-47.

[21] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure Data Deduplication," in Proceedings of the 4th ACM International Workshop on Storage Security and Survivability, New York, NY, USA, 2008, pp. 1–10

[22] "Deduplication as an attack vector" E. Marcus, E. Carl-Henrik, Project Report for Information Security Course, Linköpings Universitet, Sweden

[23] Wang, X., & Yu, H. (2005, May). How to break MD5 and other hash functions. In Annual international conference on the theory and applications of cryptographic techniques (pp. 19-35). Springer, Berlin, Heidelberg.

[24] W. van der Laan, "Dropship", 2011 (https://github.com/driverdan/dropship)

[25] Mulazzani, M., Schrittwieser, S., Leithner, M., Huber, M., & Weippl, E. R. (2011, August). Dark Clouds on the Horizon: Using Cloud Storage as Attack Vector and Online Slack Space. In USENIX security symposium (pp. 65-76).

[26] "Google Docs - create and edit documents online, for free." [Online]. Available: https://www.google.com/docs/about/. [Accessed: 10-Dec-2017].

[27] "Dropbox," Dropbox. [Online]. Available: https://www.dropbox.com/. [Accessed: 10-De-2017].

[28] "Amazon Simple Storage Service (S3) — Cloud Storage — AWS." [Online]. Available: https://aws.amazon.com/s3/. [Accessed: 10-Dec-2017].

[29] C. Kim, K. Park, and K. Park, Rethinking deduplication in cloud: From data profiling to blueprint. 2011.

[30] "How does deduplication in cloud computing work and is it beneficial?," SearchDataBackup. [Online]. Available: http://searchdatabackup.techtarget.com/answer/How-does-deduplication-in-cloud-computing-work-and-is-it-beneficial. [Accessed: 10-Dec-2017].

[31] Rasjid, Z. E., Soewito, B., Witjaksono, G., & Abdurachman, E. (2017). A review of collisions in cryptographic hash function used in digital forensic tools. Procedia Computer Science, 116, 381-392.

[32] https://crypto.stackexchange.com/questions/tagged/birthday-attack

[33] "How to Size Main Memory for ZFS Deduplication." [Online]. Available: http://www.oracle.com/technetwork/articles/servers-storage-admin/o11-113-size-zfs-dedup-1354231.html. [Accessed: 03-Dec-2017].

AUTHORS

AKM Bahalul Haque is currently working a Lecturer of Department of ECE, North South University, Bashundhara, and Dhaka 1229. He has achieved is M.Sc. in Information Technology from Fachhochschule Kiel, Germany, 2018. He achieved his Bachelor of Science (Engineering) in Computer Science and telecommunication engineering in 2014. Has published two of his papers in International Journal. He specializes in Cyber Security, Cloud Computing, Data Privacy and protection. He has one-year experience as Security Engineer and one-year experience as Vulnerability Detection Consultant in Cyber Security Division.