Multilevel Indexing Approach for Multimedia Deduplication for Cloud Data Outsourcing Environment

Rekha B , Vasudeva R

IV th Sem M.Tech , Assistant Professor in Department of Computer Science and Engineering,
C Byregowda Institute of Technology,
Kolar, India.

Abstract: Cloud computing has gained more popularity in the world as it helps store and retrieve the data anywhere in the world. Data deduplication is a technique for wiping out copy of duplicates. In this procedure a comparison of blocks of the same of different images is done against the original copy. This paper is based on storing the multimedia data files like image, text, video, audio etc. in the cloud using DICE protocol. The DICE protocol helps to divide the data into number of blocks based on the size of the files. This also helps to store Near Identical data in the cloud only once instead storing it multiple times by dividing the files into blocks and providing unique hash code to each blocks but multimedia data which of larger files are not given much attention in this paper. Multilevel indexing is the technique used to ease the process of comparing the hash codes of the blocks. Storing the data in the cloud helps to increase the storage time, capacity of the files and provides more security.

IndexTerms **- Security; Deduplication; Cloud Storage.**

I. INTRODUCTION

Distributed computing furnishes clients with the stage to profit cloud administrations on request over the internet which incorporate essentially stockpiling, database, systems administration, and programming administrations. A client examine films, tuning towards sound, taking picture, facilitating sites, making latest applications, distributed computing is a vital piece of every one of these administrations. Cloud specialist organizations (CSP's) charge their clients an ostensible expense for the utilization of these administrations. Subsequently, it is significant for the CSPs to keep up an exchange off between the expense of the transmission capacity use causes charge for the CSP's and administrations expenses they charge to their clients, as putting away and keeping up the tremendous volume of information. De-duplication methods evacuate copy information in addition to along these lines condense transfer speed and capacity necessities and CSPs depends on this method of evaluation. In any case, it is similarly significant to guarantee the protection and security of clients' information for CSP's. To address both these issues, made sure about information de-duplication was introduced, in which copy information is withdrawn while keeping up the privacy of the clients' information. A lot of examination is being done in the field of secure information de-duplication [1], [2]. Most of the protected information de-duplication techniques to set up in the writing delight information from a nonexclusive perspective. In all actuality, information is able to be one of a few unique sorts, for example, text, picture and video information. Truth be told, sight and sound substance, for example, Images and recordings include a significant segment of any information archive, as the pace of sharing has expanded with simple availability of the Internet and the spread of brilliant gadgets. Consequently, deciding copy duplicates in the encoded Multimedia information is a significant test.

A multimedia de-duplication procedure can be both on customer or server side. In the customer side de-duplication system, the calculations are first done on the customer side by creating labels and in server-side de-duplication, the customer transfers document (counting copies) to the server and afterward the server expels the copies and stores the exceptional records in like manner. Simultaneously, the server approves the customer to get to the documents and update the metadata. Thus, the visual projection is higher on the server side and this prompts more data transmission utilization and calculation cost. Both client side and server side have their own ups and downs. The servers send labels instead of the entire document, and tagging is checked for more correspondence. In the Customer Side De-duplication Methodology, this is expected to create less overhead on the server. Subsequently, the customer-side reduction method is more productive, especially when the size of the customer is improving [8].

Agarwal [3] A recently proposed DICE convention that reduces computational, remote and data transfer capability requirements by connecting only one tag. At this meeting, much of the counting was done on the customer side. There is a reduction in the record level, where the copy documents are separated by applying the hash capabilities to the entire document and then checking whether the hash estimates are the same. Based on multimedia information, it may not be appropriate to hash out all the multimedia information, as the price of the hash may be the opposite, even though the two Multimedia differs respectfully from one pixel to the other. Or, we can divide multimedia documents into categories, first by hashes, and then by looking at the similarity between the hash benefits of the respective classes, which is the strategy we have in this paper.

II. RELATED WORK

T. Jiang, X. Chen, Q. Wu, J. Ma, W. Susilo, and W. Lou proposed Towards effective completely randomized message-locked encryption [2]. Cross-user information deduplication will significantly decrease the capacity taken a toll of capacity benefit supplier. Spurred by secure information deduplication, Abadi et al. augmented the work Message-Locked Encryption (MLE) and

presented the primitive of MLE2 with decent security properties. Be that as it may, their completely randomized conspire (RMLE2) requires the wasteful equality-testing calculation to recognize all copy ciphertexts. In this way, an curiously open issue is how to diminish the overhead of R-MLE2 and propose an proficient development for R-MLE2. In this paper, we present a unused primitive called μR-MLE2, which gives a halfway positive reply to this open issue. Our primary trap is to utilize the client-assistant way based on inactive or energetic choice trees. The advantage picked up from it is that by association with clients, the server will decrease the time complexity of deduplication balance test from straight time to productive logarithmic time over the whole database items.

M. Bellare, s. Kelvedi, and T. Ristenpart Determined by the message. While MLE offers a way to realize secure commitment (space-efficient secure outsourcing capability), this goal is currently concentrated by most cloud capacity suppliers [5]. We give definitions for security and for the quantity of criteria called tag stability. Based on this establishment, we make pragmatic and hazardous commitments. On the feasible side, we offer a common family of ROM security tests of MLE schemes that contain communicated plans. On the other side, the challenge is the standard display approach, and we have combined decisive encryption, preserved hash capabilities on related inputs, and tested the delivery of messages with specific suspicion and capture a global perspective for different types of messaging resources. Our work suggests that MLE is primitive for both common sense and inary rational conspiracy.

M. Bellare and S. Keelveedhi proposed about intelligently message-locked encryption and secure deduplication [6]. This paper considers the issue of secure capacity of outsourced information in a way that licenses deduplication. We are for the primary time able to supply protection for messages that are both connected and subordinate on the open framework parameters. The modern fixing that creates this conceivable is interaction. We increment the message-bolted encryption (MLE) crude of prior work to brilliantly message-bolted encryption (iMLE) where move and download are shows. Our conspire, giving security for messages that are not only related but permitted to depend on the public framework parameters, is within the standard show. We clarify that interaction isn't an additional presumption in hone since full, existing deduplication frameworks are as of now interactive.

J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl proposed about Secure Information Deduplication [7]. As the world moves to advanced capacity for authentic purposes, there's an expanding request for frameworks that can give secure information capacity in a cost-effective way. By distinguishing common chunks of information both inside and between records and putting away them as it were once, deduplication can surrender taken a toll reserve funds by expanding the utility of a given sum of capacity. Shockingly, deduplication abuses indistinguishable substance, while encryption endeavors to form all substance show up arbitrary; the same substance scrambled with two diverse keys comes about in exceptionally diverse cipher text. In this way, combining the space effectiveness of deduplication with the mystery angles of encryption is tricky. We have created a arrangement that gives both information security and space productivity in single-server capacity and disseminated capacity frameworks. Encryption keys are created in a reliable way from the chunk information; hence, indistinguishable chunks will continuously scramble to the same cipher text. Moreover, the keys cannot be derived from the scrambled chunk information. Since the data each client must get to and unscramble the chunks that make up a record is scrambled employing a key known as it were to the client, indeed a full compromise of the framework cannot uncover which chunks are utilized by which clients.

J. R. Douceur, A. Adya, W. J. Bolosky, P. Simon, and M. Theimer proposed about Recovering Space from Copy Records in a Serverless Conveyed Record System The Distant location conveyed record framework gives accessibility by duplicating each record onto different desktop computers [9]. Since this replication devours critical capacity space, it is vital to recover utilized space where conceivable. Estimation of over 500 desktop record frameworks appears that about half of all devoured space is possessed by copy records. We show a component to reclaim space from this coincidental duplication to form it accessible for controlled record replication. Our component incorporates. 1)merged encryption, which empowers copy records to coalesced into the space of a single record, indeed on the off chance that the records are scrambled with diverse users' keys, and 2)Serving of mixed greens, a Self-Arranging, Lossy, Acquainted Database for conglomerating record substance and area data in a decentralized, versatile, fault-tolerant way. Large-scale re-enactment tests appear that the duplicate-file coalescing framework is adaptable, exceedingly successful, and fault-tolerant.

Gang et al. [10] considered a total multimedia reduction, and implemented the CE platform in conjunction with feature-based encryption to perform multimedia reduction. The secure deduplication of multimedia has been inspired by some experts. Fatima et al. [11] SPIHT pressure features and semi-encryption have been used to succumb to multimedia hashing. The Fractional Encryption Trust provides protection from CSP and helps the multimedia hashing process detect the intangible package and encode multimedia. In his work, the client first applied multimedia pressure calculation, using partial encryption at the time, which eventually led to the multimedia head signature. That mark is sent to the CSP for exemption.

A paper called Client-Based Security-Producing De-duplication of Multimedia Data (CSPD), which investigates the duplication of alcohol strategies used by multimedia, was proposed by Li et al [12]. As with various methods of performing repeated checks on multimedia technology, in their method, after the client has transferred some of the limitations of the multimedia to the CSP, the CSP is based on the limitations of putting the multimedia and applies the hamming isolation for scan database for Comparative multimedia documents.

Juan Li et al [13]. proposed for the preservation of preserved perceptual equality. Suppose they determine the likelihood of a multimedia hash in the cloud by calculating the Hamming partition between the multimedia hashes that are put away where the put count is used. Additionally they use a collection key, where all members of a conference can transfer and download multimedia using the transfer key. The keys collected to share information with customers can only be kept and used by customers. Their scheme for basic multimedia management operations is similarly powerful, for example, in size and pressure.

III. PROPOSED WORK

The close to indistinguishable Multimedia situation is characterized as at least two Multimedia which have a similar foundation as appeared in Fig 1, however there is either an adjustment in a specific square, or a portion of the pixels are unique.

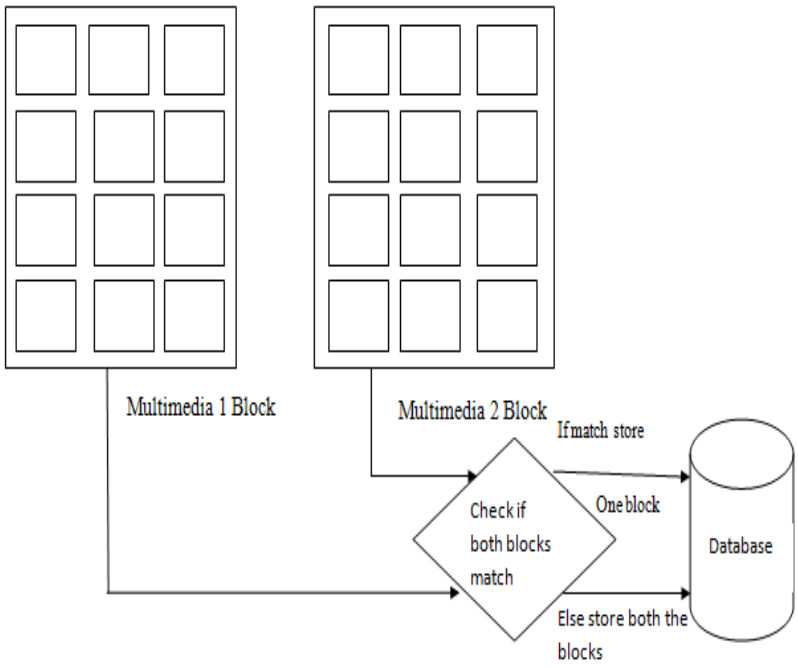

Figure 1: Block level deduplication of two Multimedia's

A. δ-NI Multimedia:

Multimedia pair (I, I_) where I is the block from multimedia 1 block and I' is the block from multimedia 2 block is called δ-NI (or δ-close to indistinguishable), if the proportion of hinders that are same in I and I_ is δ (I, I_) ∈ [0, 1]. For the protected deduplication of δ-NI Multimedia's, square astute change of the Multimedia is applied and checked if the two Multimedia coordinate square savvy, as appeared in Figure 1.The Multimedia are coordinated by planning the principal square of the Multimedia with the primary square of the subsequent Multimedia and keeping doing this until the last square. The Multimedia are changed over square savvy, at that point the DICE convention is run on the squares. For the obstructs that coordinate, just one duplicate of the square on the distributed storage is kept as appeared in Fig 1.

B. Secure Deduplication of NI-images with DICE

In this section we provide a detailed view of the secured block level image deduplication strategy based on the DICE protocol. In the following, we first describe the system and threat models and discuss the assumptions made in the DICE protocol when applied to NI-images, and then we present the adaptation of the DICE protocol for NI-images. We call this new protocol DICE-NI.

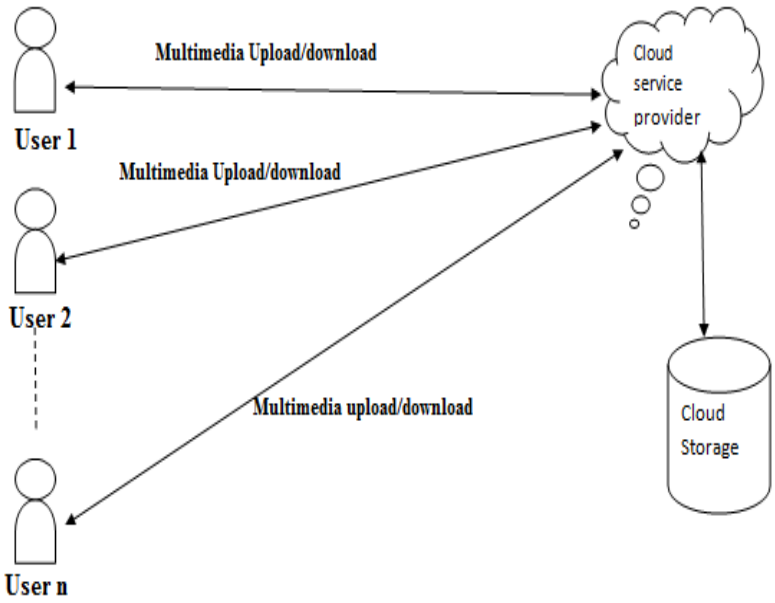

Fig 2: System Overview

1) System model:

The framework model is included clients and the CSP, where there could be various clients getting to the cloud to transfer or download Multimedia, as can be found in Figure 2.
Customer Service Point: The CSP gives the capacity administrations to the clients who have been endorsed get to and are in this way approved to utilize the cloud administrations. They are charged an insignificant expense.

Users: Users are the individuals who are approved to get to the cloud administrations. They have the scrambled record, which in this situation is included a Multimedia that they transfer to the CSP utilizing the DICE convention. Before transferring the Multimedia square insightful, the customer checks for the accessibility of the squares at the cloud. In the event that a square exists, once a connection is given to the specific client, in any case a solicitation to transfer the square is sent.

2) Threat model: In this section we consider the threats from the aspect of an adversary, where they are interested in knowing the content of the images. We consider the adversary to be malicious, the CSP to be semi-malicious and the user to be honest. An adversary could be an internal or an external adversary. An internal adversary is more interested in knowing the content of a file that might belong to the CSP. As a result, we consider the CSP to be semi-honest. Conversely, an external adversary is interested in both the content and the owner of the file. In this scenario, the threats could be generated by directly accessing the cloud services or by gaining access to the channel. If the attacker gains access to the cloud, they could try to erase the content of the file or replace it with a different file. But irrespective of the objective of the adversary, the protocols used by the CSPs should be secure enough to guarantee the detection and prevention of those attacks.

3) DICE-NI protocol:

The client first partitions the picture into a fixed number of block. Each block size could be of variable length, somewhere in the range of 4×4, 8 ×8to16 ×16. In the way of changing over the multimedia into obstructs, the client runs the customer segment of the DICE convention on each block. According to the DICE convention, the customer processes the key Ki as Ki ← H(Bi) where H is the hash capacity, and Bi is the ith block of the multimedia. Next, the customer processes the cipher text Ci as Ci ← E(Ki ,Bi) and the label T as Ti ← H(Ci) , where E is the encryption methodology.

In the way of figuring the keys, the cipher text and the labels, the client acquires the accompanying vector {K11 ,K12 , ...Kmn }, {C11 ,C12 , ...Cmn }, {T11 ,T12 ,...,Tmn } for a multimedia, where mn is the complete number of blocks. At this stage, the client sends the label vector {T11 ,T12 , ...Tmn } to the CSP and checks for its reality in the cloud as shown in Fig 3. The CSP at that point runs a quest in the label store for the presence of the labels from the label vector and sends a solicitation for just those blocks for which no match was found. The customer at that point sends the cipher text of those specific blocks to the CSP, who stores them alongside the client's accreditations and updates its label store by figuring T' ← H(Ci) .

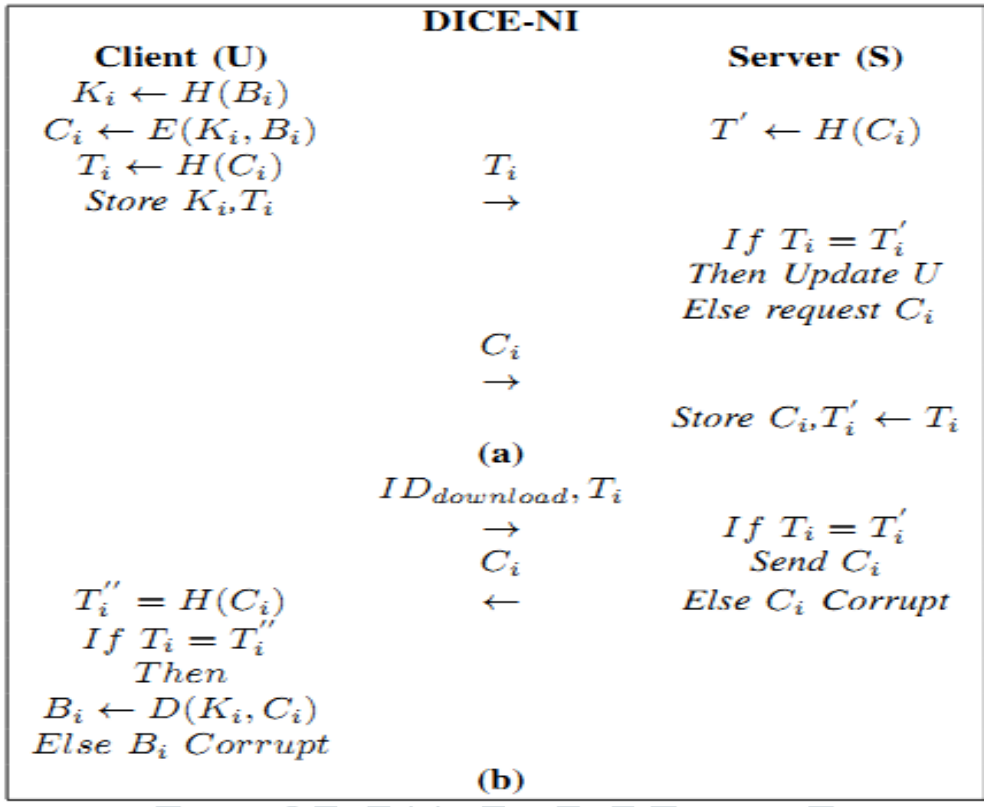

Fig 3: DICE – NI protocol (a) Upload (b) Download

At the hour of download, label vector and userid is sended by user, and the CSP look through its label store to find the relating tag and cipher text obstruct as Ti = Ti' . In the event that there is a match discovered, at that point the relating cipher text block is sent to the individual client, in any case the CSP sends an affirmation that the picture isn't found. After the client gets the cipher text, the tag is registered from the got cipher text obstruct as Ti" = H(Ci) and is coordinated with the put away tag as Ti = Ti". In the event that there is a match discovered, at that point the decoding procedure begins as Bi ← D(Ki ,Ci) , where D is the unscrambling system; in any case the client sends an affirmation to the CSP that the block has been corrupted.

IV. PERFORMANCE ANALYSIS

The total time taken for DICE-NI execution for multimedia varies in block sizes: 4 × 4, 8 × 8 and 16 × 16. It is evident that the block size is directly proportional to the execution time of DICE-NI; smaller blocks take more time to execute than the larger blocks. Although larger blocks take less execution time, they may lead to other issues as shown in Fig 4. For example, larger blocks could result in more data being stored at the CSP, or the users may not be able to retrieve the entire multimedia properly at the time of download due to the loss of some pixel values at the time of reconstruction. Although smaller block sizes could help to achieve better deduplication, the tags and the keys generated and stored at the client side are of fixed size which is irrespective of the block size. We do not want the block size to be too small because it may defeat the whole purpose of saving storage space. With a varying block size, however, some nearly identical images achieve better deduplication results, while other images do not.

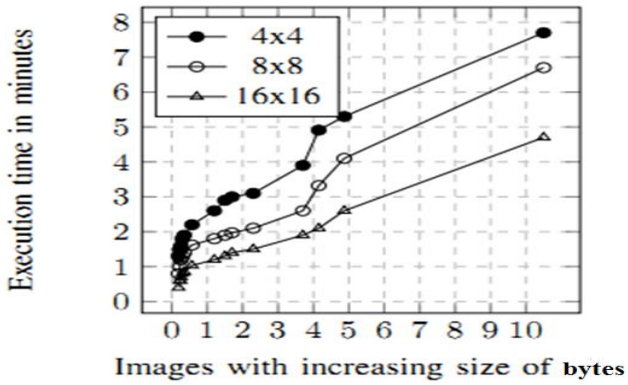

Fig 4: Comparison of DICE-NI execution time with varying number of block size

V. CONCLUSION

A Multimedia deduplication helps to store the near identical data only at once in the cloud rather than storing in multiple times to increase the space, storage capacity and to reduce the cost. The data are divided into no of blocks based the size of the files and each file is given a unique hash code using this hash code the cloud compares that whether the blocks are already present or not. If the blocks are matched then the blocks are stored only at once in the cloud instead of storing it again as user can only known the details of blocks, user and clouds it is more secure. In the future we can implement this paper for storing the larger multimedia data files by using fast internet.

REFERENCES

[1] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, "6 million spam tweets: A large ground truth for timely twitter spam detection," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 7065–7070.

[2] T. Jiang, X. Chen, Q. Wu, J. Ma, W. Susilo, and W. Lou, "Towards efficient fully randomized message-locked encryption," in *The 21st Australasian Conference on Information Security and Privacy*, Melbourne, VIC, Australia, 2016, pp. 361–375.

[3] D. Koo, J. Hur, and H. Yoon, "Secure and efficient deduplication over encrypted data with dynamic updates in cloud storage," in *Frontier and Innovation in Future Computing and Communications*, Dordrecht, 2014, pp. 229–235.

[4] A. Agarwala, P. Singh, and P. K. Atrey, "DICE: A dual integrity convergent encryption protocol for client side secure data deduplication," in *IEEE International Conference on Systems, Man, and Cybernetics*, Banff, Canada, 2017, pp. 2176–2181.

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Messagelocked encryption and secure deduplication," in *Advances in Cryptology – 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Athens, Greece, 2013, pp. 296–312.

[6] M. Bellare and S. Keelveedhi, "Interactive message-locked encryption and secure deduplication," in *Public-Key Cryptography – 18th IACR International Conference on Practice and Theory in Public-Key Cryptography*, Gaithersburg, MD, USA, 2015, pp. 516–538.

[7] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," in *Financial Cryptography and Data Security*, Berlin, Heidelberg, 2014, pp. 99–118.

[8] M. W. Storer, K. Greenan, D. D. Long, and E. L. Miller, "Secure data deduplication," in *Proceedings of the 4th ACM International Workshop on Storage Security and Survivability*, Fairfax, Virginia, USA, 2008, pp. 1–10.

[9] J. R. Douceur, A. Adya, W. J. Bolosky, P. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," in *The 22nd International Conference on Distributed Computing Systems*, Vienna, Austria,, 2002, pp. 617–624.

[10] K. Keonwoo, Y. Taek-Young, J. Nam-Su, and C. Ku-Young, "Client-side deduplication to enhance security and reduce communication costs." *ETRI Journal*, vol. 39, no. 2, pp. 116–123, 2017.

[11] H. Gang, H. Yan, and L. Xu, *Secure Image Deduplication in Cloud Storage*. Cham:Springer International Publishing, 2015, pp. 243–251.

[12] D. Li, C. Yang, C. Li, Q. Jiang, X. Chen, J. Ma, and J. Ren, "A client-based secure deduplication of multimedia data," in IEEE International Conference on Communications, Paris, France, 2017, pp. 1–6.

[13] X. Li, J. Li, and F. Huang, "A secure cloud storage sys-tem supporting privacy-preserving fuzzy deduplication," Soft Computing, vol. 20, no. 4, pp. 1437–1448, 2016.