Review of Various Encryption and Compression Mechanisms Used within Deduplication in Cloud Computing

Nancy Digra[1], **Sandeep Sharma**[2]

Post Graduate Student, Department of Computer Engineering and Technology, Guru Nanak Dev University, Amritsar, Punjab, India

Abstract: ***Data deduplication is the mechanism by which redundant data is eliminated from advanced computing architecture. Cost in advanced computing is due to storage use. Compression is the need of the hour to save bandwidth and storage space to reduce overhead. Security is also at stake since large number of users is using cloud. To enhance security, encryption mechanisms are utilized within deduplication process. Security techniques within deduplication are critical. Researchers are working towards this issue and uses different encryption mechanisms within deduplication. This paper highlighted the studies of various techniques which are part of deduplication. Comparative analysis is also presented to enhance determine optimal technique which can be used in future to enhance deduplication performance.***

Keywords: Data Deduplication, Cost, Cloud, Security, encryption

1. Introduction

Cloud computing is fast growing mechanism of using the Internet to provide software and other IT services to the user as and when desired. Users will share legions of resources online through cloud. The critical feature of cloud is lack of centralized ownership. Any user having desired resources can access cloud without restriction. Users could be many and having distinct intentions. So obligation of security exists. User can share bandwidth, power, storage space and software so cloud will allow machines to execute resource beyond the capacity of the machines [1]. Resources are provided on the basis of pay per use. Cloud computing provides user with resources such as hardware, software and information so that user can complete its task without possessing such resources dedicatedly. These resources are provided on pay per use basis hence overhead should be reduced to save cost associated with such resources. One such mechanism for reducing overhead is deduplication. Deduplication is the mechanism by which same data is not transferred over the cloud again and again hence storage is saved. Encryption mechanisms are implemented to ensure security. This is done to guard the data against the malicious users. Hence there are two aspects associated with deduplication.

- Redundancy handling in terms of compression
- Security in terms of encryption

Redundancy handling mechanisms involves both lossless and lossy compression mechanisms. Security in terms of encryption is provided through mechanisms such as RSA, DES, Digital signatures etc. The proposed work deals with the study of all listed techniques used within data deduplication. Rest of the paper is organised as follows: section II provides details of cloud computing, section III provides description of compression and encryption mechanisms, section IV provides details of types of deduplication, section V provides comparison of deduplication mechanisms.

2. Cloud Computing

The increasing demand of resources cannot be coped through single physical machine. As technology enhances techniques are furnished to accommodate resource sharing mechanism. The resource sharing and data storage capabilities become need of the hour which is provided through cloud computing. Cloud utilizes layered approach to disperse services to users. Each layer deals with distinct service.

3. Infrastructure as a Service

The critical and basic part of cloud computing is infrastructure. The basic service model included in this layer is virtual machines. Internet is integral part of cloud computing IaaS. Cloud and internet are utilized side by side. They are related since both possess lack of centralized ownership. Users having required resources can access both of these mechanisms. Internet provide gateway to store and retrieve data from within cloud. [2]Load balancing is also virtue of this layer. In cloud data centers operate by processing requests of resources and data. The load on data center having optimal performance could be large. This load degrades the performance of data center. To cope with this situation load has to be dispersed among multiple data centers which are selected on optimal performance basis. Load balancing strategies are many including genetic algorithms. The genetic algorithm disperses the load by selecting and comparing various data centers present over the cloud.[3]

Platform as a service

Cloud also provides platform services to the machines which does have compatibility issues. Some software applications require higher configuration than provided by the current machine. The problem is resolved using cloud since it provides all the required resources to devices having compatibility problems. Web servers, run time databases etc is also provided by this layer.[4] The integration flows enables customer to execute applications requiring compatibility.[5]

4. Software as a Service

This layer provides application software and database as a service to the user. Cloud user does not have access to place where software and other applications run. In order to access the applications where they are executed, price needed to be paid. Generally this service is paid. [6]The user needs to pay for amount of service they need. This amount is calculated in terms of bandwidth access given to the user. Pricing could be monthly or yearly.[7]

The base of cloud computing is based on these layers. Failsafe mechanisms are critical at every layer of cloud. Next section discusses the compression and encryption mechanisms used within deduplication mechanisms.

5. Encryption and Compression techniques Used Within Deduplication

Encryption[8]n is the mechanism used to provide security mechanism within cloud computing. Encryption uses keys in order to encrypt and decrypt the data. [9]Public key is used in order to encrypt the data at sender end and private key is used to decrypt the data at receiver end. [1]Encryption and decryption mechanisms utilize cipher text. Ciphers are divided into following categories.

- Transposition Cipher
- Substitution Cipher

Transposition cipher changes the position of characters within data to form cipher text to be transmitted. E.g suppose user is transmitting “ABC” then in transposition cipher position of characters are altered and new text to be transmitted may becomes “CAB” or “CBA”. Total of 7 possible combinations can be generated.

[10]Substitution Cipher changes the characters with certain characters. E.g suppose user transmitted “ABC” then in this case “123” may be the cipher which forms after applying substitution cipher.

Various encryption and decryption techniques using ciphers to transfer the data from source to destination is described in this section.

[11]RSA algorithm is used in order to provide security in cloud. The cloud is accessed by legions of users. The intention of the users may be uncertain. So the encryption standards are required in order to avoid the malicious attacks. This can be accomplished by the use of RSA algorithm. The RSA provide the security in private cloud domain. The RSA cloud trust authority is created for this purpose.[12] The RSA cloud trust authority is set of cloud based policies to check information and infrastructure related issues within the cloud. The transmitted information within the private cloud is checked against the compliance rules. The compliance rules if satisfied then only the information is accepted otherwise the information will be rejected. [13] RSA cloud security software are also available to ensure security within the cloud. The cloud security system utilizes RSA token based system. The token will be given to the packet satisfying certain constraints assisted with the security premises. The RSA algorithm will utilize the plain text. The plain text will be given to the encoder and then cipher text will be obtained. The cipher text will be encoded information. The encoding will takes place according to prime numbers which are assumed at the sender and receiver end. The information which is transferred will be in encoded form which is not understood by the malicious node if any. The process is described as

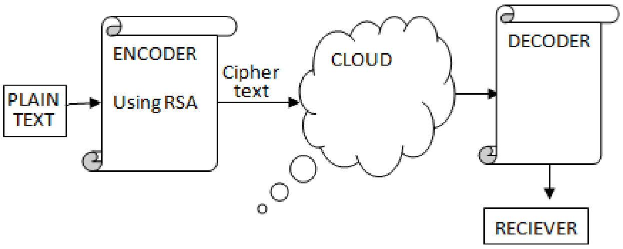

Figure 1: Describing encryption process and transfer process to cloud along with decryption process

The receiver will get the decoded information. The decoded information will be received by the receiver in the user understandable form. The information in the cloud will remains in the encoded form where the malicious attacker may distort the data. The concept of parity bits are also used in this case which will detect the disturbance in the bit formation. Another slandered which is commonly used in cloud security is Data Encryption slandered. This is a symmetric method of encryption.[14] DES was created in order to avoid brute force attacks within the system. The public key method is used in order to ensure the prevention of malicious attacks within the cloud. Through the DES cloud control matrix is used. The public keys are shared among source and destination. The size and compliance related information is contained within the cloud control matrix. The cloud control matrix contains logical values. The values can be true or false. The most of the values should be true in order for the packet to be accepted. The packets which are accepted are said to be compliance. The DES will be used by dividing the data part into subsections each subsection will be checked separately. The code block is generally divided into block of 32 bits each. The keys in this case are known as sub-keys.

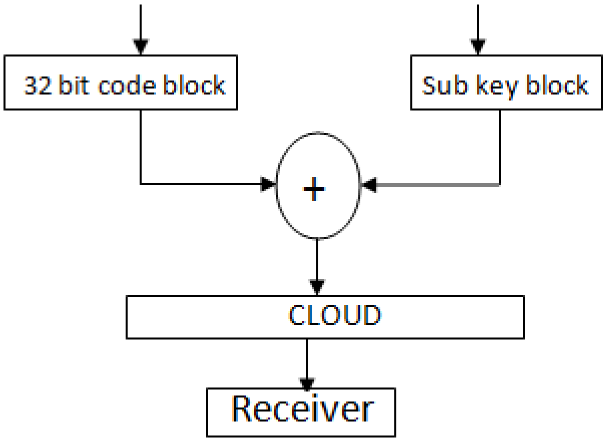

Figure 2: Partitioning of code block along with sub key

The DES algorithm is quite useful in a situation where simple encoding is required within the cloud. The parity bits are not used in this case hence security is least in this system.

The DES is replaced by more advanced standard of encryption known as advanced encryption slandered. The AES is used in the cloud in order to ensure the attacks on data do not occur.

[15][16][17][18]With the tremendous growth of sensitive information on cloud, cloud security is getting more important than even before. The cloud data and services reside in massively scalable data centers and can be accessed everywhere. The growth of the cloud users has unfortunately been accompanied with a growth in malicious activity in the cloud. More and more vulnerabilities are discovered, and nearly every day, new security advisories are published. Millions of users are surfing the Cloud for various purposes, therefore they need highly safe and persistent services. The future of cloud, especially in expanding the range of applications, involves a much deeper degree of privacy, and authentication. We propose a simple data protection model where data is encrypted using Advanced Encryption Standard (AES) before it is launched in the cloud ensuring confidentiality of information. The AES and cloud utilised certification agencies. The certification agency will identify the malicious node and allocate the tag as malicious. The malicious sender will be blocked and not allowed to transmit the data forward. The data and packets which are transmitted to the cloud will be checked for maliciousness and certification agency and monitoring nodes will play the role of patrolling parties.

[19]Deffie hellman algorithm is another mechanism to encrypt the data. Encryption technique following deffie hellman provides more secure encryption as compared to other encryption mechanisms. Prime numbers are used in this case to generate keys. Secret keys are used to decrypt the data. Public keys are used to during encryption. Hence this algorithm uses both public and private keys for the purpose of encryption and decryption.

Compression mechanism is employed in order to save space within advanced computing to save from space overhead consumption. Various compression techniques available to be used in cloud computing are described in this section.

In deduplication only duplicate copies of the data is identified and eliminated hence only one copy of the data is maintained. [20], [21]Hence more storage space is saved by using compression with deduplication. By using compression within deduplication, data density is enhanced. Legions of advantages introduced when compression is used along with deduplication.

- Less storage space requirements
- Faster read and write cycle time.
- Transfer speed is enhanced.
- Overhead in terms of cost is reduced.
- Storage capacity of medium almost becomes doubled the original capacity.

With deduplication mainly lossless data compression techniques are used. Lossless compression techniques involve Huffman compression is one such technique used to compress the data which is based on lossless compression. In Huffman compression mechanism probability of character occurring is determined. Higher the probability of occurrence less bits is required to represent character. Statistical coding is used within Huffman coding scheme. Probability of occurrence of characters has direct impact on length of bit representing the character. Considering a file, certain characters occur more than other characters. In such situations, binary representation is used. „0‟ is used to represent first character and „1‟ is used to represent second character. Using two bits we can represent four characters and so on. Huffman compression is a variable length coding scheme in which longer code is assigned to less frequently occurring characters and smaller codes are assigned to more frequently occurring characters.

[22]–[25]Run length encoding is another compression technique which is used to compress the data within deduplication mechanism. In RLE scheme characters repeated are replace with the number representing number of times characters repeats. E.g character string „abcabc‟ can be represented in RLE as „a2b2c2‟. Length of data is sufficiently reduced using RLE. Longer the length of characters shorted will be length of compressed data. Hence length of original data has trade off with compression ratio.

In future, collaboration of Deffie Hellman along with deduplication can be used to increase the performance of the system. More security can be provided using Deffie Hellman in deduplication mechanism.

6. Deduplication Mechanisms

Deduplication mechanism is used to provide effective space utilization along with encryption mechanism. Deduplication mechanisms available are described in this section.

- Post Process Deduplication
- In Line Deduplication
- Source Deduplication
- Target Deduplication

[13]Post process deduplication mechanism allows the data to store within the storage device. After storage deduplication mechanism is applied in order to reduce the storage space consumption. The advantage is that storage performance is not degraded. Disadvantage is that redundant data is stored with storage device even for small period of time.

Inline deduplication is the mechanism in which hash calculations is performed on target machine. If it detect the block which is already stored with the disk then that block is not stored again. Storage space is efficiently used which is a merit of this technique. the disadvantage is that hash calculations takes long time. Hence overall process is slow.

Source deduplication is the technique in which deduplication is performed close to the location where data is created. It is performed within the file system. Hashes are compared against already present hashes to ensure duplicated hashed are not again stored within memory. This process is slow in nature but efficient in terms of storage.

Target deduplication is performed at a location where files are stored. It is efficient since operation is performed without has function calculations. Security mechanisms are poorly utilized in this approach.

7. Comparison of Deduplication Mechanims

This section provides the comparison of various techniques associated with deduplication. Various parameters are also considered in this case.

Table 1: Comparison in terms of parameters

Paper	Parameters used	Parameter Usage	Gaps
[26]Online lazy migration and randomized fixed horizon control	Average latency Cost Execution time	1. Minimization of cost 2. Execution time is reduced	No fault tolerant mechanism is considered
[27]Load balancing on virtual data centre	Load Balancing Round trip time Bandwidth	1. Efficient Load balancing 2. Execution time is reduced	Performance studies are exempted from research
[28]Dynamic partially reconfigurable system	Resource utilization Cost Time	1. Efficient resource utilization 2. Cost and time assonated with execution is reduced	Faults considered at run time only
[29]Securityin real time cloud computing	Fault rate Pass rate	1. Handling faults in IaaS of Cloud 2. Reducing fault rate and enhancing pass rate	Recovery mechanism does not include voluminous data
[30]Security and QoS in CAN	QOS Replication Load Balancing	1. Quality of service is ensured 2. Load balancing for Securityis designed	No voluminous data is handled and technique may not produce optimal result in case of Big Data
[31]β Misclassification	Reduction Cores Decision rules	1. Effectively use Securityin bijective soft set. 2. Effectively discover fault data	No voluminous data is considered
[32]Cloud Therapy and Entropy weight	Entropy Accuracy	1. Cloud therapy is efficient technique for Security related to hardware 2. Accuracy is included as parameter hence system is efficient	Big Data cannot be used in this environment since soft set faults is not considered.
[33]Check pointing and restart approach	Overhead Bandwidth	1. Overhead reduced 2. Bandwidth is least utilized	Big data analysis is not considered

Table 2: Comparison of various deduplication mechanism used in cloud computing

Reference And Title	*Technqiues*	*Merits*	*Demerits*
[26]Public Auditing for encrypted data at client side deduplication in cloud	Client side deduplication	Overhead in storage is reduced.	Data is redundantly stored for small period of time
[27] HEDup: Secure Deduplication with Homomorphic Encryption	Target Side deduplication	Overhead in terms of hash function is reduced	Hash function at target side consume more time
[28] A secure cloud storage system supporting privacy-preserving fuzzy deduplication	Client side deduplication	Fuzzy deduplication provide more secure environment however consume more space	Consume more space
[29] BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication	Block Level deduplication	Storage is efficiently utilized	Calculations of hash function consume more time
[30] Dynamic Data Deduplication in Cloud Storage	Dynamic deduplication approach	Client side deduplication approach is supported for better performance in terms of speed	Storage space is consumed heavily.
[31] Secure Enterprise Data Deduplication in the Cloud	Secure Client side deduplication	Security enhancement is the target of this approach	Storage space is not efficiently used
[32] Data Deduplication Cluster Based on Similarity-Locality Approach	Cluster based similarity locality approach	Storage space is efficiently used	Overhead in terms of cost is high
[33] A secure data deduplication framework for cloud environments	Security based mechanism for cloud is used	Security is a key feature is introduced	Overhead in terms of cost is high
[34] ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage	Encryption Based technique is used to enhance security	Security is greatly enhanced	Redundancy is not handled efficiently
[35] Hybrid data deduplication in cloud environment	Hybrid approach	Security as well as deduplication is efficiently performed	Overhead can further be reduced
[36] A Hybrid Cloud Approach for Secure Authorized Deduplication	Hybrid Approach	Secure and efficient deduplication	Overhead is high
[37] Secure Deduplication with Efficient and Reliable Convergent Key Management	Convergent key mechanism	Security is enhanced	Redundancy is not handled properly
[38] Private data deduplication protocols in cloud storage	Private key mechanism	Security is of prime concern in this approach	Redundancy is not handled properly
[39] Exploiting Data Deduplication to Accelerate Live Virtual Machine Migration	Deduplication to enhance Live VM migration	Storage is efficiently handled	Security is poorly handled

[40] An Intelligent Data De-duplication Based Backup System	Intelligent data deduplication mechanism	Storage space is efficiently handled	Cost due to intelligent behaviour is more
[13] Optimized Storage Approaches in Cloud Environment	Optimized storage mechanism	Storage is efficiently managed	Cost encountered is high

8. Conclusions and Future Scope

Cloud and advanced computing architecture support resource sharing so that users with limited resources can access resources to complete their tasks. Users of distinct nature access the cloud. Cloud storage is shared on the basis of pay per use basis. Storage hence has to be efficiently used. This is accomplished by the use of deduplication mechanism. The proposed work comprehensively describes various techniques used within deduplication. Both compression and encryption mechanisms are critical in this approach.

In the future approaches described in this paper can be hybridised to improve the performance in terms of security and compression. Deffie hellman algorithm is found to be most secure as compared to other security mechanism hence this approach can be used along with deduplication mechanism.

References

[1] A. Kumar, K. Bhushan, and M. C. Pandey, “SECURE ENVIRONMENT FOR CLOUD COMPUTING USING MODIFIED THIRD PARTY AUDITING (TPA) SYSTEM,” pp. 339–345.
[2] Y. Charalabidis, M. Janssen, and O. Glassey, “Introduction to Cloud Infrastructures and Interoperability Minitrack,” p. 7695, 2012.
[3] M. Janssen, “Introduction to the Cloud Infrastructures and Interoperability Minitrack,” p. 2013, 2013.
[4] C. Pahl, S. Helmer, L. Miori, J. Sanin, and B. Lee, “A Container-based Edge Cloud PaaS Architecture based on Raspberry Pi Clusters,” 2016.
[5] R. Glitho, M. Morrow, and P. Polakos, “A Cloud Based - Architecture for Cost-Efficient Applications and Services Provisioning in Wireless Sensor Networks,” 2013.
[6] M. Armbrust, A. D. Joseph, R. H. Katz, and D. A. Patterson, “Above the Clouds : A Berkeley View of Cloud Computing,” 2009.
[7] J. Footen, A. V. P. B. Consulting, C. T. Solutions, and F. W. B. Blvd, “Service Oriented Architecture & Cloud Computing in Media Industry,” vol. 1100, 2011.
[8] G. L. Prakash, M. Prateek, and I. Singh, “Data encryption and decryption algorithms using key rotations for data security in cloud system,” Int. Conf. Signal Propag. Comput. Technol. (ICSPCT 2014), vol. 3, no. 4, pp. 624–629, 2014.
[9] S. S. M. Chow and R. H. Deng, “Key-Aggregate Cryptosystem for Scalable Data Sharing in Cloud Storage,” IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 2, pp. 468–477, Feb. 2014.
[10] “Polyalphabetic Substitution Ciphers,” in Coding for Data and Computer Communications, New York: Springer-Verlag, 2005, pp. 243–267.
[11] K. Govinda and E. Sathiyamoorthy, “Identity Anonymization and Secure Data Storage using Group Signature in Private Cloud,” Procedia Technol., vol. 4, pp. 495–499, 2012.
[12] S.-H. Yang, “WSN Security,” pp. 187–215, 2014.
[13] I. Journal and O. F. Engineering, “Optimized Storage Approaches in Cloud Environment,” vol. 3, no. 12, pp. 601–605, 2014.
[14] Intel IT Center, “Planning Guide Cloud Security - Seven Steps for Building Security in the Cloud from the Ground Up,” no. May, 2012.
[15] K. Yang and X. Jia, “Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage,” IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 7, pp. 1735–1744, Jul. 2014.
[16] N. Ahmad, A. Kanwal, and M. A. Shibli, “Survey on Secure Live Virtual Machine (VM) Migration in Cloud,” no. Vm, pp. 101–106, 2013.
[17] O. Harfoushi, B. Alfawwaz, N. a. Ghatasheh, R. Obiedat, M. M. Abu-Faraj, and H. Faris, “Data Security Issues and Challenges in Cloud Computing: A Conceptual Analysis and Review,” Commun. Netw., vol. 06, no. 01, pp. 15–21, 2014.
[18] R. Marchany, V. a Tech, and I. T. Security, “Cloud Computing Security Issues Something Old , Something New,” Computer (Long. Beach. Calif)., 2010.
[19] Y. Rahulamathavan, S. Veluru, J. Han, F. Li, M. Rajarajan, and R. Lu, “User Collusion Avoidance Scheme for Privacy-Preserving Decentralized Key-Policy Attribute-Based Encryption,” vol. 9340, no. c, 2015.
[20] T. Y. J. Nagamalleswari, “Deduplication Techniques : A Technical Survey,” vol. 1, no. 7, pp. 318–325, 2014.
[21] D. T. Meyer and W. J. Bolosky, “A Study of Practical Deduplication.”
[22] M. K. Abdmouleh, A. Masmoudi, and M. S. Bouhlel, “A New Method Which Combines Arithmetic Coding with RLE for Lossless Image Compression,” J. Softw. Eng. Appl., vol. 5, no. 01, p. 41, 2012.
[23] M. Sharma, “Compression Using Huffman Coding,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 10, no. 5, (May, pp. 133–141, 2010.
[24] M. Kaur, “A Review of Various Data Compression Techniques to form a New Technique for Text Data Compression,” no. 5, pp. 1–5, 2015.
[25] M. R. Desai, “Efficient Virtual Machine Migration in Cloud Computing,” no. Vm, pp. 1015–1019, 2015.
[26] K. He, C. Huang, H. Zhou, J. Shi, X. Wang, and F. Dan, “Public auditing for encrypted data with client-side deduplication in cloud storage,” Wuhan Univ. J. Nat. Sci., vol. 20, no. 4, pp. 291–298, Jul. 2015.
[27] R. Miguel, “HEDup: Secure Deduplication with Homomorphic Encryption,” in 2015 IEEE International Conference on Networking, Architecture and Storage (NAS), 2015, pp. 215–223.
[28] X. Li, J. Li, and F. Huang, “A secure cloud storage system supporting privacy-preserving fuzzy deduplication,” Soft Comput., Jan. 2015.
[29] R. Chen, Y. Mu, G. Yang, and F. Guo, “BL-MLE: Block-Level Message-Locked Encryption for Secure

Large File Deduplication," IEEE Trans. Inf. Forensics Secur., vol. 10, no. 12, pp. 2643–2652, Dec. 2015.

[30] W. Leesakul, P. Townend, and J. Xu, "Dynamic Data Deduplication in Cloud Storage," in 2014 IEEE 8th International Symposium on Service Oriented System Engineering, 2014, pp. 320–325.

[31] F. Rashid, A. Miri, and I. Woungang, "Secure Enterprise Data Deduplication in the Cloud," in 2013 IEEE Sixth International Conference on Cloud Computing, 2013, pp. 367–374.

[32] X. Zhang and J. Zhang, "Data Deduplication Cluster Based on Similarity-Locality Approach," in 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, 2013, pp. 2168–2172.

[33] F. Rashid, A. Miri, and I. Woungang, "A secure data deduplication framework for cloud environments," in 2012 Tenth Annual International Conference on Privacy, Security and Trust, 2012, pp. 81–87.

[34] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage," in 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, 2013, vol. 1, pp. 363–370.

[35] C.-I. Fan, S.-Y. Huang, and W.-C. Hsu, "Hybrid data deduplication in cloud environment," in 2012 International Conference on Information Security and Intelligent Control, 2012, pp. 174–177.

[36] J. Li, Y. K. Li, X. Chen, P. P. C. Lee, and W. Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 5, pp. 1206–1216, May 2015.

[37] J. Li, X. Chen, M. Li, J. Li, P. P. C. Lee, and W. Lou, "Secure Deduplication with Efficient and Reliable Convergent Key Management," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 6, pp. 1615–1625, Jun. 2014.

[38] W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC "12, 2012, p. 441.

[39] X. Zhang, Z. Huo, J. Ma, and D. Meng, "Exploiting Data Deduplication to Accelerate Live Virtual Machine Migration," 2010.

[40] G. Zhu, X. Zhang, L. Wang, Y. Zhu, and X. Dong, "An Intelligent Data De-duplication Based Backup System," in 2012 15th International Conference on Network-Based Information Systems, 2012, pp. 771–776.