ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Detection and Elimination Scheme for Data Reduction with Low Overheads in Multi Cloud Storage

Sayali H. Ahire, Y. B. Hembade

Department of Computer Engineering, ZCOER, Narhe, Pune, India

Department of Computer Engineering, ZCOER, Narhe Pune, India

ABSTRACT: ecently data reduction or trimming has become more and more important in cloud storage systems due to the explosive growth of digital data in the world that has ushered in the big data era. One of the main challenges facing large-scale data reduction or trimming is how to maximally detect and eliminate redundancy at very low overheads. We design DARE, a low-overhead deduplication-aware resemblance detection and elimination scheme that effectively exploits existing duplicate-adjacency information for highly efficient resemblance detection in data deduplication based backup/archiving storage systems. The main idea behind DARE is Duplicate-Adjacency based Resemblance Detection (DupAdj), by considering any two data chunks to be similar (i.e., candidates for delta compression) if their respective adjacent data chunks are duplicate in a deduplication system, and then further enhance the resemblance detection efficiency by an improved super-feature approach. In existing system Deduplication technique is used only in-house computer, in our proposed system you can use DARE Deduplication technique in cloud storage also and you can perform DARE Deduplication technique on encrypted data. Our experimental results based on real-world and synthetic backup datasets show that DARE only consumes about 1/4 and 1/2 respectively of the computation and indexing overheads required by the traditional super-feature approaches while detecting 2-10 percent more redundancy and achieving a higher throughput.

KEYWORDs: Data deduplication, Delta compression, Storage system, Index structure, Performance evaluation, Multi cloud, File fragmentation.

I. INTRODUCTION

THE amount of digital data is growing explosively, as evidenced in part by an estimated amount of about 1.2 zettabytes and 1.8 zettabytes respectively of data produced in 2010 and 2011. As a result of this “data deluge”, managing storage and reducing its costs have become one of the most challenging and important tasks in mass storage systems. According to a recent IDC study [3], almost 80 percent of corporations surveyed indicated that they were exploring data deduplication technologies in their storage systems to increase storage efficiency. Data deduplication is an efficient data reduction approach that not only reduces storage space by eliminating duplicate data but also minimizes the transmission of redundant data in low-bandwidth network environments In general, a chunk-level data deduplication scheme splits data blocks of a data stream (e.g., backup files, databases, and virtual machine images) into multiple data chunks that are each uniquely identified and duplicate-detected by a secure SHA-1 or MD5 hash signature (also called a fingerprint). Storage systems then remove duplicates of data chunks and store only one copy of them to achieve the goal of space savings. While data deduplication has been widely deployed in storage systems for space savings, the fingerprint-based deduplication approaches have an inherent drawback: they often fail to detect the similar chunks that are largely identical except for a few modified bytes, because their secure hash digest will be totally different even only one byte of a data chunk was changed. It becomes a big challenge when applying data deduplication to storage datasets and workloads that have frequently modified data, which demands an effective and efficient way to

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

eliminate redundancy among frequently modified and thus similar data. Delta compression, an efficient approach to removing
redundancy among similar data chunks has gained increasing attention in storage systems. For example, if chunk A2 is similar to chunk A1 (the basechunk), the delta compression approach calculates and then only stores the differences (delta) and mapping relation between A2 and A1. Thus, it is considered a promising technique that effectively complements the fingerprint-based deduplication approaches by detecting similar data missed by the latter. One of the main challenges facing the application of delta compression in deduplication systems is how to accurately detect the most similar candidates for delta compression with low overheads. The state-of-the-art solutions detect similarity for delta compression by computing several Rabin fingerprints as features and grouping them into super-fingerprints, also referred to as super-features (SF). Nevertheless, to index a dataset of 80 TB and assuming an average chunk size of 8 KB and 16 bytes per index entry, for example, about 200 GB worth of super-feature index entries must be generated, which will still be too large to fit in memory [12]. Since the random accesses to on-disk index are much slower than that to RAM, the frequent accesses to on-disk super-features will cause the system throughput to become unacceptably low for the users.

II. BACKGROUND AND MOTIVATION

Data deduplication is becoming increasingly popular in data-intensive storage systems as one of the most efficient data reduction approaches in recent years. Fingerprint based deduplication techniques eliminate duplicate chunks by checking their secure-fingerprints (i.e., SHA-1/ SHA-256 signatures), which has been widely used in commercial backup and archiving storage systems. Another challenge for data deduplication is how to maximally detect and eliminate data redundancy in storage systems by determining appropriate data chunking schemes. In order to find more redundant data, the content-defined chunking (CDC) approach was proposed in LBFS to find the proper cut-point of each chunk in the files and address the boundary-shift problem. Re-chunking approaches were also proposed to divide those non-duplicate chunks into smaller ones to expose and detect more redundancy. Resemblance detection with delta compression as another approach to data reduction in storage systems, was proposed more than 10 years ago but was later overshadowed by fingerprint-based deduplication due to the former's scalability issue. Table 1 compares these two data reduction approaches. Resemblance detection detects redundancy among similar data at the byte level while duplicate detection finds totally identical data at the chunk level, which makes the latter much more scalable than the former in mass storage systems.

Fact of Duplicate Adjacency

The modified chunks may be very similar to their previous versions in a backup system while unmodified chunks will remain duplicate and are easily identified by the deduplication process. For those non-duplicate chunks that are location-adjacent to known duplicate data chunks in a deduplication system, it is intuitive and quite possible that only a few bytes of them are modified from the last backup, making them potentially excellent delta compression candidates. Rethinking of the Super-Feature Approaches Similar data, like duplicate data, are in wide existence in backup systems. Meister and Brinkmann find that small semantic changes on documents may result in big modifications in the binary representation of files, and delta compression is more effective in eliminating redundancy in such cases. To support delta compression, resemblance detection will be required for selecting suitable similar candidates.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

III. LITERATURE SURVEY

Sr. No.	Author & Title	Proposed Scheme	We have referred
1	B. Zhu, K. Li, and R. H. Patterson, "Avoiding the disk bottleneck in the data domain deduplication file system," in Proc. 6th USENIX Conf. File Storage Technol., Feb. 2008, vol. 8, pp. 1–14.	Describes three techniques employed in the production Data Domain deduplication file system to relieve the disk bottleneck. These techniques include: (1) the Summary Vector, a compact in-memory data structure for identifying new segments; (2) Stream-Informed Segment Layout, a data layout method to improve on-disk locality for sequentially accessed segments; and (3) Locality Preserved Caching, which maintains the locality of the fingerprints of duplicate segments to achieve high cache hit ratios. Together, they can remove 99% of the disk accesses for deduplication of real world workloads. These techniques enable a modern two-socket dual-core system to run at 90% CPU utilization with only one shelf of 15 disks and achieve 100 MB/sec for single-stream throughput and 210 MB/sec for multi-stream throughput.	(1) A compact in-memory data structure for identifying new segments; (2) Stream-Informed Segment Layout, a data layout method to improve on-disk locality for sequentially accessed segments; and (3) Locality Preserved Caching, which maintains the locality of the fingerprints of duplicate segments to achieve high cache hit ratios.
2	D. Meister, J. Kaiser, and A. Brinkmann, "Block locality caching for data deduplication," in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.	Propose a novel approach, called Block Locality Cache (BLC), that captures the previous backup run significantly better than existing approaches and also always uses up-to date locality information and which is, therefore, less prone to aging. We evaluate the approach using a trace-based simulation of multiple real-world backup datasets. The simulation compares the Block Locality Cache with the approach of Zhu et al. [37] and provides a detailed analysis of the behavior and IO pattern. Furthermore, a prototype implementation is used to validate the simulation.	Simplified example of chunk index entries and block recipes and Simplified example of the cache data structures of the BLC approach.
3	A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, "Primary data deduplication-large scale study and system design," in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp. 285–296.	We present a large scale study of primary data deduplication and use the findings to drive the design of a new primary data deduplication system implemented in the Windows Server 2012 operating system. File data was analyzed from 15 globally distributed file servers hosting data for over 2000 users in a large multinational corporation. The findings are used to arrive at a chunking and compression approach which maximizes	Architecture of a new primary data deduplication system and evaluate the deduplication performance and chunking aspects of the system.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

		deduplication savings while minimizing the generated metadata and producing a uniform chunk size distribution. Scaling of deduplication processing with data size is achieved using a RAM frugal chunk hash index and data partitioning – so that memory, CPU, and disk seek resources remain available to fulfill the primary workload of serving IO.	
4	S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," in Proc. USENIX Conf. File Storage Technol., Jan. 2002, pp. 89–101.	describes a network storage system, called Venti, intended for archival data. In this system, a unique hash of a block's contents acts as the block identifier for read and write operations. This approach enforces a write-once policy, preventing accidental or malicious destruction of data. In addition, duplicate copies of a block can be coalesced, reducing the consumption of storage and simplifying the implementation of clients. Venti is a building block for constructing a variety of storage applications such as logical backup, physical backup, and snapshot file systems. Built a prototype of the system and present some preliminary performance results. The system uses magnetic disks as the storage technology, resulting in an access time for archival data that is comparable to non-archival data. The feasibility of the write-once model for storage is demonstrated using data from over a decade's use of two Plan 9 file systems.	Choice of Storage Technology , A tree structure for storing a linear sequence of blocks and Build a new version of the tree.
5	A. Venish and K. Siva Sankar, "Study of Chunking Algorithm in Data Deduplication " Proceeding of the International Conference on Soft Computing System, ICSCS , Volume 2.	In the deduplication technology, data are broken down into multiple pieces called "chunks" and every chunk is identified with a unique hash identifier. These identifiers are used to compare the chunks with previously stored chunks and verified for duplication. Since the chunking algorithm is the first step involved in getting efficient data deduplication ratio and throughput, it is very important in the deduplication scenario. In this paper, we discuss different chunking models and algorithms with a comparison of their performances.	Chunking algorithm Data deduplication Fixed and variable chunking method
6	Bo Mao, Member, IEEE, Hong Jiang, Fellow, IEEE, Suzhen Wu, Member, IEEE,	propose a performance-oriented I/O deduplication, called POD, rather than a	Select the every requests to deduplicate or do not bypass

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

	and Lei Tian, Senior Member, IEEE "**Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud**" IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016	capacity-oriented I/O deduplication, exemplified by iDedup, to improve the I/O performance of primary storage systems in the Cloud without sacrificing capacity savings of the latter. POD takes a two-pronged approach to improving the performance of primary storage systems and minimizing performance overhead of deduplication, namely, a request-based selective deduplication technique, called Select-Dedupe, to alleviate the data fragmentation and an adaptive memory management scheme, called iCache, to ease the memory contention between the bursty read traffic and the bursty write traffic. We have implemented a prototype of POD as a module in the Linux operating system.	small requests (e.g., 4 KB, 8 KB or less).
7	Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "**A Hybrid Cloud Approach for Secure Authorized De-duplication**" IEEE Transactions on Parallel and Distributed Systems: PP Year 2014	In the proposed system we are achieving the data de-duplication by providing the proof of data by the data owner. This proof is used at the time of uploading of the file. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files.	New de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing.
8	Shweta D. Pochhi, Prof. Pradnya V. Kasture "**Encrypted Data Storage with De-duplication Approach on Twin Cloud** " International Journal of Innovative Research in Computer and Communication Engineering	The data and the Private cloud where the token generation will be performed for each file. Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file. Private clouds then generate a hash and a token and send the token to client. Token and hash keep in the private cloud itself so that whenever next file comes for token generation, the private clod can refer the same token. Once Client gets token for a particular file, public cloud search for the similar token if it exists or not. If the token exist public cloud will return a pointer of the already existing file otherwise it will send a message to upload a file.	A system which achieves confidentiality and enables block-level de-duplication at the same time.Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file. Private cloud then generate a hash and a token and send the token to client. Token and hash keep in the private cloud itself so that whenever next file comes for token generation, the private clod can refer the same token.
9	BhushanChoudhary, AmitDravid "**A Study On Secure Deduplication Techniques In Cloud Computing**" International Journal of	A de-duplication system in the cloud storage proposed to reduce the storage size of the tags for integrity check. To	A de-duplication system in the cloud storage proposed to reduce the storage size of the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

	Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3, Issue 12, April 2014	upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.	tags for integrity check. To upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.
10	Wee Keong Ng SCE, NTU Yonggang Wen SCE, NTU Huafei Zhu "**Private Data De-duplication Protocols in Cloud Storage**" SAC12 March 2529, 2012, Riva del Garda, Italy. Copyright 2011 ACM 9781450308571/12/03	This paper studies private data de-duplication technique for cloud storages. Intuitively, a private data de-duplication protocol allows a client who holds a private data proves to a server who holds a summary string of the data that he/she is the owner of that data without revealing further information to the server.	The proposed private data de-duplication protocol is provably secure in the simulation based framework assuming that the underlying hash function is collision resilient, the discrete logarithm is hard and the erasure coding algorithm E can erasure up to fraction of the bits in the presence of malicious adversaries.

IV. S/W REQUIREMENT SPECIFICATION

1. Operating System - Windows95/98/2000/XP
2. Application Server - Tomcat5.0/6.X
3. Front End - HTML, JDK 1.7, JSP
4. Scripts - JavaScript.
5. Server side Script - Java Server Pages.
6. Database - My SQL 5.0
7. IDE - Eclipse Luna

V. PROPOSED SYSTEM

Propose DARE, a low-overhead Deduplication- Aware Resemblance detection and Elimination scheme for deduplication based backup and archiving storage system. The main idea of DARE is to effectively exploit existing duplicate-adjacency information to detect similar data chunks (DupAdj), refine and supplement the detection by using an improved super-feature approach (Low-Overhead Super-Feature) when the existing duplicate-adjacency information is lacking or limited. In addition, we present an analytical study of the existing super-feature approach with a mathematic model and conduct an empirical evaluation of this approach with several real-world workloads in data deduplication systems. In existing system DARE Deduplication technique is used only in-house computer, in our proposed system you can use DARE Deduplication technique in cloud storage also and you can perform DARE Deduplication technique on encrypted data. Our experimental evaluation results, based on real-world and synthetic backup datasets, show that DARE significantly outperforms the traditional Super-Feature approach. More specifically, the DupAdj approach achieves a similar data reduction efficiency to the pure super-feature approach and DARE detects 2-10 percent more redundant data while achieving a higher throughput of data reduction than the pure super-feature approach.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

V. SYSTEM ARCHITECTURE

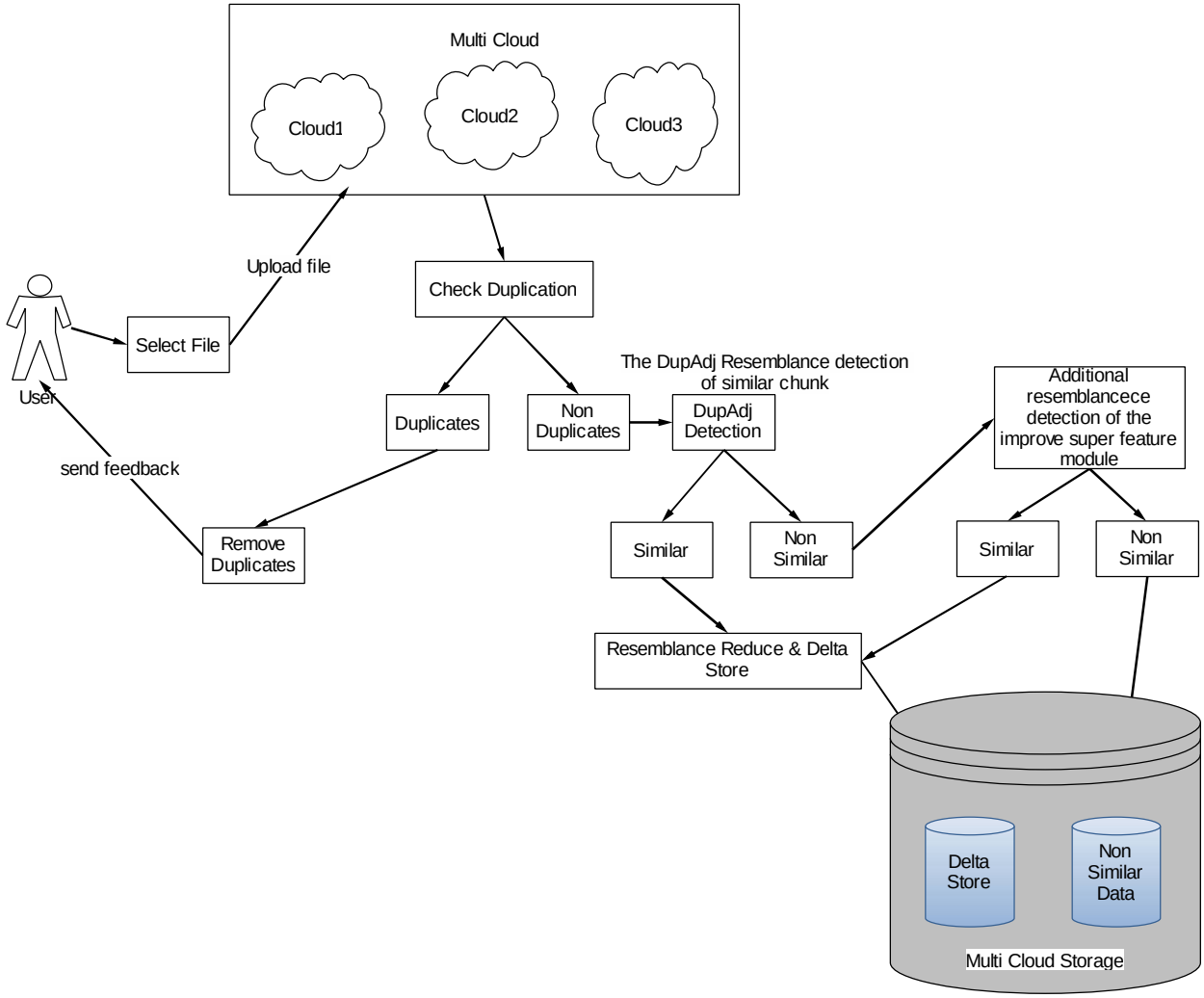


Fig.1. Proposed System Architecture

VI. MATHEMATICAL MODEL AND DESIGN

Relevant mathematics associated with the Project

Input

Input given to the system is: -Encrypted File in any format.

Output

Whenever user wants to upload the file on cloud or store the file on secondary storage then we check or test the duplication and Resemblance Detection and Elimination or not.

Process

Step 1: Data owner Select File

Step 2: Encrypt File (For encryption we user AES or RSA).

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Step 3: Upload file on cloud or store the file on secondary storage.
Step 4: CSP or Controller check the duplicate file available on cloud or secondary storage.
Step 5: If found then remove the duplication and maintain index.
Step 6: On non duplicate data CSP again check resemblance detection of similar chunk.
Step 7: If resemblance found then reduce it create delta store.
Step 8: Again on non similar data check resemblance detection.
Step 9: If resemblance found then reduce it store similar data in delta store.
Step 10: Finally non similar data stored into secondary storage or Cloud storage

VII. ALGORITHMS USED

a. File Encryption

$$\mathrm{Encrypt}(pk, i)$$

This algorithm is run by the data owner to encrypt the i^{th} document and generate its keywords' ciphertexts. For each document, this algorithm will create a delta Δi for its searchable encryption keyk_i. On input of the owner's public key pk and the file index i, this algorithm outputs data ciphertext and keyword ciphertextsCi.

b. Encrypted Data Upload

If data duplication check is negative, the data holder encrypts its data using a randomly selected symmetric key DEK in order to ensure the security and privacy of data, and stores the encrypted data at CSP together with the token used for data duplication check. The data holder encrypts DEK with pk AP and passes the encrypted key to CSP.

c. Data Deduplication

Data duplication occurs at the time when data holder u tries to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive, CSP contacts AP for deduplication by providing the token and the data holder's PRE public key. The AP challenges data ownership, checks the eligibility of the data holder, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.

VIII. RESULT TABLE

Percentage of found similer block in Existing System & Proposed System

File\System	Proposed System	Existing System
File 1	90	65
File 2	80	60
File 3	90	60
File 4	50	20

Proposed system work on low overheads means existing system only find those file which have 100% duplicate if file is 50% duplicate then existing system directly allocate the storage space for file in secondary storage, But in proposed work system reduce the 50% data of file using block and byte level duplication checking and maintain the index and store only 50% data on cloud storage which is not duplicate. Above result table and graph shows the Performance Mesurment and Comparitive analysis between Proposed and Existing System.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

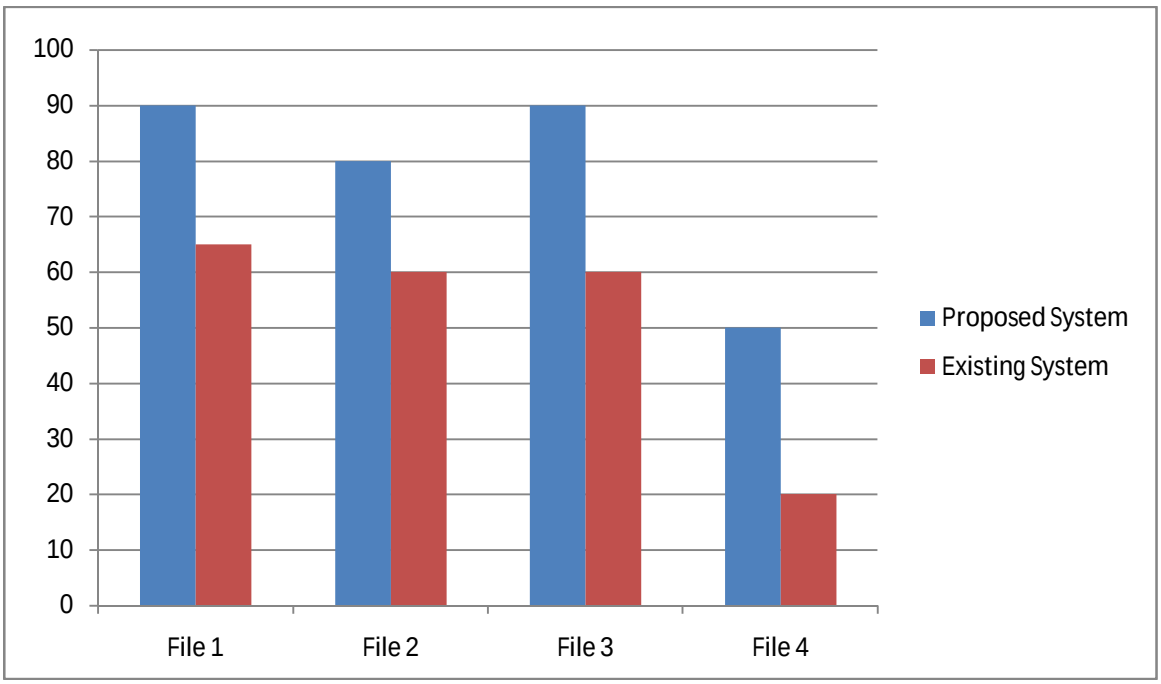
International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

IX. PERFORMANCE MEASURE OR EFFICIENCY EVALUATION



X. CONCLUSION

In this paper, we present DARE, a deduplication-aware, low-overhead resemblance detection and elimination scheme for data reduction in backup/archiving storage systems. DARE uses a novel approach, DupAdj, which exploits the duplicate-adjacency information for efficient resemblance detection in existing deduplication systems, and employs an improved super-feature approach to further detecting resemblance when the duplicate-adjacency information is lacking or limited.

Results from experiments driven by real-world and synthetic backup datasets suggest that DARE can be a powerful and efficient tool for maximizing data reduction by further detecting resembling data with low overheads. Specifically, DARE only consumes about 1/4 and 1/2 respectively of the computation and indexing overheads required by the traditional super-feature approaches while detecting 2-10 percent more redundancy and achieving a higher throughput.

Furthermore, the DARE-enhanced data reduction approach is shown to be capable of improving the data-restore performance, speeding up the deduplication-only approach by a factor of 2(2X) by employing delta compression to further eliminate redundancy and effectively enlarge the logical space of the restoration cache. Our preliminary results on the data-restore performance

suggest that supplementing delta compression to deduplication can effectively enlarge the logical space of the restoration cache.

REFERENCES

1. B. Zhu, K. Li, and R. H. Patterson, “Avoiding the disk bottleneck in the data domain deduplication file system,” in Proc. 6th USENIX Conf. File Storage Technol., Feb. 2008, vol. 8, pp. 1–14.
2. D. Meister, J. Kaiser, and A. Brinkmann, “Block locality caching for data deduplication,” in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.
3. A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, “Primary data deduplication-large scale study and system design,” in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp. 285–296.
4. S. Quinlan and S. Dorward, “Venti: A new approach to archival storage,” in Proc. USENIX Conf. File Storage Technol., Jan. 2002, pp. 89–101.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

5. A. Venish and K. Siva Sankar, "Study of Chunking Algorithm in Data Deduplication " Proceeding of the International Conference on Soft Computing System, ICSCS , Volume 2.
6. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized De-duplication" IEEE Transactions on Parallel and Distributed Systems: PP Year 2014
7. Bo Mao, Member, IEEE, Hong Jiang, Fellow, IEEE, Suzhen Wu, Member, IEEE, andLeiTian, Senior Member, IEEE "Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud " IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016
8. Mr Vinod B Jadhav Prof Vinod S Wadne Secured Authorized De-duplication Based Hybrid Cloud Approach International Journal of Advanced Research in Computer Science and Software Engineering
9. JadapalliNandini, Rami reddyNavateja Reddy Implementation De-duplication System with Authorized Users International Research Journal of Engineering and Technology (IRJET)
10. Backialakshmi. N Manikandan. M SECURED AUTHORIZED DE-DUPLICATION IN DISTRIBUTED SYSTEM IJIRST International Journal for Innovative Research in Science and Technology— Volume 1 — Issue 9 — February 2015