LATENCY AWARE FOG CENTRIC DE-DUPLICATION SYSTEM TO REDUCE DATA REDUNDANCY IN IOT APPLICATIONS

[1]Dr. A.GODAVARI,[2]MOTHUKURI PRANAV CHANDRA, [3]G.SHIVANI ,[4]D.SRENIKA [5]SYED ASGAR ALI NAVEED

1Assistant Professor, Dept of CSE(Networks), Kakatiya Institute of Technology and Science Warangal-506015, Telangana, India, Email: ag.csn@kitsw.ac.in

2Student, Dept of CSE(Networks), Kakatiya Institute of Technology and Science, Warangal-506015, Telangana, India . Email: b20in037@kitsw.ac.in

3Student, Dept of CSE(Networks), Kakatiya Institute of Technology and Science, Warangal-506015, Telangana, India. Email: b20in049@kitsw.ac.in

4Student, Dept of CSE(Networks), Kakatiya Institute of Technology and Science, Warangal-506015, Telangana, India. Email: b20in054@kitsw.ac.in

5Student, Dept of CSE(Networks), Kakatiya Institute of Technology and Science, Warangal-506015, Telangana, India . Email: b21in066L@kitsw.ac.in

Abstract-Now-a-days IoT applications are generating exponential amount of data. According to IDC, 75% of the data is duplicate data. This leads to storage inefficiency. So, there is a need of technique to reduce redundant data. Deduplication is one the technique which helps to reduce data redundancy and improves data management. In cloud, the data has to be sent for redundancy checking results in high communication overhead. Instead of sending a entire data to the cloud compute a convergent key base and sent to the cloud for redundancy checking. IoT-based medical devices track patients health conditions and periodically upload them to the cloud servers now and then. The increasing popularity and its real-time applications have made IoT healthcare monitoring as a promising technology. However, the redundant healthcare data produced by the IoT medical devices are exponentially growing and reduce the storage efficiency. Executing a deduplication algorithm, identifying the redundant data, and stopping them from entering into the cloud servers helps in improving the storage efficiency. However, in existing approaches, the communication overhead is heavy. Traditional hash-based and convergent key-based deduplication approaches use the entire hash values of the data chunks to perform deduplication as well as encryption.

KEYWORDS:IoT, data, redundancy, deduplication, cloud, communication, overhead, healthcare, storage, efficiency, hash-based, convergent key-based, encryptio

1. INTRODUCTION

Data deduplication has been demonstrated to be an effective technique in Cloud backup and archiving applications to reduce the backup window, improve the storage-space efficiency and network bandwidth utilization. Recent studies reveal that moderate to high data redundancy clearly exists in VM (Virtual Machine) , enterprise ,and High-Performance Computing (HPC) storage systems. These studies have shown that by applying the data deduplication technology to large-scale data sets, an average space saving of 30%, with up to 90% in VM and 70% in HPC storage systems, can be achieved. For example, the time for the live VM migration in the Cloud can be significantly reduced by adopting the data deduplication technology .The existing data deduplication schemes for primary storage, such as iDedup and Offline-Dedupe , are capacity oriented in that they focus on storage capacity savings and only select the large requests to deduplicate and bypass all small requests (*e.g.*, 4KB, 8KB or less). The rationale is that the small I/O requests only account for a tiny fraction of the storage capacity requirement, making deduplication on them unprofitable and potentially counterproductive considering the substantial deduplication overhead involved. However, previous workload studies have revealed that small files dominate in primary storage systems (more than 50%) and are at the root of the system performance bottleneck.

II. LITERATURE SURVEY

Previous works in the realm of data management and storage have laid foundational groundwork for addressing challenges such as redundancy and storage efficiency. One notable attempt was by B. Mao and S. Wu from Xiamen University, China, along with H. Jiang and L. Tian from the University of Nebraska Lincoln, USA. Their work focused on exploring techniques to mitigate redundancy in data generated by IoT applications, acknowledging the exponential growth in data volume and the ensuing storage inefficiency. By introducing deduplication as a solution, they aimed to improve data management and alleviate the burden on storage systems.

Furthermore, N. Agrawal, William J. Bolosky, John R. Douceur, and Jacob R. Lorch proposed a comprehensive study on

file-system metadata in their work presented at FAST'07 in February 2007. This study provided insights into the intricacies of file-system management, shedding light on potential areas for optimization and efficiency enhancement.

Another noteworthy contribution came from Anand, S. Sen, A. Krioukov, F. Popovici, A. Akella, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, and S. Banerjee, presented at OSDI'08 in December 2008. Their research focused on streamlining file system operations to minimize micromanagement efforts, particularly through the utilization of range writes. This approach aimed to improve system performance and alleviate the overhead associated with file system management tasks.

$$Deduplication\ Ratio = \frac{Size\ \text{of the}\ input\ data - Storage\ utilized}{size\ \text{of the}\ input\ data}.$$

Building upon these previous works, it becomes evident that the need for efficient data management techniques, especially in the context of IoT-generated data and cloud storage, remains a pressing concern. The exponential growth of data volume exacerbates challenges related to redundancy, storage efficiency, and communication overhead. Therefore, there is a continual drive within the research community to innovate and develop solutions that can address these challenges effectively.

$$Throughput = \frac{Input\ data\ size}{De\text{-}duplication\ time}$$

In the pursuit of enhancing storage efficiency and reducing redundancy, deduplication emerges as a prominent technique. By identifying and eliminating duplicate data, deduplication not only optimizes storage utilization but also minimizes the overhead associated with data transmission and processing. However, traditional approaches to deduplication, such as hash-based and convergent key-based methods, may still incur significant communication overhead, particularly in cloud environments where data must be transmitted for redundancy checking.

Parameters	FileLevel	FixedSize	VariableSize
De-duplicationRatio	Less	Medium	High
ProcessingTime	Medium	Less	High
IndexOverhead	Better	Worst	Worst

To address these challenges, recent research

efforts have focused on refining deduplication algorithms and exploring novel approaches to data management. For instance, there is ongoing exploration into the integration of deduplication with encryption techniques to ensure data security without compromising storage efficiency. Additionally, advancements in distributed computing architectures and parallel processing methodologies hold promise for optimizing deduplication processes and reducing communication overhead.

Moreover, the integration of machine learning and artificial intelligence techniques into deduplication algorithms presents an exciting avenue for further improvement. By leveraging pattern recognition and predictive analytics, these advanced algorithms can enhance the accuracy and efficiency of duplicate data detection, thereby optimizing storage utilization and reducing communication overhead.

TABLE 2: False-positive errors in the distributed index table.

Number of data blocks	Number of the hash function used (k)	Total bits in index table	Prob. of false-positive error in the index table
1024	3	9801	0.003319
2048	3	19600	0.003319
3072	3	29241	0.002969
4096	3	39204	0.003319
5120	3	48841	0.003058
10240	3	97969	0.00319
102400	3	980100	0.00321
204800	3	1960000	0.00325
512000	3	4906225	0.00312
1024000	3	9809424	0.00317
2048000	3	19624900	0.00319

EXISTING SYSTEM:

Existing methodologies for addressing data redundancy in IoT applications primarily rely on deduplication techniques, which aim to identify and eliminate duplicate data to improve storage efficiency and data management. However, traditional approaches often incur high communication overhead, particularly in cloud environments where data redundancy checking necessitates transmitting entire datasets. Moreover, the exponential growth of redundant healthcare data from IoT medical devices further exacerbates storage inefficiency issues.

FileLevel	BlockLevel	ByteLevel
Whole file isconsidered as a singleinstance and hashvalueforwhole fileis generated.	More fine grain levelde-duplication,dividin geachfileintoblocks.	Compares data chunkbytebybyte& checksfor redundantfragments.
Indexisvery small.	Indexsizeisgreater than filelevel.	Indexsizeishuge.
Lesscomputation overhead.	Morecomputation overhead.	Morecomputation overhead.

DISADVANTAGES OF EXISTING SYSTEM:

Transmitting entire datasets for redundancy

checking leads to increased network traffic and latency.

Traditional approaches may struggle to scale effectively to handle the growing volume of IoT data.

Despite deduplication techniques, storage capacity is underutilized, especially with the exponential growth of redundant healthcare data.

PROPOSED SYSTEM:

ADVANTAGES OF PROPOSED SYSTEM:

Reduced Communication Overhead: By leveraging a fog-centric architecture, the proposed system minimizes communication overhead by processing deduplication tasks closer to data sources, reducing latency and network traffic.

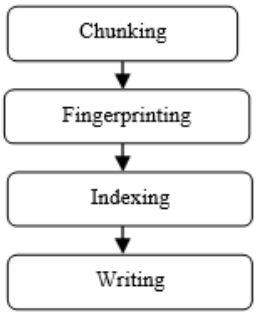


Figure 1: De-duplication process

Improved Scalability: With a focus on performance-oriented deduplication and adaptive memory management, the proposed system enhances scalability, allowing it to efficiently handle the increasing volume of

IoT data.

Enhanced Storage Efficiency: Through request-based selective deduplication and a performance-oriented approach, the system optimizes storage utilization, mitigating storage inefficiency issues prevalent in the existing system.

Input: Data chunks $DC_{i=1, 2...m}$ (each with the size of 1024KB)
Output: Partial hash values ($p\alpha$)
Begin:

1. Data owner HASHES the data chunks ($DC_{i=1, 2...m}$) using 'n' number of hash functions.
2. Data owners stores resultant message digests (hash values) of the data chunks in their local storage.
3. Generate a partial hash value
   Begin
   a. Divide the hash values into 'p' partitions (i.e.) (L bits of message digest / 'p' partitions).
   b. To derive the partial hash values '$p\alpha$', choose odd partitions from the message digest and take the even bits from each partition.
   c. Combines the even bits from each odd partition and create a partial hash value '$p\alpha$' for each data chunk.
   d. Creates chunk id (C_{id}), file id (F_{id}) for partial hash values.
   End
4. Transfer the chunk id (C_{id}), file id (F_{id}), and its corresponding partial hash values '$p\alpha$' to the nearby fog nodes.

Algorithm 1: Data owner—generating partial hash values.

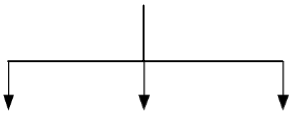
Lower Performance Overhead: By employing efficient deduplication algorithms and adaptive memory management, the proposed system minimizes performance overhead, ensuring smoother operation and improved system performance.

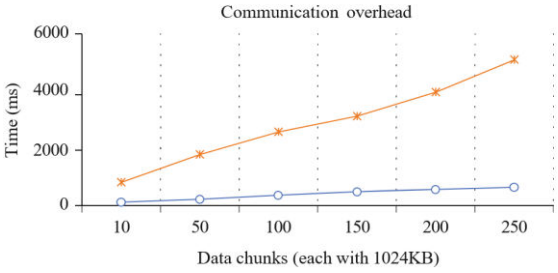
5

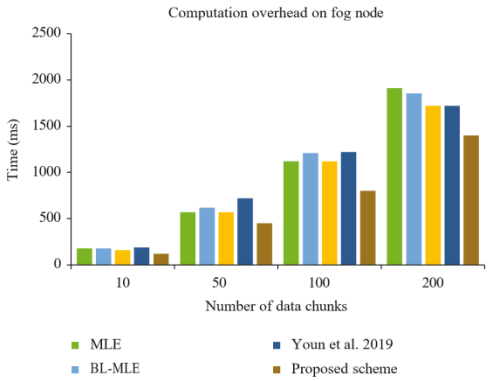
Real-Time Processing Capabilities: The integration of fog computing and prioritization of real-time data processing enables the proposed system to deliver timely insights and responses, crucial for applications such as healthcare monitoring.

4. RESULTS:

LevelsofDe-duplication







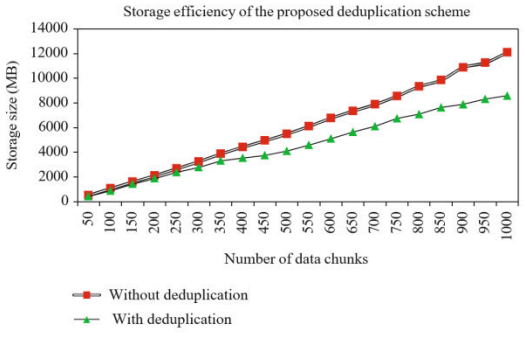


Figure 5: Storage efficiency of cloud storage servers.

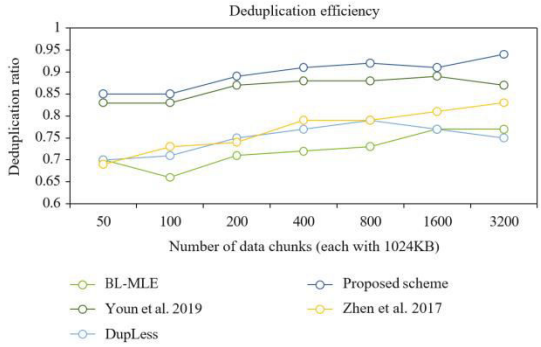


Figure 6: Comparison of deduplication ratio.

5. CONCLUSION

In this paper, we propose POD, a performance-oriented deduplication scheme, to improve the performance of primary storage systems in the Cloud by leveraging data deduplication on the I/O path to remove redundant write requests while also saving storage space. It takes a request-based selective deduplication approach (SelectDedupe) to deduplicating the I/O redundancy on the critical I/O path in such a way that it minimizes the data fragmentation problem. In the meanwhile, an intelligent cache management (iCache) is employed in

POD to further improve read performance and increase space saving, by adapting to I/O burstiness. Our extensive trace driven evaluations show that POD significantly improves the performance and saves capacity of primary storage systems in the Cloud. POD is an ongoing research project and we are currently exploring several directions for iCache into other deduplication schemes, such as IDedup, to investigate how much benefit iCache can bring to saving extra storage capacity and improving read performance. Second, we will build a power measurement module to evaluate the energy efficiency of POD. By reducing write traffic and saving storage space, POD has the potential to save the power that disks consume. We will compare the extra power that CPU consumes for computing fingerprints with the power that the storage saves, thus systematically investigating the energy efficiency of POD.

6.

REFERENCES

1. B.Alouffi,M.Hasnain,A.Alharbi,W.Alosaimi,H.Alyami, and M. Ayaz, "A systematic literature review on cloud com-putingsecurity:threatsandmitigation strategies,"*IEEEAccess*, vol. 9, pp. 57792–57807, 2021.
2. IDC Data Age, "Whitepaper," 2025, Available from, Error! Hyperlink reference not valid [Accessed on July 2022]
3. "NewspaperArticle," Available from,https://waterfordtechnologies.com/banish-your-redundant-data-like-st-patrick-banished-snakes-out-of-ireland/ , [Accessed on July 2022].
4. B.T.Devi,S.Shitharth,andM.A.Jabbar,"Anappraisalover intrusion detection systems in cloud computing security attacks,"in*20202ndInternationalConferenceonInnovative Mechanisms for Industry Applications (ICIMIA)*, pp. 722–727, Bangalore, India, 2020.
5. R.Chen,M.Yi,G.Yang,andF.Guo," BL-MLE:block-level message-lockedencryptionforsecurelargefile deduplication," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2643–2652, 2015
6. T.R.Gadekallu,Q.V.Pham,D.C.Nguyenetal.,"Blockchain foredgeofthings:applications,opportunities,andchallenges," *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 964–988, 2021.
7. B. Prabadevi, N. Deepa, Q.-V.

7

Pham et al., “Toward block- chain for edge-of-things: a new paradigm, opportunities, and future directions,”*IEEE Internet of Things Magazine*, vol. 4, no. 2, pp. 102–108, 2021

8. M. Liyanage, Q. V. Pham, K. Dev et al., “A survey on Zero touch network and Service Management (ZSM) for 5G and beyondnetworks,”*JournalofNetworkandComputerApplica- tions*, vol. 203, article 103362, 2022.

9. H. Qi, Y. Han, X. Di, and F. Sun, “Secure data deduplication scheme based on distributed random key in integrated net- works,”in*In20173rdIEEEInternationalConferenceonCom- puterandCommunications(ICCC)*,pp.1308–1312,Chengdu, China, 2017.

10. R.MiguelandK.M.M.Aung,“Hedup :securededuplication with homomorphic encryption,” in *2015 IEEE International ConferenceonNetworking,ArchitectureandStorage(NAS)*pp.215–223,Boston,MA,USA,2015.,

11. R.L.Rivest,L.Adleman,andM.L.Dertouzos,“Ondatabanks andprivacyhomomorphisms,”*Foundationsofsecurecompu- tation*, vol. 4, no. 11, pp. 169–180, 1978.