ACCEL DEDUP

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

Computer Science and Technology

SUBMITTED BY

ROBAN SINGH	RUPALLI DEVI	SIMRAN TIWARI
2104169	2104172	21041

August 2024

UNDER THE GUIDANCE OF

ER.SHAILJA

GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA

INDEX

Sno.	Topic	Page Number
1.	Introduction	**3**
2.	Rationale	**4**
3.	Objectives	**5**
4.	Literature Review	**6**
5.	Feasibility Study	**7**
6.	Methodology	**8**
7.	Facilities Required	**9**
8.	Expected Outcomes	**10**
9	References	**11**

1.INTRODUCTION

We propose the Acceldedup system, a data deduplication solution designed to eliminate redundant data copies, conserve storage space, and reduce costs. Efficient data storage management is crucial in today's cloud computing landscape. Traditional deduplication schemes, such as iDedup and Offline-Dedupe, primarily target large I/O requests for capacity savings, often neglecting smaller requests (e.g., 4KB or 8KB). This capacity-oriented approach can overlook performance considerations, especially in primary storage systems where both read and write operations are critical.

The Acceldedup system addresses performance gaps in traditional deduplication with a dual-focused strategy that enhances I/O performance while maintaining storage efficiency. Acceldedup integrates two core innovations: Select-Dedupe and iCache.

- Select-Dedupe: This request-based selective deduplication technique considers small I/O requests, deduplicating sequential write requests to minimize data fragmentation and optimize performance.
- iCache : This component dynamically adjusts the partitioning of cache space between the index cache and the read cache based on workload characteristics.
- Acceldedup ensures data consistency and integrity, prevents incorrect data overwriting, and manages bursty read and write traffic to ensure efficient operation under various workloads

2.RATIONALE

The AccelDedup addresses the need for solution solves the demand for increased performance in cloud-based main storage systems.

Imrpoving the performance of current systems.

1. Ignoring minor I/O Requests: Conventional systems ignore minor I/O requests, which affects overall performance and misses possibilities for improvement.
2. Read efficiency Focus: When write optimization is prioritized above read efficiency,
3. Memory Contention: Memory contention and decreased performance are caused by single index cache management.

AccelDedup overcomes the performance constraints by –

- Balancing Performance and Efficiency: To optimize all data processes, take into account both tiny and large I/O requests.
- Adaptive memory management -iCache dynamically adjusts cache space based on workload dmands , enhancing system performace.
- Minimizing Fraagmentation- Select -Dedupe reduces data fragmentation by deduplicating sequential write requests.

3.OBJECTIVES

1. To create an interface for storing data on the cloud.
2. To perform De-duplication at the cloud Platform.
3. To compare and analyze the existing de-duplication technique with proposed solution.

4.LITERATURE REVIEW

The evolution of data deduplication has significantly impacted storage systems, especially in cloud environments where efficient data management is crucial. This review explores key advancements and challenges in data deduplication, focusing on performance-oriented approaches, memory management, and optimization strategies.

Deduplication research initially focused on reducing data redundancy in storage systems. Rabin's fingerprinting technique (1981) was foundational, introducing a method for breaking data into variable-sized chunks, a principle still relevant in deduplication. By the mid-2000s, studies like Zhu et al. (2008) proposed the use of content-defined chunking (CDC), which adapts chunk sizes based on data patterns, improving deduplication efficiency. Simultaneously, hash-based approaches, such as SHA-1 and MD5, emerged for identifying duplicate data segments, although their computational cost and scalability became significant challenges.3. Adaptive Memory Management

The next decade saw efforts to address computational bottlenecks and scalability issues. Debnath et al. (2010) introduced techniques leveraging Bloom filters to optimize the detection of duplicates, reducing memory overhead. Around this time, Sparse Indexing by Lillibridge et al. (2011) gained attention for enabling scalable deduplication in large storage systems by indexing only a subset of fingerprints.

In parallel, hardware-based acceleration began to take shape. Guo et al. (2014) explored GPU-accelerated deduplication, which significantly reduced computation time for chunk hashing and comparison. These studies marked a shift toward leveraging specialized hardware for performance enhancement.

5.FEASIBILITY STUDY

The AccelDedup system is technically feasible with its modest hardware requirements (Pentium i3, 4GB RAM, 500GB disk) and use of established technologies like Node.js and SQL Server 2005. The system's modular design supports smooth integration and incremental development, and the proposed deduplication and caching techniques are compatible with current storage technologies.

Operationally, AccelDedup enhances I/O performance and reduces fragmentation, benefiting existing storage systems by improving efficiency and reducing latency. It is adaptable to various environments and requires manageable user training due to its reliance on familiar technologies. Risks are mitigated through phased implementation and targeted troubleshooting.

Economically, the AccelDedup system is cost-effective, with development and implementation costs offset by significant performance gains and potential storage savings. The return on investment is supported by efficiency improvements and scalability, making Acceldedup a viable option for enhancing primary storage systems.

6. METHODOLOGY

The development and implementation of the AccelDedup system involve a structured approach encompassing project planning, system design, development, testing, deployment, and evaluation. Initially, project planning defines the objectives, scope, and resource allocation, setting clear milestones and timelines. The system design phase includes requirements analysis, architectural design, and technology selection, ensuring a robust framework for the AccelDedup system.

The development phase focuses on backend implementation using Node.js, frontend development with Express.js, HTML, CSS, and JavaScript, and database integration with SQL Server 2005. Integration ensures seamless interaction between components and existing storage infrastructures. Testing involves unit, system, and user acceptance testing to ensure functionality and performance meet the defined requirements. Deployment includes installation, data migration, and user training, followed by post-deployment evaluation and ongoing maintenance to monitor performance, address issues, and incorporate feedback for continuous improvement.

By adhering to this comprehensive methodology, the project aims to deliver a high-performance deduplication system that enhances storage efficiency and user satisfaction in modern cloud environments.

7. FACILITIES REQUIRED FOR PROPOSED WORK

Hardware-

- System : Pentium i3 Processor
- Hard Disk : 500 GB..
- Monitor : 15’’ LED
- Input Devices : Keyboard, Mouse
- RAM : 4 GB.

Software-

- Operating system : Windows 10/11.
- Coding Language : Node.js.
- Frontend : Express.js , HTML, CSS, JavaScript.
- IDE Tool : VISUAL STUDIO.
- Database : SQL SERVER 2005.

8. EXPECTED OUTCOMES

The successful implementation of the Performance-Oriented Deduplication (POD) system is expected to yield significant improvements in system performance, resource utilization, and user satisfaction. By optimizing both read and write operations, POD will reduce latency, improve throughput, and efficiently handle small I/O requests, leading to better storage capacity utilization and minimized redundant data. The adaptive memory management of iCache will optimize cache utilization and reduce memory contention, enhancing overall performance. Additionally, POD will minimize I/O overhead and data fragmentation, improve RAID reconstrussction performance, and ensure efficient storage usage, scalability, and enhanced user experience. The system's comprehensive performance evaluation will validate its effectiveness and set new benchmarks for future data deduplication advancements.

9.REFERENCES

[1]Author(s), "De-Duplication Over Cloud Data to Enhance the Storage Systems," *JP Infotech*, [Online]. Available: https://jpinfotech.org/de-duplication-over-cloud-data-to-enhance-the-storage-systems/

[2] Author(s), " De-Duplication Over Cloud Data to Enhance the Storage Systems " International Arab Journal of Information Technology, vol.16, no. 5,Sep. 2019. [Online]. Available: https://iajit.org/portal/PDF/September%202019,%20No.%205/15822.pdf.

[3]"Data Deduplication," *Data Intell*, Feb. 2023. [Online]. Available: https://dataintell.io/2023/02/data-deduplication/.

[4] Author(s), " De-Duplication of Data in Cloud Storage," *International Journal of Advanced Networking and Applications* [Online]. Available: https://www.ijana.in/papers/84.pdf.