IJETIR
ISSN (Online) 2583-0554
Impact Factor: 5.731
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025

Efficient In-Line Data Deduplication for Non-Volatile Memory Storage Using I/O Causality

Mrs. P. Anusha[1], Mr. B.Venkata Abhinav[2], Ms. D. Sai Kalpana[3]. Ms. D. Srilatha[4]
Assistant Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4]
Guru Nanak Institute of Technology, Hyderabad

Abstract*: I/O causality-based in-line data deduplication (ICID) is a novel approach designed to enhance deduplication efficiency in non-volatile memory (NVM)-enabled storage systems. Traditional chunk-based offline deduplication methods suffer from significant performance overhead due to data chunking and indexing, making them unsuitable for NVM-based storage, which benefits from fine-grained, byte-addressable operations. Unlike conventional in-line deduplication schemes that rely on hash indexes, ICID leverages a B-tree structure to record memory copy operations, enabling causality-based deduplication. To optimize efficiency, ICID introduces two key techniques: grouping memory-copy records targeted to the same page within a B-tree node to enhance data locality and exploiting spatial locality to identify and remove outdated memory-copy records, thereby reducing memory consumption. Evaluations conducted on a system with Intel Optane DC Persistent Memory Modules demonstrate that ICID significantly improves deduplication performance, achieving up to 16× higher deduplication ratios and reducing deduplication time by an average of 47% compared to state-of-the-art methods, particularly benefiting key-value stores like LevelDB.*

Keywords: ICID, Non-Volatile Memory, Data Locality Optimization, Intel Optane DC Persistent Memory

I. INTRODUCTION

DATA deduplication techniques have long been studied for space saving in storage systems. They eliminate duplicate data at file or chunk levels by identifying the same data using cryptographic hash functions such as SHA-1 or MD5. Data deduplication can be performed in an offline or inline mode. The offline deduplication is conducted after the data has been written to the storage device, and thus usually causes write amplification. In contrast, the inline deduplication detects the duplicate data before writing it back into storage, and thus can reduce data redundancy and the wear of storage devices. More importantly, inline deduplication has the potential to explore the correlation among I/O operations at runtime, and may significantly improve the deduplication ratio. Non-volatile memory (NVM) technologies such as Intel 3D Xpoint offer much lower cost, higher density and energy efficiency than traditional DRAM technologies. They have become a promising complement to DRAM to bridge the performance gap between main memory and SSD/HDD storage devices However, NVMs often suffer from limited write endurance, a typical PCM cell can only sustain 107-108 writes. Inline deduplication is a promising approach for reducing the storage consumption and the wear of NVM devices. Unfortunately, existing inline deduplication schemes are not efficient for new NVM devices. Since previous data deduplication techniques are designed for block devices such as HDD and SSD, they suffer from high computation and storage overhead due to data chunking and indexing. For example, to chunk 1 TB data without any redundancy using 4 KB blocks, Content-Defined Chunking (CDC) algorithm has to compute rolling hashes for almost 1012 times during the chunking stage. Moreover, the SHA-1 algorithm introduces 5 GB fingerprints (20 bytes per fingerprint) to index those chunks. Since DRAM resource is expensive and limited, many recent proposals store most fingerprints on disk. Thus, it is expensive to search those fingerprints in storage. Moreover, existing inline deduplication schemes have to make a tradeoff between the deduplication ratio and the read throughput. Moreover, using big chunks can also mitigate the total size of fingerprints and the cost of indexing. In contrast, using small chunks can improve the deduplication ratio, but lowers the read throughput and increase the cost of indexing.

Currently, most deduplication systems tend to use big chunks to improve the read throughput and to mitigate the cost of indexing, at the expense of a lower deduplication ratio.

II. LITERATURE SURVEY

Xu et. al (2022) : Persistent Memory (PM) is increasingly supplementing or substituting DRAM as main memory. Prior work have focused on reusability and memory leaks of persistent memory but have not addressed a problem amplified by persistence, persistent memory fragmentation, which refers to the continuous worsening of fragmentation of persistent memory throughout its usage. This paper reveals the challenges and proposes the first systematic crash-consistent solution, Fence-Free Crash-consistent Concurrent Defragmentation (FFCCD). FFCCD resues persistent pointer format, root nodes and typed allocation provided by persistent memory programming model to enable concurrent defragmentation on PM. FFCCD introduces architecture suppor`t for concurrent defragmentation that enables a fence-free design and fast read barrier, reducing two major overheads of defragmenting persistent memory. The techniques is effective (28–73% fragmentation reduction) and fast (4.1% execution time overhead).

C. Ji et. al(2021): Mobile applications exhibit unique file access patterns, often involving random accesses of write-mostly files and read-only files. The high write stress of mobile applications significantly impacts on the lifespan of flash-based mobile storage. To reduce write stress and save space without sacrificing user-perceived latency, this study introduces FPC, file access pattern guided compression. FPC is optimized for the random-writes and fragmented-reads of mobile applications. It features dual-mode compression: Foreground compression handles write-mostly files for write stress reduction, while background compression packs random-reading file blocks for boosted read performance. FPC exploits the out-of-place updating design in F2FS, a log-structured file system for mobile devices, for the best effect of the proposed dual-mode compression. Experimental results showed that FPC reduced the volume of total write traffic and executable file size by 26.1% and 23.7% on average, respectively, and improved the application launching time by up to 14.8%.

D. Yang el. Al (2021): Emerging non-volatile memory (NVM) technologies promise high density, low cost and dynamic random access memory (DRAM)-like performance, at the expense of limited write endurance and high write energy consumption. It is more practical to use NVM combining with the traditional DRAM. However, the hybrid memory management such as page migration becomes more challenging in a virtualization environment because virtual machines (VMs) are unaware of the memory heterogeneity. In this paper, we propose HMvisor, a hypervisor and VM coordinated hybrid memory management mechanism to better utilize DRAM and NVM resources. HMvisor exposes the memory heterogeneity to VMs by mapping virtual NUMA nodes to different physical NUMA nodes. We propose a lightweight and efficient page migration mechanism by decoupling page hotness tracking from page migration. HMvisor performs those operations in the hypervisor and VMs separately, without disrupting the execution of VMs. We also propose a memory resource trading policy to adjust the capacity of DRAM and NVM for each VM, with the monetary cost unchanged. We implement our prototype system based on QEMU/KVM and evaluate it with several benchmarks. Experimental results show that HMvisor can reduce 50% of write traffic to NVM with less than 5% performance overhead. Moreover, the hybrid memory adjustment scheme in HMvisor can significantly improve application performance by up to 30 ×

III. METHODOLOGIES

MODULES NAMES:

1. User Interface Design
2. Admin
3. Hospital
4. Health Sector
5. Encryption

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Impact Factor: 5.731
Volume 5, Issue 6, June 2025

MODULES EXPLANATION

- **USER INTERFACE DESIGN:** Users must register with details or log in to access the server, manage data rates, and perform query-based searches.
- **ADMIN**: Admin module controls all project roles, managing operations to ensure efficiency, supervision, and organizational success.
- **HOSPITAL**: Hospital module allows login, views patients, approves requests, sends symptoms, receives scan data, and delivers encrypted reports securely.
- **HEALTH SECTOR:** Health Sector scans patients based on symptoms and sends reports to Hospital Management for further processing.
- **ENCRYPTION:** Medical image encryption using ECC and Blum-Goldwasser ensures data confidentiality, integrity, and efficiency in resource-limited healthcare environments..

PROPOSED SYSTEM

Data deduplication saves storage by removing duplicate data using hash functions like SHA-1 or MD5. This project proposes I/O Causality-based In-line Deduplication (ICID) for NVM systems, using a B-tree to track memory-copy operations. Unlike hash-based methods, ICID identifies duplicates before writing, reducing redundancy and wear on storage devices while improving efficiency and performance.

IV. SYSTEM ARCHITECTURE

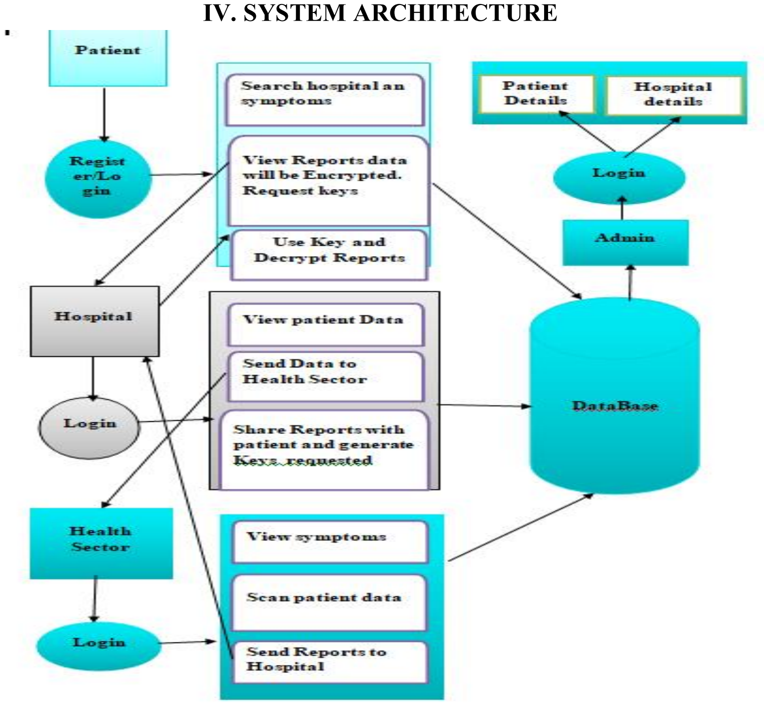

Fig 1: System Architecture

This diagram explains how a digital healthcare system connects patients, hospitals, and the health sector to manage and share medical data securely. Patients start by registering or logging in, then they can search for hospitals and view their health reports. These reports are encrypted for security, and patients need a key to decrypt and read them. This ensures that only the patient can access their private medical information.

IJETIR
ISSN (Online) 2583-0554
Impact Factor: 5.731
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025

V. EXPERIMENTAL RESULTS

Fig 2: Welcome Page

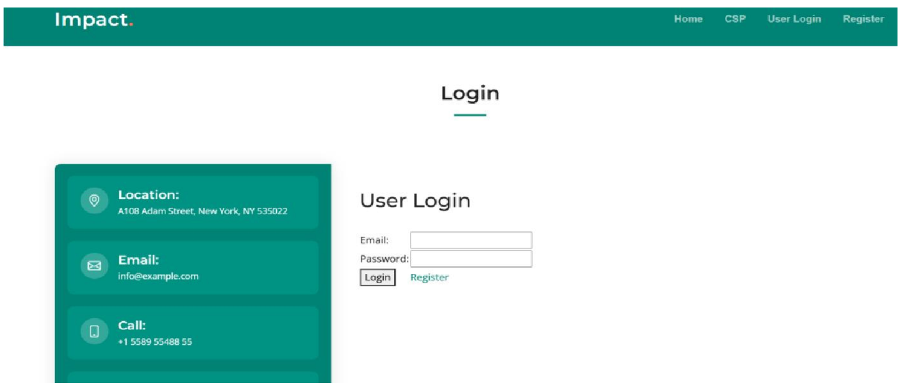

Fig 3:User Login Page

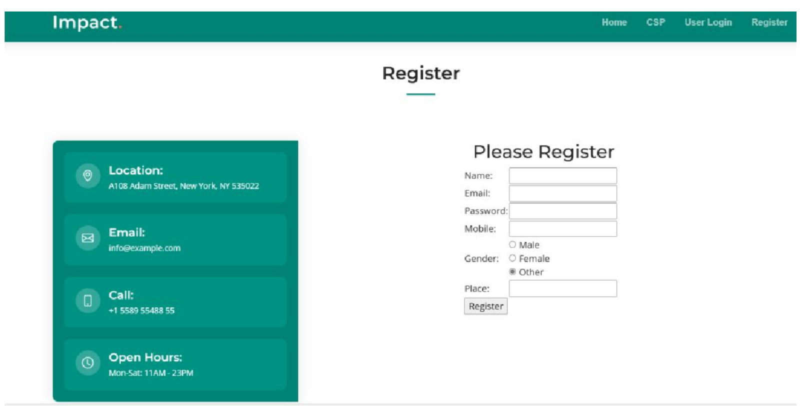

Fig 4: User Registration Page

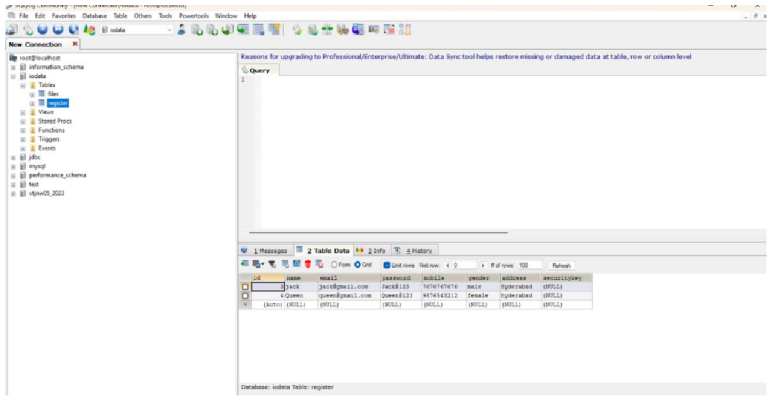

Fig 5: Data Base Page

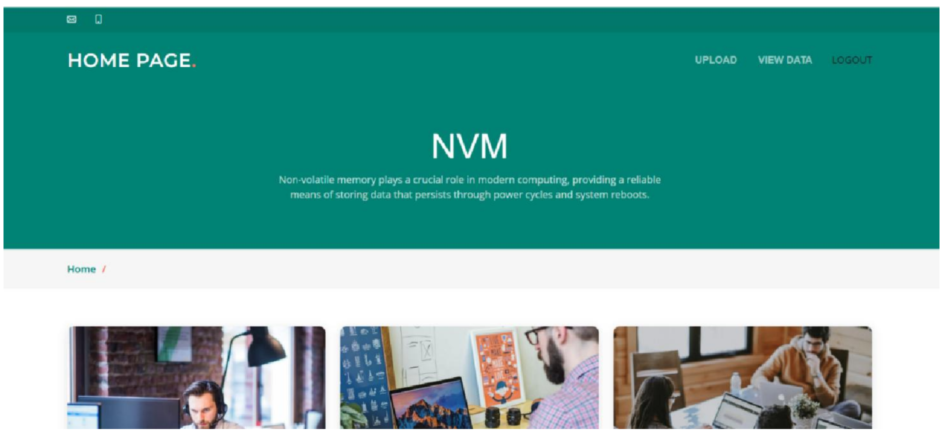

Fig 6:Home Page

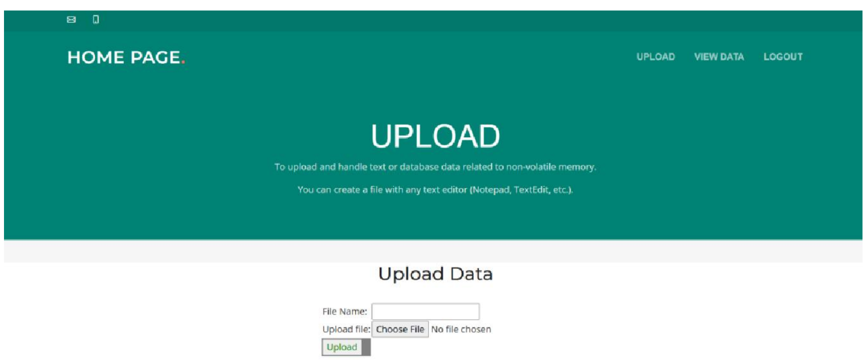

Fig 7: Upload Page

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Impact Factor: 5.731
Volume 5, Issue 6, June 2025

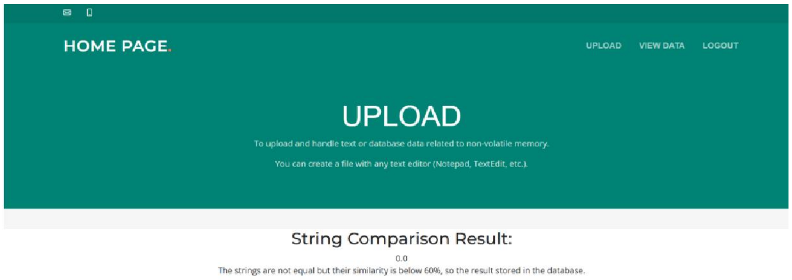

Fig 8 : String Comparison Result Page

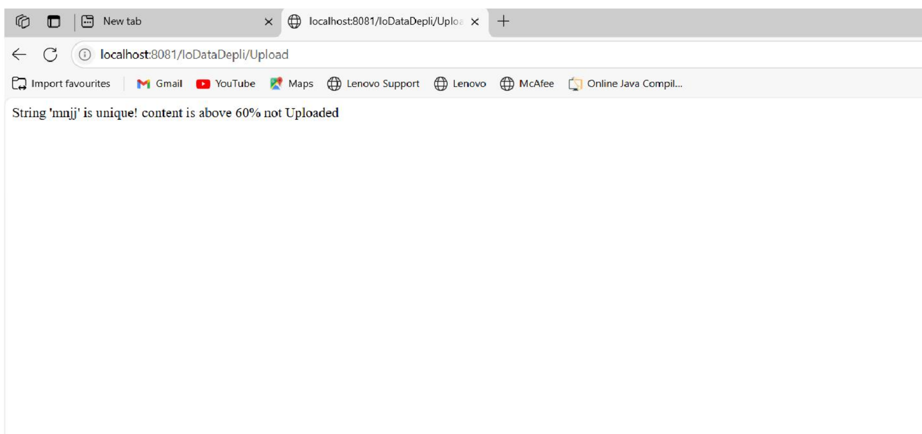

Fig 9: String Upload Validation Message

VI. CONCLUSION

In this paper, we present an I/O Causality-based Inline Deduplication (ICID) scheme for NVM-based storage. ICID records the I/O causality among in-memory file operations to achieve in-line data deduplication, and avoid the time-consuming calculation of fingerprints caused by previous chunk-based deduplication schemes. We advocate two key technologies to manage memory-copy operations in a B-tree efficiently, i.e., a hybrid data structure to store memory-copy records, and a location-dependent garbage collection scheme for deleting outdated memory-copy records. Our experimental result demonstrates that ICID achieves up to 16× higher deduplication ratio than state-of-the-art deduplication schemes, and also reduces the time overhead of data deduplication by 47% on average.

VII. FUTURE ENHANCEMENT

Existing data deduplication systems face severe fragmentation problems. The fragmentation can cause random reads and lowers the throughput of data restoration. Some proposals exploit rewriting schemes to relocate unique chunks, and thus can mitigate the fragmentation problem to some extent. A few proposals exploit other optimizations such as caching to accelerate the data restoration due to fragmentation. However, the fragmentation problem is always accompanied with the data deduplication. We can explore defragmentation technologies to mitigate this problem in an offline manner. On the other hand, since the random read performance of NVMs is much higher than that of SSDs/HDDs the fragmentation is no longer a critical problem for NVM-based storage systems.

REFERENCES

[1] K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti, "iDedup: Latency-aware, inline data deduplication for primary storage," in Proc. 10th USENIX Conf. File Storage Technol. (FAST), vol. 12, 2012, pp. 1–14.

[2] "IntelR OptaneTM Technology." 2021. Accessed: Dec. 29, 2023. [Online]. Available: https://www.intel.com/content/www/us/en/products/ docs/memory-storage/optane-persistent-memory/overview.html

[3] D. Yang, H. Liu, H. Jin, and Y. Zhang, "HMvisor: Dynamic hybrid memory management for virtual machines," Sci. China Inf. Sci., vol. 64, no. 9, pp. 192 104 1–16, 2021.

[4] Y. Xie, "Modeling, architecture, and applications for emerging memory technologies," IEEE Des. Test Comput., vol. 28, no. 1, pp. 44–51, Jan./Feb. 2011.

[5] C. Tian, H. Liu, X. Liao, and H. Jin, "UCat: Heterogeneous memory management for unikernels," Frontiers Comput. Sci., vol. 17, no. 1, pp. 171204–171215, 2022.

[6] M. K. Qureshi, J. Karidis, M. Franceschini, V. Srinivasan, L. Lastras, and B. Abali, "Enhancing lifetime and security of PCM-based main memory with start-gap wear leveling," in Proc. Annu. IEEE/ACM Int. Symp. Microarchit. (MICRO), 2009, pp. 14–23.

[7] M. Fu et al., "Design tradeoffs for data deduplication performance in backup workloads," in Proc. 13th USENIX Conf. File Storage Technol. (FAST), 2015, pp. 331–344.

[8] S. Mandal et al., "Using hints to improve inline block-layer deduplication," in Proc. 14th USENIX Conf. File Storage Technol. (FAST), 2016, pp. 315–322.

[9] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," Proc. 18th ACM Symp. Operating Syst. Princ. (SOSP), 2001, pp. 174–187.

[10] W. Xia et al., "A comprehensive study of the past, present, and future of data deduplication," Proc. IEEE, vol. 104, no. 9, pp. 1681–1710, Sep. 2016.

[11] W. Xia, H. Jiang, D. Feng, and Y. Hua, "Similarity and locality based indexing for high performance data deduplication," IEEE Trans. Comput., vol. 64, no. 4, pp. 1162–1176, Apr. 2015. 1340 IEEE TRANSACTIONS ON COMPUTERS, VOL. 73, NO. 5, MAY 2024

[12] F. Guo and P. Efstathopoulos, "Building a high-performance deduplication system," in Proc. USENIX Conf. USENIX Annu. Tech. Conf. (ATC), 2011, pp. 25–39.

[13] C. Yu, C. Zhang, Y. Mao, and F. Li, "Leap-based content defined chunking—Theory and implementation," in Proc. IEEE 31st Symp. Mass Storage Syst. Technol. (MSST), 2015, pp. 1–12.

[14] J. Park, J. Kim, Y. Kim, S. Lee, and O. Mutlu, "DeepSketch: A new machine learning-based reference search technique for post-deduplication delta compression," in Proc. 20th USENIX Conf. File Storage Technol. (FAST), 2022, pp. 247–264.

[15] "libfuse." GitHub. Accessed: Dec. 29, 2023. [Online]. Available: https:// github.com/libfuse/libfuse.git

[16] G. Fan, L. Yongkun, X. Yinlong, J. Song, and J. C. S. Lui, "SmartMD: A high performance deduplication engine with mixed pages," in Proc. USENIX Annu. Tech. Conf. (ATC), 2017, pp. 733–744.

[17] X. Nai, T. Chen, L. Yan, L. Hang, and W. Xiaoliang, "UKSM: Swift memory deduplication via hierarchical and adaptive memory region distilling," in Proc. 16th USENIX Conf. File Storage Technol. (FAST), 2018, pp. 325–340.

[18] F. Chen, T. Luo, and X. Zhang, "CAFTL: A Content-Aware flash translation layer enhancing the lifespan of flash memory based solid state drives," in Proc. 9th USENIX Conf. File Stroage Technol. (FAST), 2011, pp. 6–20.

[19] S. Quinlan and S. Dorward, "Venti: A new approach to archival data storage," in Proc. Conf. File Storage Technol. (FAST), 2002, pp. 7–21.

[20] C. Wang, Q. Wei, J. Yang, C. Chen, Y. Yang, and M. Xue, "NV-Dedup: High-performance inline deduplication for non-volatile memory," IEEE Trans. Comput., vol. 67, no. 5, pp. 658–671, May 2018.