ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

RIPPLE: A DECENTRALIZED EDGE-BASED DATA DEDUPLICATION FRAMEWORK

SHAGAM GEETHA[1], ERRABOTHU VIKAS REDDY[2], PINNINTI SAI VARUN REDDY[3], NAVULURI NIVAS 4, THATI ARUN KUMAR5

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— With its advantages in ensuring low data retrieval latency and reducing backhaul network traffic, edge computing is becoming a backbone solution for many latency-sensitive appli cations. An increasingly large number of data is being generated at the edge, stretching the limited capacity of edge storage systems. Improving resource utilization for edge storage systems has become a significant challenge in recent years. Existing solutions attempt to achieve this goal through data placement optimization, data partitioning, data sharing, etc. These approaches overlook the data redundancy in edge storage systems, which produces substantial storage resource wastage. This motivates the need for an approach for data deduplication at the edge. However, existing data deduplication methods rely on centralized control, which is not always feasible in practical edge computing environments. Ripple is a novel framework that enables edge servers to perform data deduplication in a fully decentralized manner. Unlike traditional approaches relying on centralized coordination, Ripple establishes a local data index on each edge server, empowering them to independently identify and eliminate redundant data. This decentralized design allows Ripple to:

1. efficiently detect and remove duplicate data,
2. uphold low-latency data retrieval requirements, and (removing Unwanted Data)
3. ensure data availability after deduplication.

Extensive trace-driven experiments on a real-world testbed validate the effectiveness of Ripple. Compared to existing state-of-the-art techniques, Ripple achieves a 60.42% reduction in data retrieval latency and enhances the deduplication ratio by up to 16.79%, demonstrating its practical advantages in edge computing environments.

Index Terms – Decentralized Data Deduplication, Edge Computing, Edge Storage, Cloud-Edge Computing, Storage Optimization, Distributed Systems, IoT, Resource Management, Secure Data Management, Cybersecurity.

I. INTRODUCTION

POWERED by the swift evolution of the web of things (WoT) and artificial intelligence (AI), an escalating multitude of mobile and smart devices are generating substantial data at the network edge. According to IDC projections, the global data volume is expected to reach 180ZB by 2025, with a noteworthy 70% originating from the network edge. The transmission of these data to the cloud for processing consumes excessive network resources, resulting in a considerable burden on network traffic. Concurrently, the conventional cloud computing paradigm falls short in meeting latency-sensitive applications' escalating demands for low latency, e.g., autonomous driving, AR, VR, etc. Multi-access edge computing (MEC), as a new computing paradigm, offers a remedy by relocating data and processing power from the cloud to the network edge. This ensures low-latency data processing and web services. Consequently, app vendors like YouTube and Tik Tok can deploy popular data on edge servers to minimize data retrieval latency and improve users' quality of experience. Comprise dof connected edge servers in a specific geographic area, an edge storage system (ESS) is capable of delivering data to users in the area with low latency. Unlike cloud servers, edge servers are commonly deployed at 5G/6G base stations to nearby users. Their limited physical sizes dictate their constrained resources. Their constrained storage capacities, in particular, impact the system's performance in data delivery. To improve storage resource utilization, some recent studies have attempted to optimize data placement reduce storage overheads with erasure coding and transmit data between edge servers on demand. In this study, we investigate edge data redundancy, a fundamental and uncharted issue that contributes greatly to storage resource wastage at the network edge. Edge Data

Redundancy: Powered by the mmWave communication technology, 5G base stations offer high bandwidth, and support massive concurrent connections, but suffer from limited coverage range. To avoid blank areas, 5G base stations (and edge servers) are deployed densely, reaching up to 50 base stations per km2. The users in the same specific area, e.g., a university or a CBDarea, often share similar data demands. In response to these data demands, app vendors cache popular data like viral videos and software updates on the edge servers covering the area. This inevitably results in data redundancy across these edge servers. In September 2023, we collected information on the like counts of 16,342 popular Tik Tok videos worldwide of different lengths. summarizes the results. We can see in the figure that Tik Tok users are clearly more interested in popular videos. For example, 8.53% of the most popular videos contribute 98.6% of the total likes. Cached on multiple edge servers, these popular data result in substantial data redundancy in the system. Challenges: The reisal so significant data redundancy in cloud storage systems and many deduplication approaches have been proposed in the past decade. However, suffering the following three limitations, these approaches are not suitable for edge storage systems. Limitation

1. Deduplication Granularity: The main idea shared by many of these systems is to identify and reduce common data blocks shared by different storage nodes in the system. This idea may also be implemented in ESSs. However, the downside of this idea is that data blocks have to be fetched from different storage nodes (a few hundred or thousands of data blocks) for data reconstruction in response to users' data retrieval requests. It takes time and does not align with users' requirements for low latency in MEC environments. Limitation

2. Deduplication Scope: In pursuit of a high reduplication ratio, cloud deduplication approaches try to minimize the number of replicas of any data blocks in the system to only one. It is feasible in cloud storage systems because data blocks can be fetched from any storage nodes in the system for data reconstruction. However, in an edge storage system, if an edge server has to fetch from other edge servers far away in the system for data construction, the delays introduced in users' data retrieval are undesirable. Limitation

3. Deduplication Control: Data deduplication techniques for cloud storage systems are designed based on central control and global storage information. Implementing these

techniques for edge storage systems is impractical for two main reasons. First, central control and global storage information provided from the cloud introduce an inevitable delay in data deduplication at the edge. This also goes against the low-latency requirement in MEC environments. Second, even if the delay is acceptable, maintaining the global storage information in the cloud demands constant updates from edge.

II. RELATED WORK

A. *FEUAGame: Fairness aware edge user allocation for app vendors*

Mobile edge computing (MEC) offers a new computing paradigm that turns computing and storage resources to the network edge to provide minimal service latency compared to cloud computing. Many research works have attempted to help app vendors allocate users to appropriate edge servers for high-performance service provisioning. However, existing edge user allocation (EUA) approaches have ignored fairness in users' data rates caused by interference, which is crucial in service provisioning in the MEC environment. To pursue fairness in EUA, edge users need to be assigned to edge servers so their quality of experience can be ensured at minimum costs without significant service performance differences among them. In this paper, we make the first attempt to address this fair edge user allocation (FEUA) problem. Specifically, we formulate the FEUA problem, prove its NP-hardness, and propose an optimal approach to solve small-scale FEUA problems. To accommodate large-scale FEUA scenarios, we propose a game-theoretic approach called FEUAGame that transforms the FEUA problem into a potential game that admits a Nash equilibrium. FEUA employs a decentralized algorithm to find the Nash equilibrium in the potential game as the solution to the FEUA problem. A widely-used real-world data set is utilised to experimentally compare the performance of FEUAGame to four representative approaches. The numerical outcomes show the effectiveness and efficiency of the proposed approaches in solving the FEUA problem.

B. *Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G.*

Pushing artificial intelligence (AI) from central cloud to network edge has reached board consensus in both industry and academia for materializing the vision of artificial intelligence of things (AIoT) in the sixth-generation (6G) era. This gives rise to an emerging research area known as edge intelligence, which concerns the distillation of human-like intelligence from the vast amount of data scattered at the wireless network edge. Typically, realizing edge intelligence corresponds to the processes of sensing, communication, and computation, which are coupled ingredients for data generation, exchanging, and processing, respectively. However, conventional wireless networks design the three mentioned ingredients separately in a task-agnostic manner, which leads to difficulties in accommodating the stringent demands of ultra-low latency, ultrahigh reliability, and high capacity in emerging AI applications like auto-driving and metaverse. This thus prompts a new design paradigm of seamlessly integrated sensing, communication, and computation (ISCC) in a task-oriented manner, which comprehensively accounts for the use of the data in downstream AI tasks. In view of its growing interest, this study provides a timely overview of ISCC for edge intelligence by introducing its basic concept, design challenges, and enabling techniques, surveying the state-of-the-art advancements, and shedding light on the road ahead.

C. Cost-effective data placement in edge storage systems with erasure code

Edge computing, as a new computing paradigm, brings cloud computing's computing and storage capacities to network edge for providing low latency services for users. The networked edge servers in a specific area constitute edge storage systems (ESSs), where popular data can be stored to serve the users in the area. The novel ESSs raise many new opportunities as well as unprecedented challenges. Most existing studies of ESSs focus on the storage of data replicas in the system to ensure low data retrieval latency for users. However, replica-based edge storage strategies can easily incur high storage costs. It is not cost-effective to store massive replicas of large-size data, especially those that do not require real-time access at the edge, e.g., system upgrade files, popular app installation files, videos in online games. It may not even be possible due to the constrained storage resources on edge servers. In this article, we make the first attempt to investigate the use of erasure codes in cost-effective data storage at the edge. The focus is to find the optimal strategy for placing coded data blocks on the edge servers in an ESS, aiming to minimize the storage cost while serving all the users in the system. We first model this novel Erasure Coding based Edge Data Placement (EC-EDP) problem as an integer linear programming problem and prove its NP-hardness. Then, we propose an optimal approach named EC-EDP-O based on integer programming. Another approximation algorithm named EC-EDP-V is proposed.

III. PROPOSED METHODOLOGY

The proposed system, Ripple, introduces a ground breaking approach to data deduplication in edge computing environments by eliminating the need for centralized control. This decentralized framework allows each edge server to autonomously maintain a local data index and perform deduplication tasks independently. By enabling each server to detect and eliminate redundant data at the edge, Ripple significantly improves storage efficiency while preserving the low-latency requirements essential for real-time applications. The system is designed to operate under the constraints of edge computing, such as limited resources and high distribution, and ensures high data availability even after deduplication. Ripple stands out as a scalable and practical solution for modern edge infrastructures, addressing the shortcomings of traditional centralized methods.

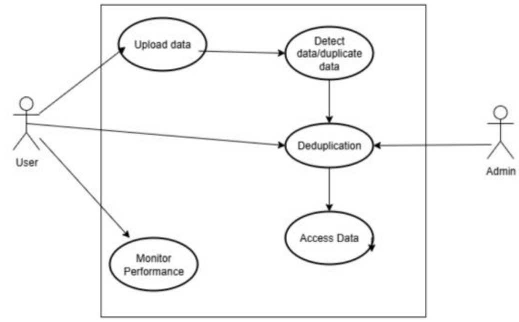
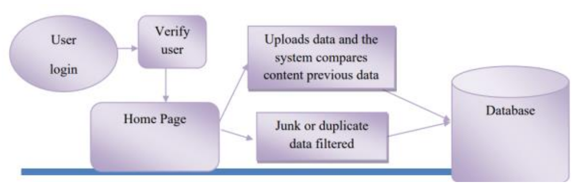


Fig. 1: System Overview

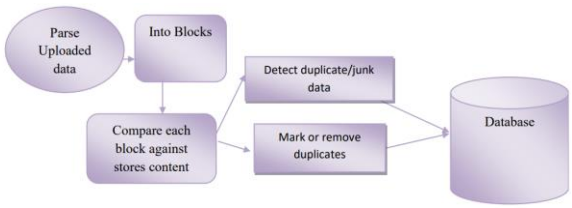
1. User Login & Data Upload

This module provides a secure login system for users who want to upload their data into the edge storage framework. Once logged in, users can upload IoT or text data, which is first checked for duplicate or junk content before storing. The system ensures that unnecessary storage consumption is avoided by validating uploaded files. It also maintains user authentication records to ensure accountability. Overall, this module acts as the entry point for data into the deduplication system.



2. Data Deduplication

The data deduplication module scans uploaded content to identify and eliminate duplicate information. It compares new files with previously stored data, checking for repeated words, sentences, or file chunks. By doing so, it saves storage space and improves overall system efficiency. The module also tracks junk data separately to maintain storage cleanliness. This ensures that only unique and useful content remains in the system.



3. Graphical Performance Representation

The graphical performance representation module visualizes the results of data deduplication through charts and graphs. It shows metrics such as deduplication ratio, storage savings, and data retrieval latency. These visual outputs make it easier for users and administrators to understand system performance. It also helps in comparing different deduplication strategies. Overall, this module ensures transparency and provides insights into the efficiency of the framework.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental results demonstrated that the framework effectively eliminated duplicate data across distributed edge nodes, significantly reducing storage requirements and minimizing redundant data transmission. The decentralized architecture enabled efficient data management without relying on a centralized server, improving system reliability and fault tolerance.

The analysis showed that the proposed framework achieved lower communication overhead and faster data synchronization compared to conventional centralized deduplication methods.

Fig. 2: Homepage

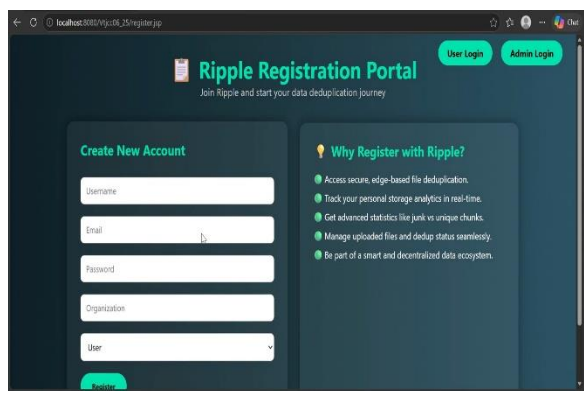


Fig. 3: Register

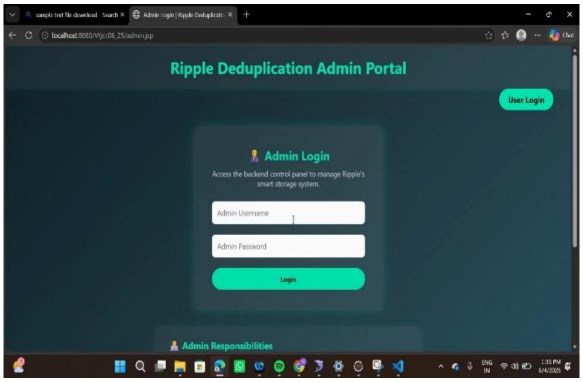


Fig. 4: Login

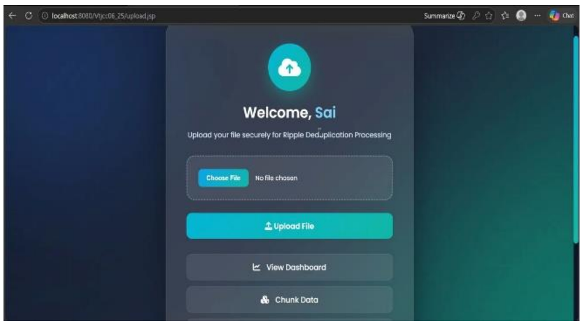


Fig. 5: Upload

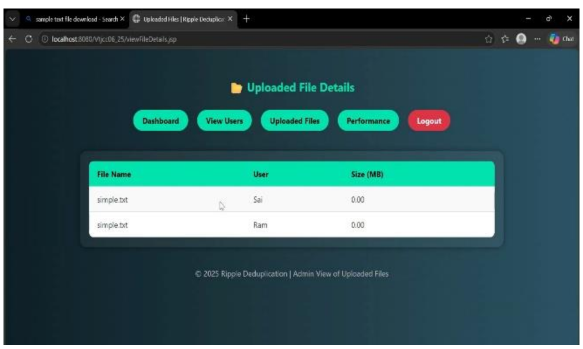


Fig.6: View Uploaded File

Fig.7: System Overview

Fig.8: System Activity

V. CONCLUSION

This work introduced Ripple, the first practical decentralized framework for deduplicating data in edge storage systems. By leveraging the Ripple tree data structure, the framework efficiently indexes data across neighbouring edge servers, enabling duplicate elimination without compromising data availability or violating latency constraints. Ripple's dynamic tree expansion and contraction mechanisms effectively reduce memory overhead, ensuring efficient resource utilization. Experimental results demonstrate its superiority over existing methods, achieving up to 16.79% improvement in deduplication ratio and an average reduction of 60.42% in data retrieval latency. These results confirm Ripple's capability to deliver both high storage efficiency and low latency performance in MEC environments.

REFERENCES

[1] W. Shi, G. Pallis, and Z. Xu, "Edge computing," Proc. IEEE, vol. 107, no. 8, pp. 1474–1481, Aug. 2019.

[2] AWS, "AWS Wavelength for media & entertainment," 2021. [Online]. Available: https://d1.awsstatic.com/Wavelength2020/AWS-Wavelength for-Media-EntertainmentSolutionBrief-Feb2021-Final.pdf

[3] G. Zhu et al., "Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G," Sci. China Inf. Sci., vol. 66, no. 3, 2023, Art. no. 130301.

[4] Q. He, Z. Dong, F. Chen, S. Deng, W. Liang, and Y. Yang, "Pyramid: Enabling hierarchical neural networks with edge computing," in Proc. ACMWebConf., 2022, pp. 1860–1870.

[5] H. Jin, R. Luo, Q. He, S. Wu, Z. Zeng, and X. Xia, "Cost-effective data placement in edge storage systems with erasure code," IEEE Trans. Serv. Comput., vol. 16, no. 2, pp. 1039–1050, Mar./Apr. 2023.

[6] W. Xiao, Y. Hao, J. Liang, L. Hu, S. A. Alqahtani, and M. Chen, "Adaptive compression offloading and resource allocation for edge vi sion computing," IEEE Trans. Cogn. Commun. Netw., to be published, doi: 10.1109/TCCN.2024.3400820.

AUTHORS Profile

Ms. SHAGAM GEETHA has received her B.Tech in CSE from Nagarjuna Institute Of Technology JNTUH in 2012, M. Tech in Computer Science and Engineering Vidya Vikas Institute of Technology JNTUH, in 2014 and She has been dedicated to the teaching field for the last 8 years. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.

Mr. ERRABOTHU VIKAS REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. PINNINTI SAI VARUN REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. NAVULURI NIVAS pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. THATI ARUN KUMAR pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.