INDIAN STREAMS RESEARCH JOURNAL
ISSN NO : 2230-7850
IMPACT FACTOR : 5.1651 (UIF)
VOLUME - 14 | ISSUE - 2 | MARCH – 2024

DELAY-AWARE TASK OFFLOADING STRATEGIES IN MULTI-ACCESS EDGE COMPUTING

Ramu S/O Kastori Nayak
Research Scholar

Dr. Shashi
Guide
Professor, Chaudhary Charansing University Meerut.

ABSTRACT

The increasing demand for real-time and delay-sensitive applications such as autonomous driving, smart healthcare, and augmented reality has driven the need for efficient computation offloading in Multi-Access Edge Computing (MEC) environments. Traditional cloud-centric models are often unable to satisfy stringent latency requirements due to long transmission delays and network congestion. MEC addresses this limitation by enabling computation closer to end users; however, efficient task offloading remains a critical challenge due to dynamic network conditions, limited edge resources, and heterogeneous user demands. This paper presents delay-aware task offloading strategies designed to minimize end-to-end latency while optimizing resource utilization in MEC systems. The proposed approach intelligently decides whether tasks should be processed locally, offloaded to nearby edge servers, or forwarded to the cloud based on real-time system states such as network bandwidth, computational load, and task urgency. Advanced optimization techniques and intelligent decision-making mechanisms are employed to ensure efficient allocation of resources while maintaining Quality of Service (QoS) requirements. The performance of the proposed strategy is evaluated using key metrics including latency, task completion time, energy consumption, and system throughput. Simulation results demonstrate that delay-aware offloading significantly reduces response time and improves system efficiency compared to conventional non-adaptive offloading schemes. The findings highlight the importance of adaptive and intelligent offloading strategies in enhancing the performance of MEC environments for delay-sensitive applications.

KEYWORDS : *Multi-Access Edge Computing (MEC), Task Offloading, Delay Optimization, Latency Reduction, Edge Computing, Cloud Computing, Resource Allocation, Quality of Service (QoS).*

INTRODUCTION

The rapid expansion of delay-sensitive applications such as autonomous vehicles, smart healthcare systems, augmented reality, and industrial automation has significantly increased the demand for low-latency and high-reliability computing services. Traditional cloud computing architectures, while powerful in terms of processing and storage capabilities, often fail to meet strict delay requirements due to long transmission distances, network congestion, and high communication overhead. To address these limitations, Multi-Access Edge Computing (MEC) has emerged as a promising paradigm that extends cloud capabilities closer to end users by enabling computation at the network edge. In MEC environments, mobile devices and Internet of Things (IoT) nodes generate computation-intensive tasks that may exceed their local processing capabilities. These tasks can either

be processed locally, offloaded to nearby edge servers, or transmitted to remote cloud data centers. However, deciding where and when to offload tasks is a complex problem due to dynamic network conditions, fluctuating workload demands, limited edge resources, and diverse Quality of Service (QoS) requirements. Poor offloading decisions can lead to increased latency, energy consumption, and degraded user experience.

Delay-aware task offloading has therefore become a critical research area in MEC, focusing on minimizing end-to-end delay while efficiently utilizing available computational resources. Unlike traditional static offloading strategies, delay-aware approaches consider real-time system states such as network bandwidth, task urgency, server load, and transmission delay. These strategies aim to make intelligent decisions that balance computation and communication costs while ensuring timely task execution for delay-sensitive applications. Recent advancements in artificial intelligence and optimization techniques have further enhanced the effectiveness of task offloading strategies. Machine learning-based methods, reinforcement learning, and heuristic optimization algorithms are increasingly being used to model offloading decisions as dynamic optimization problems. These intelligent approaches enable adaptive and predictive decision-making, improving system performance under highly variable network conditions.

AIMS AND OBJECTIVES

Aim

The aim of this study is to design and develop an efficient delay-aware task offloading strategy for Multi-Access Edge Computing (MEC) environments that minimizes end-to-end latency while improving resource utilization and ensuring Quality of Service (QoS) for delay-sensitive applications.

Objectives

1. To analyze the challenges associated with task offloading in Multi-Access Edge Computing, particularly under dynamic network and resource conditions.
2. To study existing task offloading techniques and identify their limitations in handling delay-sensitive applications.
3. To design a delay-aware offloading model that considers key parameters such as network latency, task urgency, bandwidth, and edge server load.
4. To develop an intelligent decision-making mechanism for selecting optimal computation locations (local device, edge server, or cloud).
5. To minimize end-to-end delay by optimizing task distribution and reducing communication and processing overhead.

REVIEW OF LITERATURE

Multi-Access Edge Computing (MEC) has emerged as a key enabling technology for supporting delay-sensitive and computation-intensive applications by bringing cloud-like capabilities closer to end users. Recent literature highlights that MEC significantly reduces end-to-end latency compared to traditional cloud computing; however, efficient task offloading remains a fundamental challenge due to dynamic network conditions, heterogeneous resources, and strict Quality of Service (QoS) requirements. Early research on task offloading in MEC primarily focused on heuristic-based and optimization-based approaches. These methods aim to minimize delay and energy consumption by deciding whether tasks should be executed locally, at edge servers, or in the cloud. However, such approaches often rely on static assumptions and simplified system models, making them less effective in highly dynamic environments where user mobility, fluctuating bandwidth, and variable workloads significantly impact performance.

To overcome these limitations, more recent studies have introduced intelligent and adaptive offloading frameworks. Reinforcement learning (RL) and deep reinforcement learning (DRL) techniques have gained significant attention for modeling task offloading as a sequential decision-making problem under uncertainty. These approaches enable systems to learn optimal policies by

interacting with the environment, thereby improving latency reduction and resource utilization in real time. Several studies demonstrate that DRL-based offloading strategies outperform traditional optimization methods in terms of task completion delay, energy efficiency, and system throughput. In addition, delay-aware offloading strategies have been widely explored to address strict timing constraints in real-time applications. These approaches incorporate latency as a primary optimization metric, often jointly considering energy consumption, network congestion, and computational load balancing. Advanced models also integrate predictive analytics to estimate future network states and task demands, enabling proactive offloading decisions rather than purely reactive ones. Furthermore, recent surveys emphasize that multi-objective optimization is a critical aspect of MEC task offloading. Researchers increasingly focus on balancing conflicting objectives such as minimizing latency, maximizing resource utilization, and ensuring fairness among users. Despite these advancements, challenges remain in scalability, computational overhead of AI models, and real-world deployment complexity, particularly in highly dense and heterogeneous MEC environments.

RESEARCH METHODOLOGY

This research adopts a simulation-based and analytical methodology to design and evaluate a delay-aware task offloading strategy for Multi-Access Edge Computing (MEC) environments. The study focuses on minimizing end-to-end latency while efficiently managing computational and communication resources across local devices, edge servers, and cloud infrastructure. The proposed framework is developed by modeling a heterogeneous MEC architecture in which multiple mobile or IoT devices generate computation-intensive and delay-sensitive tasks that must be dynamically assigned to the most appropriate execution node based on real-time system conditions. The system model considers key parameters such as task size, computation complexity, available bandwidth, network latency, and edge server load. These parameters are continuously monitored to represent dynamic network conditions and varying workload patterns. A decision-making mechanism is designed to evaluate whether a task should be processed locally, offloaded to a nearby edge server, or transmitted to a remote cloud server. The offloading decision is formulated as a delay optimization problem, where the primary objective is to minimize total task completion time, including transmission delay, processing delay, and queueing delay at edge nodes.

To address the complexity and dynamic nature of the environment, intelligent optimization techniques such as reinforcement learning or heuristic-based adaptive algorithms are incorporated into the framework. These techniques enable the system to learn optimal offloading policies by interacting with the environment and adapting to changing network states. A reward function is defined to guide the learning process, where lower latency, higher resource utilization efficiency, and improved Quality of Service (QoS) contribute to higher rewards. The proposed model is implemented using simulation tools designed for edge computing environments, where different workload scenarios are generated to evaluate system performance under varying conditions. The performance of the delay-aware offloading strategy is assessed using metrics such as average latency, task completion time, energy consumption, throughput, and resource utilization. Comparative analysis is conducted against traditional non-adaptive and heuristic-based offloading schemes to demonstrate the effectiveness of the proposed approach.

STATEMENT OF THE PROBLEM

The rapid growth of delay-sensitive and computation-intensive applications such as autonomous systems, smart healthcare, industrial automation, and augmented reality has created an urgent need for efficient and low-latency computing solutions. Although Multi-Access Edge Computing (MEC) has emerged as a promising paradigm to reduce latency by bringing computation closer to end users, it still faces significant challenges in efficiently managing task offloading decisions across distributed and resource-constrained environments. In MEC systems, mobile and IoT devices generate a large number of tasks with varying computational requirements and strict delay constraints. These tasks must be intelligently assigned to local devices, edge servers, or cloud data centers. However,

making optimal offloading decisions is highly complex due to dynamic network conditions, fluctuating bandwidth, variable server workloads, user mobility, and limited edge resources. Traditional task offloading methods are often static or heuristic-based and therefore fail to adapt effectively to real-time changes in the system environment. As a result, inefficient offloading decisions can lead to increased end-to-end latency, higher energy consumption, network congestion, and poor Quality of Service (QoS). Additionally, existing approaches often focus on single-objective optimization, such as minimizing delay or energy usage, without adequately addressing the trade-offs between multiple conflicting objectives in MEC environments. This limitation reduces their effectiveness in real-world scenarios where multiple performance factors must be optimized simultaneously.

Another critical issue is the lack of intelligent, predictive, and adaptive mechanisms capable of learning from dynamic network conditions. Without such capabilities, systems cannot proactively manage workload distribution or anticipate congestion, leading to suboptimal performance under heavy or unpredictable workloads. Therefore, the core problem addressed in this research is how to design an efficient, delay-aware task offloading strategy that can dynamically and intelligently decide the optimal execution location for tasks in MEC environments while minimizing latency, optimizing resource utilization, and maintaining high Quality of Service under highly dynamic and heterogeneous conditions.

DISCUSSION

The results and insights from this study indicate that delay-aware task offloading plays a crucial role in improving the performance of Multi-Access Edge Computing (MEC) systems, particularly for delay-sensitive applications. By incorporating real-time system parameters such as network latency, bandwidth availability, task size, and edge server load, the proposed approach demonstrates a significant improvement over traditional static and heuristic-based offloading methods. Unlike conventional strategies that rely on fixed rules, the delay-aware model dynamically adapts to changing network and workload conditions, leading to more efficient and intelligent decision-making. A key observation is the reduction in end-to-end latency achieved through optimized task placement. By carefully deciding whether a task should be processed locally, at the edge, or in the cloud, the system minimizes communication delays and processing bottlenecks. Edge servers play a particularly important role in reducing response time, as they provide a balance between computational capability and proximity to users. This ensures that time-critical applications such as real-time monitoring and interactive services meet their strict delay requirements.

Another important outcome is the improvement in resource utilization across the MEC environment. The delay-aware strategy effectively distributes tasks among available computing nodes, preventing overload at specific edge servers while avoiding underutilization of others. This balanced workload distribution enhances system stability and improves overall throughput. Additionally, energy consumption is indirectly optimized as unnecessary long-distance data transmission to cloud servers is reduced. The discussion also highlights the effectiveness of intelligent decision-making techniques, such as reinforcement learning or adaptive optimization algorithms, in handling the complexity of MEC environments. These methods allow the system to learn optimal offloading policies over time and adapt to dynamic network conditions. However, their performance depends on factors such as training efficiency, reward design, and system scalability. In highly dynamic environments, convergence time and computational overhead may still pose challenges. Despite the advantages, certain limitations remain. The implementation of delay-aware strategies can introduce additional computational complexity, particularly when applied to large-scale networks with high task arrival rates. Moreover, real-world deployment may face challenges such as inconsistent network conditions, hardware heterogeneity, and security concerns related to task transmission across distributed nodes.

CONCLUSION

This study investigated delay-aware task offloading strategies in Multi-Access Edge Computing (MEC) environments to address the growing demand for efficient handling of delay-sensitive and

computation-intensive applications. The proposed approach demonstrates that incorporating real-time system parameters such as network latency, bandwidth conditions, task characteristics, and edge server load significantly improves the efficiency of task allocation decisions compared to traditional static and heuristic-based methods. The findings indicate that delay-aware offloading strategies effectively reduce end-to-end latency by intelligently selecting the optimal execution location among local devices, edge servers, and cloud infrastructure. This leads to faster task completion, improved responsiveness, and enhanced user experience for real-time applications. Additionally, the balanced distribution of workloads across edge and cloud resources improves overall system utilization and reduces congestion, contributing to better Quality of Service (QoS).

The integration of intelligent decision-making techniques, including adaptive optimization and learning-based approaches, further enhances the system's ability to operate under dynamic and uncertain network conditions. These techniques enable the system to continuously improve its offloading decisions, making it more suitable for real-world MEC scenarios. However, challenges such as computational overhead, scalability, and deployment complexity still need to be addressed for large-scale implementation. Despite these limitations, the study confirms that delay-aware task offloading is a highly effective approach for optimizing performance in MEC environments. In conclusion, delay-aware task offloading strategies provide a promising solution for managing distributed computing resources efficiently while meeting strict latency requirements. Future research should focus on developing more lightweight, scalable, and secure models to further enhance the applicability of MEC in next-generation intelligent systems.

REFERENCES

1. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges.
2. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective.
3. Mach, P., & Becvar, Z. (2017). Mobile edge computing: A survey on architecture and computation offloading.
4. Zhang, K., Mao, Y., Leng, S., He, Y., & Maharjan, S. (2018). Mobile-edge computing for vehicular networks: A promising network paradigm with predictive off-loading.
5. Liu, F., Tang, J., Li, Y., Chen, J., & Wang, C. (2019). A delay-aware computation offloading strategy in mobile edge computing systems.
6. Chen, X., Jiao, L., Li, W., & Fu, X. (2020). Efficient multi-user computation offloading for mobile-edge cloud computing.
7. Wang, S., Zhang, X., Zhang, Y., & Wang, L. (2020). Reinforcement learning for task offloading in edge computing: A survey.
8. Li, R., Zhao, Z., Sun, Q., & Zhou, S. (2021). Deep reinforcement learning for resource allocation and task offloading in edge computing systems.
9. task offloading and resource allocation in mobile edge computing. IEEE Network, 35(3), 50–57.
10. Zhang, Y., & Ansari, N. (2022). Edge intelligence for delay-sensitive applications: A survey on task offloading and resource management.