International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03640
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

"Delay-Constrained Task Offloading and Resource Optimization in Edge-Cloud Networks"

Mr. Mula Mahender*1, M.Hithesh*2, G.Aswitha*3, Nithin Goud*4,

*[1]Associate Professor, CSE (AI & ML), ACE Engineering College, Hyderabad, India.

*[2,3,4] Students of Department CSE (AI & ML), ACE Engineering College, Hyderabad, India.

ABSTRACT-Delay-Constrained Task Offloading and Resource Optimization in Edge-Cloud Networks A joint optimization model is presented to reduce task latency and energy expenditure while satisfying application deadlines. The problem is cast as a mixed-integer nonlinear program (MINLP), and a hybrid solution integrating deep reinforcement learning and convex optimization is formulated to handle its complexity. This paper is extremely applicable to latency-sensitive IoT, augmented reality, and smart transportation applications. Our method is shown through extensive simulations to significantly reduce latency and save energy compared to conventional methods. In comparison to random or cloud-only approaches, our model guarantees timely completion of delay-sensitive tasks while optimizing resource usage throughout the network. This framework has great potential for facilitating faster, greener, and smarter future computing services in hybrid edge-cloud networks.

Keywords: Edge Computing, Cloud Computing, Task Offloading, Delay Constraints.

I.INTRODUCTION

The fast growth of smart technologies and real-time applications has raised an emergent demand for computing systems that can provide low-latency responses while keeping energy efficiency. Hybrid edge-cloud networks that leverage the local proximity of edge servers and the enormous computing capacity of centralized clouds have come out as a potential solution to address these demands. But distributing tasks between edge and cloud servers efficiently is not simple, particularly when stringent delay constraints and energy optimization are needed. Task offloading decisions need to consider multiple factors carefully such as task size, server capacity, delay of transmission, and power usage. Static offloading approaches tend to ignore dynamic network conditions and thus cause increased delays along with energy wastage. In this project, we introduce a smart task allocation framework that selects dynamically between cloud and edge processing. Employing a hybrid of deep reinforcement learning and convex optimization, our system tries to minimize task delay as well as energy consumption.

II LITERATURE REVIEW

2.1 C. Zhang, P. Patras, H. Haddadi(2019): Deep Learning in Mobile and Wireless Networking: A Survey

provides a thorough overview of how deep learning (DL) methods are being applied to mobile and wireless networking. It presents the difficulties brought by the rising complexity and heterogeneity of mobile environments and examines the ways DL can tackle problems such as traffic forecasting, resource allocation, and anomaly detection. Existing research is classified under several domains in the survey and emphasizes the role of cost-effective deployment strategies on mobile systems. It also suggests present challenges and delineates future research avenues to promote the use of DL in this area.

2.2 Rafia Malik(2020):Energy-Efficient Computation Offloading in Delay-Constrained Massive MIMO Enabled Edge Network Using Data Partitioning.

Tackles energy minimization in wireless edge computing systems. Such systems allow several users offload compute-intensive applications to huge MIMO access points with each having a collocated multi-access edge

computing (MEC) server. Formulates a problem of minimizing weighted sum of energy expenditure at both the users and the MEC server subject to a round-trip latency constraint, using data partitioning, transmit power control, and CPU frequency scaling at both sides. A new nested primal-dual algorithm is proposed to efficiently solve this problem. The results show that partial offloading with data partitioning greatly saves total energy consumption compared to binary offloading, particularly under hard latency constraints. The research also points out the effects of massive MIMO channel estimation errors and pilot contamination on energy efficiency.

2.3 Nguyen TT, Le LB, Le-Trung Q (2019): Computation Offloading in MIMO Based Mobile Edge Computing Systems Under Perfect and Imperfect CSI Estimation

Explores the offloading computation issue in big multiple-input multiple-output (MIMO)-based mobile edge computing (MEC) systems. The research focuses on reducing power consumption and offloading latency at mobile devices with random system environments. To this end, the authors define the problem as a Markov Decision Process (MDP) and suggest two Deep Reinforcement Learning (DRL) algorithms: a Deep Q-Network (DQN)-based algorithm to mitigate the curse of state space explosion and a Proximal Policy Optimization (PPO)-based algorithm to tackle discrete action spaces. Simulation outcomes reveal that the designed DRL-based techniques perform better than top-notch algorithms, with the PPO algorithm providing steady performance and optimal offloading outcomes in comparison to the DQN and Double DQN (DDQN) approaches.

2.4 Xiong X, Zheng K, Lei L, Hou L (2019): Resource Allocation Based on Deep Reinforcement Learning in IoT Edge Computing

Overcomes the problem of effective resource allocation in Internet of Things (IoT) edge computing systems. The research sets up the resource allocation problem as a Markov Decision Process (MDP) and suggests a Deep Q-Network (DQN)-based solution to optimize computation, communication, and storage resource allocation. The suggested technique minimizes latency and energy consumption while satisfying Quality of Service (QoS) requirements. Simulation results show that the DQN-based approach performs better than conventional methods regarding system performance and adaptability to dynamic environments.

2.5 Chen Z, Wang X (2020): Decentralized Computation Offloading for Multi-User Mobile Edge Computing: A Deep Reinforcement Learning Approach

Overcomes the challenge of computation offloading in multi-user mobile edge computing (MEC) systems. The paper presents a decentralized dynamic computation offloading scheme based on Deep Deterministic Policy Gradient (DDPG), a deep reinforcement learning (DRL) technique. Every user learns a best offloading policy from local information independently, without access to global system information. The new method aims to reduce long-term average computation cost, such as power and buffering delay. Simulation results show that the DDPG-based approach is better than conventional approaches like Deep Q-Network (DQN) and greedy strategies in both energy efficiency and delay performance. The work indicates the superiority of DRL in decentralized MEC systems with low feedback and changing environments.

2.6 Chubo Liu, Fan Tang, Yikun Hu. (2021): Distributed Task Migration Optimization in MEC by Extending Multi-Agent Deep Reinforcement Learning Approach

presents a cooperative task scheduling algorithm called CA-DTS for Mobile Edge Computing (MEC) systems. The algorithm formulates task scheduling between distributed Road Side Units (RSUs) as a Markov game and uses Counterfactual Multi-Agent Policy Gradient (COMA) to support decentralized decision-making. To improve cooperation between multiple RSUs, CA-DTS uses an Action Semantics Network (ASN), which helps to decrease training expenses and enhance performance stability in large-scale scenarios. Simulation outcomes verify that CA-DTS converges around 35% more quickly and decreases average task delay by 6.7% to 9.8% over baseline algorithms in different scenarios.

2.7Haipeng Yao, Xiangjun Xin, Ran Gao, Mohsen Guizani (2023): Joint Multi-Task Offloading and Resource Allocation for Mobile Edge Computing Systems in Satellite IoT

Faced the issue of effective handling of multi-task offloading and resource allocation in satellite-based Mobile Edge Computing (MEC) systems. In such systems, tasks tend to have dependencies and must be offloaded jointly to achieve optimal resource utilization.

2.8 MChen, Y. Zhang, Y. Li, M. Shikh-Bahaei (2019): Dynamic Computation Offloading in Edge Computing for Internet of Things

The problem of offloading computationally intensive tasks from Internet of Things (IoT) devices into edge computing facilities. Due to the random nature of task generation as well as the resource-constrained nature of IoT devices, the authors introduce a Dynamic Computation Offloading Algorithm (DCOA) that breaks

down the optimization problem into a sequence of subproblems and solves them online and in parallel. Theoretical analysis proves that DCOA attains a balance between offloading cost and performance and thus can be a potential solution for real-time applications in IoT environments.

2.9 Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, J. Zhang (2019): Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing

Discovers the unification of Artificial Intelligence (AI) and Edge Computing, a paradigm referred to as Edge Intelligence (EI). This cross-disciplinary domain seeks to make AI nearer to the source of data, e.g., Internet of Things (IoT) devices, through computation near the network edge instead of centralized cloud-only computation.

2.11. Comparison Table: Literature Review on Task Offloading and Resources optimization.

Author(s)	Titles	Methodology	Contribution	Limitation
C. Zhang, P. Patras, H. Haddadi(2019)	Deep Learning in Mobile and Wireless Networking: A Survey	Survey of deep learning applications (deep reinforcement learning for dynamic offloading in edge computing included)	DL enables the forecasting of network traffic and user behavior, enabling proactive network management. This feature is essential in forecasting congestion and optimizing resource allocation .	DL models need considerable amounts of quality data for their training. Obtaining this may be difficult within mobile networks since privacy becomes a major concern while the dynamic setting poses issues too.
Rafia Malik, Mai Vu(2020)	Energy-Efficient Computation Offloading in Delay-Constrained Massive MIMO Enabled Edge Network Using Data Partitioning	State delay constraints (how much delay can be tolerated for the offloading process). Establish the energy consumption model for edge servers and edge devices and the communication power consumption for offloading	The authors introduce a system model comprising three phases: data offloading from users to the Multi-Access Edge Computing (MEC) server, computation at both the user and MEC, and result downloading. This model facilitates a comprehensive analysis of energy	Perfect CSI is assumed by the analysis, which might not be true in real-world applications owing to causes such as mobility and environmental dynamics. This can affect the performance of the solutions suggested.

			consumption across the network	
Nguyen TT, Le LB, Le-Trung Q(2019)	Computation Offloading in MIMO Based Mobile Edge Computing Systems Under Perfect and Imperfect CSI Estimation	Analytical models and simulations to measure computation offloading performance	The work presents a decision framework for computation offloading based on considerations such as CSI quality, energy cost, and latency.	Although offloading may lower device energy consumption, the energy used by the MEC server and network equipment is not always considered.
Xiong X, Zheng K, Lei L, Hou L(2019)	Resource Allocation Based on Deep Reinforcement Learning in IoT Edge Computing	Deep Reinforcement Learning (DRL) is a subfield of machine learning that combines reinforcement learning (RL) ideas with deep learning techniques	The paper breaks new ground in using DRL for intelligent and dynamic resource allocation in IoT edge computing systems to maximize performance parameters like latency, energy use, and throughput.	Training DRL models is computationally intensive and may not be practical in resource-limited IoT setting.
Chen Z, Wang X(2020)	Decentralized Computation Offloading for Multi-User Mobile Edge Computing: A Deep Reinforcement Learning Approach	Deep Reinforcement Learning (DRL) is a subfield of machine learning that combines reinforcement learning (RL) ideas with deep learning techniques.	The paper proposes a decentralized method whereby every mobile user learns a global optimum computation offloading policy separately from local observations without any requirement of global system knowledge.	Training DRL models to be particularly in decentralized environments, however, can be computationally expensive, thus limiting scalability in big-scale MEC systems.

Chubo Liu, Fan Tang, Yikun Hu.(2021)	Distributed Task Migration Optimization in MEC by Extending Multi-Agent Deep Reinforcement Learning Approach	Multi-Agent Deep Reinforcement Learning	The research formulates task scheduling among distributed RSUs as a Markov game, allowing each RSU to independently make choices based on the actions of other RSUs.	Though the method proves to be effective in simulated scenarios, scalability to a greater number of RSUs and more intricate network topologies is still an issue.
Furong Chai, Qi Zhang, Haipeng Yao, Xiangjun Xin, Ran Gao, Mohsen Guizani(2023)	Joint Multi-Task Offloading and Resource Allocation for Mobile Edge Computing Systems in Satellite IoT	Joint multi-task offloading and resource allocation scheme using attention mechanism and proximal policy optimization (PPO)	The research proposes an integrated framework under which several tasks with dependencies are gathered and offloaded together by Unmanned Aerial Vehicles (UAVs) as airbase stations, with edge computing offered by satellites.	The suggested framework relies on perfect communication conditions between satellites and UAVs, which might not be the case in real-world environments with changing channel qualities and interference.
M. Chen, Y. Zhang, Y. Li, M. Shikh-Bahaei(2019)	Dynamic Computation Offloading in Edge Computing for Internet of Things	Reinforcement Learning (RL)-based offloading policy	The offloading computation problem as an optimization problem that minimizes offloading cost with the guarantee of performance. This is done while taking into consideration the dynamic nature of task	The model presupposes that task arrivals are stochastic processes, which might not hold for all actual situations. Periodic or predictable task arrivals might need alternative handling strategies.

			arrivals and network conditions.	
Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, J. Zhang(2019)	Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing	Review of AI-powered offloading and optimization	By shifting AI computations to the edge, the paper highlights less data transmission to centralized cloud servers, resulting in reduced latency and enhanced real-time decision-making performance.	Edge devices usually lack high processing capabilities, memory, and power supply, which limits the size and complexity of AI models that can be utilized optimally.

3. Research Gaps:

i.Energy Efficiency Enhancement:
The findings indicate that combined MU-MIMO precoding and computation offloading greatly enhance the energy efficiency of IoT systems. Through the application of MU-MIMO for interference management and offloading computation to the edge, the energy consumption is greatly minimized compared to conventional single-user communication models or local processing only. Memory Organization and Encoding.

ii.Delay Constrained Offloading:
The joint optimization guarantees delay constraints are satisfied even when computationally intensive tasks are delegated to the edge servers. This is particularly critical in industrial applications when real-time processing is imperative. Inability to Handle Topic Shifts. The research demonstrates that the offloading decisions, when optimally made, satisfy the delay constraints without sacrificing the task completion times, even in situations with strict real-time processing requirements.

iii.Energy Consumption Models:
The energy consumption models are specified for both the edge devices and the edge servers, such as the computation power consumed and the communication costs of offloading tasks to the edge server.
Energy models comprise:
Edge devices and server computation energy.
Communication energy for offloading, considering the energy used in transferring data to and from the edge server.

iv.Scalability and Practical Applicability:
The approach is scalable to big systems, hence can be used in industrial IoT networks with a high number of devices. The method is also practical for use in situations with different load and task sizes, offering flexibility for industrial applications of different.

4. CONCLUSION

In this work, we analyzed a delay-constrained edge-cloud computing network with the purpose of enhancing the computation and energy resource utilization efficiency. Assuming the delay constraint, two MINLP problems were presented by simultaneously optimizing offloading task decision and computation resource allocation. In order to address the quantity driven issue, we designed a BTTO scheme that has the capability of searching for the optimal task offloading decisions with low complexity. Paving on convex optimization and BnB methodology, we made an AO algorithm proposal for tackling the energy driven problem. Experimental results confirm that proposed algorithms can be well proved through performance results.

5. REFERENCES

1. Z. Sharif, L. T. Jung, I. Razzak, and M. Alazab, "Adaptive and priority based resource allocation for efficient resources utilization in mobile edge computing," IEEE Internet Things J., vol. 10, no. 4, pp. 3079–3093, Feb. 2023.

2.L. Li et al., "Data-driven optimization for cooperative edge service provisioning with demand uncertainty," IEEE Internet Things J., vol. 8, no. 6, pp. 4317–4328, Mar. 2021.

3.Q. Peng, C. Wu, Y. Xia, Y. Ma, X. Wang, and N. Jiang, "DoSRA: A decentralized solution to online edge task scheduling and resource allocation," IEEE Internet Things J., vol. 9, no. 6, pp. 4677–4692, Mar. 2022.

4.Y. Liu, H. Yu, S. Xie, and Y. Zhang, "Deep reinforcement learning

for offloading and resource allocation in vehicle edge computing and networks," IEEE Trans. Veh. Technol., vol. 68, no. 11, pp. 11158–11168, Nov. 2019.

5.Z. Wu and D. Yan, "Deep reinforcement learning-based computation offloading for 5G vehicle-aware multi-access edge computing network," China Commun, vol. 18, no. 11, pp. 26–41, Nov. 2021.

6.Q. Yuan, B. Chen, G. Luo, J. Li, and F. Yang, "Integrated route plan ning and resource allocation for connected vehicles," China Commun., vol. 18, no. 3, pp. 226–239, Mar. 2021.

7.L. Tan, Z. Kuang, L. Zhao, and A. Liu, "Energy-efficient joint task offloading and resource allocation in OFDMA-based collaborative edge computing," IEEE Trans. Wireless Commun., vol. 21, no. 3,pp. 1960–1972, Mar. 2022.

8.L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "NOMA-enabled mobile edge computing for Internet of Things via joint communicationand computation resource assignments," IEEE Internet Things J., vol. 7, no. 1, pp. 718–733, Jan. 2020.