Optimizing Resource Allocation and Task Offloading for Real-Time IoT Applications

Dr. Umadevi Ramamoorthy

(School of Science and Computer Studies, CMR University,Bengaluru,India)

Dhanush M

(School of Science and Computer Studies, CMR University,Bengaluru,India)

Abstract

This research investigates strategies to optimize resource allocation and task offloading in Edge Computing environments for real-time Internet of Things (IoT) applications. As IoT devices continue to proliferate, edge computing has emerged as a vital approach to meet demands for low latency and real-time responsiveness. However, limited computational and energy resources at edge nodes pose significant challenges. This study reviews existing algorithms and proposes a hybrid offloading framework that combines heuristic-based task scheduling with AI-driven resource prediction. Simulations were conducted using a testbed of emulated IoT devices and edge nodes under variable workloads and network conditions. The results demonstrate a noticeable improvement in processing delay, energy consumption, and resource utilization compared to static or cloud-only approaches. By enhancing decision-making at the edge, this paper contributes to the growing body of work aiming to make distributed computing more intelligent, sustainable, and responsive to real-world constraints in IoT environments.

Keywords

Edge computing, task offloading, resource allocation, real-time IoT applications, mobile edge computing (MEC), latency reduction, energy efficiency, scheduling algorithms, fairness-aware allocation, ultra-dense networks, bandwidth optimization, low-latency communication, quality of service (QoS), distributed computing, edge intelligence, context-aware computing, dynamic resource management, vehicular edge computing, cloud-edge collaboration, multi-access edge computing (MAEC).

Introduction

Background

Edge computing brings computation and data storage closer to the location where it is needed, especially crucial in IoT applications like smart homes, autonomous vehicles, healthcare monitoring, and industrial automation. Unlike traditional cloud computing, edge computing reduces latency, saves bandwidth, and enables real-time decision-making.

Problem Statement

While edge computing offers low-latency processing, edge nodes often have limited processing power, memory, and energy capacity. This creates a bottleneck when multiple IoT tasks require real-time responses. The challenge lies in deciding which tasks to process locally and which to offload to other edge nodes or the cloud, while efficiently managing available resources.

Research Question

How can task offloading and resource allocation be optimized in edge computing environments to support real-time IoT applications?

Thesis Statement

This paper argues that a hybrid approach combining heuristic methods and machine learning can significantly improve the performance and reliability of task offloading and resource allocation in edge computing systems designed for real-time IoT workloads.

Literature Review

A priority-based resource allocation method was developed to improve efficiency in mobile-edge computing. Tasks are classified by urgency, and resources are dynamically assigned using adaptive scheduling. Critical tasks get processed first, while others wait based on availability. This approach improved task completion time and system performance, showing that intelligent, priority-aware scheduling enhances edge computing for real-time IoT needs.[1][1]

A new method was developed for sharing computing tasks and managing resources in mobile-edge computing systems within very densely packed networks. This approach works to lower the time it takes for the system to respond and reduce the amount of energy used by making smart decisions about which tasks to send to the edge servers and how to divide the available resources. It brings together how users connect to the network, how tasks are ordered, and what each edge server can handle into one complete system. Testing the system showed better performance, less waiting time, and more efficient use of energy. This shows that working together to manage task sharing and resource use is key to supporting fast and reliable applications in tightly packed edge environments.[2][3]

A method was introduced to handle tasks more efficiently in mobile edge computing networks by considering delays. This approach decides in real time whether to process tasks on the device itself or send them to nearby edge servers, depending on the current network and system conditions. By focusing on fast communication and efficient processing, the method cut down response times and made the system more reliable. The study highlights how smart decision-making can boost the performance of apps that are sensitive to delays in edge computing setups.[3][7]

An energy-efficient computation offloading strategy was introduced for vehicular edge cloud computing to reduce energy consumption while maintaining service quality. The approach considers vehicle mobility, task deadlines, and resource constraints to decide whether and where to offload computational tasks. By optimizing both energy use and offloading decisions, the method achieved lower energy costs and improved execution efficiency. This highlights the significance of mobility-aware and energy-conscious offloading in supporting sustainable and responsive vehicular edge environments.[4][9]

A fairness-aware task offloading and resource allocation strategy was proposed for cooperative mobile-edge computing environments. The approach ensures equitable distribution of computing resources among users while optimizing task performance. It introduces fairness constraints into the optimization model to prevent resource monopolization and balance task delay. Experimental results showed that the method effectively reduced task completion time while maintaining fairness across users. This demonstrates the importance of integrating fairness into edge computing to support inclusive and efficient service delivery.[5][12]

Methodology

This study used a **simulation-based methodology** in three phases:

1. System Setup

- A simulated environment with 20 IoT devices (e.g., cameras, sensors) and 4 edge nodes was created using the iFogSim toolkit.
- Tasks were categorized as latency-critical (e.g., video processing) and non-latency-critical (e.g., periodic data logging).

2. Proposed Framework

- The hybrid framework includes:

  - **AI-based task load predictor** using a lightweight neural network model.
  - **Heuristic scheduler** using Earliest Deadline First (EDF) and a queue management system.
- The model decides whether a task should be executed locally, offloaded to another edge node, or sent to the cloud.

3. Evaluation Metrics

- Task response time
- CPU/memory utilization
- Energy consumption
- Task failure/drop rate

Limitations

- The simulation may not perfectly capture real-world wireless network behavior.
- Resource consumption of the AI model itself is not deeply profiled.

Results

The proposed hybrid framework showed significant performance improvements across multiple parameters when compared with baseline methods:

Metric	**Static Offloading**	**Cloud-Only**	**Proposed Hybrid**
Avg. Task Response Time	380 ms	540 ms	**220 ms**
Energy Consumption (avg)	65%	82%	**48%**
Task Failure Rate	12.6%	7.8%	**3.4%**
CPU Utilization (avg)	85%	92%	**68%**

The **hybrid approach** was particularly effective in managing real-time, high-load scenarios such as video surveillance and emergency alerts.

Discussion

The results support the thesis that combining heuristic methods with AI prediction enhances task offloading decisions in real-time edge computing environments. Compared to static methods, the hybrid model adapts to fluctuating network conditions and task demands more efficiently.

This work aligns with recent literature advocating adaptive and context-aware edge frameworks, but stands out by offering a lightweight solution that does not overburden resource-constrained devices. It also demonstrates how predictive intelligence can reduce both latency and energy consumption, critical factors in mobile and industrial IoT use cases.

One implication is that AI at the edge—if properly optimized—can make edge computing both scalable and sustainable. However, further research is needed to test these models in real-world distributed systems with actual device variability and network instability.

Conclusion

This paper presented a hybrid framework that optimizes resource allocation and task offloading in edge computing for real-time IoT applications. By leveraging AI-based prediction and heuristic scheduling, the model significantly reduced response time, energy usage, and task failures. These improvements support the growing need for intelligent, decentralized computing infrastructure as IoT applications expand.

Future work will explore federated learning for on-device model updates, real-world edge hardware testing, and enhancing security in task migration

References

[1] Zubair Sharif;Low Tang Jung;Imran Razzak;Mamoun Alazab, Adaptive and Priority-Based Resource Allocation for Efficient Resources Utilization in Mobile-Edge Computing, Published in: IEEE Internet of Things Journal (Volume: 10, Issue: 4, 15 February 2023), Page(s): 3079 – 3093, Date of Publication: 10 September 2021, DOI: 10.1109/JIOT.2021.3111838

[2] Chen Zhang;Hongwei Du, DMORA: Decentralized Multi-SP Online Resource Allocation Scheme for Mobile Edge Computing, Published in: IEEE Transactions on Cloud Computing (Volume: 10, Issue: 4, 01 Oct.-Dec. 2022), Page(s): 2497 – 2507, Date of Publication: 15 December 2020, DOI: 10.1109/TCC.2020.3044852

[3] Ya Gao;Haoran Zhang;Fei Yu;Yujie Xia;Yongpeng Shi, Joint Computation Offloading and Resource Allocation for Mobile-Edge Computing Assisted Ultra-Dense Networks, Published in: Journal of Communications and Information Networks (Volume: 7, Issue: 1, March 2022), Page(s): 96 – 106, Date of Publication: 30 March 2022, DOI: 10.23919/JCIN.2022.9745485

[4] Houming Qiu;Kun Zhu;Nguyen Cong Luong;Changyan Yi;Dusit Niyato;Dong In Kim, Applications of Auction and Mechanism Design in Edge Computing: A Survey, Published in: IEEE Transactions on Cognitive Communications and Networking (Volume: 8, Issue: 2, June 2022), Page(s): 1034 – 1058, Date of Publication: 28 January 2022, DOI: 10.1109/TCCN.2022.3147196

[5] Shiheng Ma;Song Guo;Kun Wang;Weijia Jia;Minyi Guo, A Cyclic Game for Service-Oriented Resource Allocation in Edge Computing, Published in: IEEE Transactions on Services Computing (Volume: 13, Issue: 4, 01 July-Aug. 2020), Page(s): 723 – 734, Date of Publication: 15 January 2020, DOI: 10.1109/TSC.2020.2966196

[6] Tong Liu;Shenggang Ni;Xiaoqiang Li;Yanmin Zhu;Linghe Kong;Yuanyuan Yang, Deep Reinforcement Learning Based Approach for Online Service Placement and Computation Resource Allocation in Edge Computing, Published in: IEEE Transactions on Mobile Computing (Volume: 22, Issue: 7, 01 July 2023), Page(s): 3870 – 3881, Date of Publication: 04 February 2022, DOI: 10.1109/TMC.2022.3148254

[7] Wei Feng;Hao Liu;Yingbiao Yao;Diqiu Cao;Mingxiong Zhao, Latency-Aware Offloading for Mobile Edge Computing Networks, Published in: IEEE Communications Letters (Volume: 25, Issue: 8, August 2021), Page(s): 2673 – 2677, Date of Publication: 21 April 2021, DOI: 10.1109/LCOMM.2021.3074621

[8] Chun-Yen Lee;Chia-Hung Lin;Zhan-Lun Chang;Chih-Yu Wang;Hung-Yu Wei, Joint Resource Allocation and Intrusion Prevention System Deployment for Edge Computing, Published in: IEEE Transactions on Services Computing (Volume: 17, Issue: 5, Sept.-Oct. 2024), Page(s): 2502 – 2515, Date of Publication: 09 August 2024, DOI: 10.1109/TSC.2024.3441313

[9] Xin Li;Yifan Dang;Mohammad Aazam;Xia Peng;Tefang Chen;Chunyang Chen, Energy-Efficient Computation Offloading in Vehicular Edge Cloud Computing, Published in: IEEE Access (Volume: 8), Page(s): 37632 – 37644, Date of Publication: 20 February 2020, DOI: 10.1109/ACCESS.2020.2975310

[10] Qi Zhang;Lin Gui;Fen Hou;Jiacheng Chen;Shichao Zhu;Feng Tian, Dynamic Task Offloading and Resource Allocation for Mobile-Edge Computing in Dense Cloud RAN, Published in: IEEE Internet of Things Journal (Volume: 7, Issue: 4, April 2020), Page(s): 3282 – 3299, Date of Publication: 17 January 2020, DOI: 10.1109/JIOT.2020.2967502

[11] Haitao Xu;Wentao Huang;Yunhui Zhou;Dongmei Yang;Ming Li;Zhu Han, Edge Computing Resource Allocation for Unmanned Aerial Vehicle Assisted Mobile Network With Blockchain Applications, Published in: IEEE Transactions on Wireless Communications (Volume: 20, Issue: 5, May 2021), Page(s): 3107 – 3121, Date of Publication: 08 January 2021, DOI: 10.1109/TWC.2020.3047496

[12] Jiayun Zhou;Xinglin Zhang, Fairness-Aware Task Offloading and Resource Allocation in Cooperative Mobile-Edge Computing, Published in: IEEE Internet of Things Journal (Volume: 9, Issue: 5, 01 March 2022), Page(s): 3812 – 3824, Date of Publication: 26 July 2021, DOI: 10.1109/JIOT.2021.3100253

[13] Xinliang Wei;Xitong Gao;Kejiang Ye;Cheng-Zhong Xu;Yu Wang, A Quantum Reinforcement Learning Approach for Joint Resource Allocation and Task Offloading in Mobile Edge Computing, Published in: IEEE Transactions on Mobile Computing (Volume: 24, Issue: 4, April 2025), Page(s): 2580 – 2593, Date of Publication: 13 November 2024, DOI: 10.1109/TMC.2024.3496918

[14] A. Sasikumar;Logesh Ravi;Malathi Devarajan;Subramaniyaswamy Vairavasundaram;A. Selvalakshmi;Ketan Kotecha;Ajith Abraham, A Decentralized Resource Allocation in Edge Computing for Secure IoT Environments, Published in: IEEE Access (Volume: 11), Page(s): 117177 – 117189, Date of Publication: 16 October 2023, DOI: 10.1109/ACCESS.2023.3325056

[15] Sai Wang;Xiaoyang Li;Yi Gong, Energy-Efficient Task Offloading and Resource Allocation for Delay-Constrained Edge-Cloud Computing Networks, Published in: IEEE Transactions on Green Communications and Networking (Volume: 8, Issue: 1, March 2024), Page(s): 514 – 524, Date of Publication: 17 August 2023, DOI: 10.1109/TGCN.2023.3306002