# EFFICIENT RESOURCE ALLOCATION FOR IOT APPLICATIONS IN MOBILE EDGE COMPUTING VIA DYNAMIC REQUEST SCHEDULING OPTIMIZATION

**Yusupov Jasurbek Shuxrat o'g'li[1] Parthasarathi Murugesan[2]**
**1.Master of MScIT , 2. Assoc. Professor & Dean- BS, Sambhram University Jizzakh - Uzbekistan**

## Annotation

This study investigates the challenges of task scheduling and resource allocation in an ultra-dense edge cloud (UDEC) network incorporating micro (macro) base stations and different equipment users (UE) under 5G technology. By utilizing multiple access interference (MAI), we enable multiple access capabilities to the UEs under the Non-Orthogonal Multiple Access (NOMA) protocol. To accommodate the dynamic nature of user tasks, we propose a two-level scheduling framework. At the upper level, the non-convex, non-linear power allocation problem for mobile users is formulated and solved using the Inertia Weight Particle Swarm Optimization (IW-PSO) algorithm to minimize the transmission energy consumption. Simultaneously, at the lower level, the joint task offloading and resource allocation problem is formulated as MINP and solved with the binary Particle Swarm Optimization (PSO) algorithm to find an optimal schedule that maximizes the response rate and the overall welfare of the system while minimizing the energy consumption. In this study, various sensitivity analyses are conducted for different parameters, including the number of mobile users, channel power, and request profile parameters, including workload and request size. Comparative analyses with other methods from recent literature consistently reveal the comparable performance of the proposed framework and confirm its effectiveness in addressing the challenges of dynamic task scheduling.

## Introduction

The rapid growth of mobile devices connecting to wireless networks has been driven by the proliferation of cloud computing and advances in wireless communication technology (Pang et al., 2024). In the domains of mobile computing and the Internet of Things (IoT), mobile devices have become integral to modern life (Li et al., 2024). The integration of a wide array of sensors with wireless communication technology allows IoT to connect mobile devices to the Internet through various networks (Wang et al., 2024). Furthermore, user equipment (UE) is becoming

increasingly intelligent, with applications extending to areas such as smart cell phones (Suryani et al., 2024), smart vehicle fleets (Quyet & Kha, 2024), unmanned aerial vehicles (Dai et al., 2021), and augmented/virtual reality (AR/VR) (Paramasivam et al., 2024). While these emerging applications and services are advancing, they are also constrained by the limited computing capacity and battery life of end devices. Simultaneously, the proliferation of diverse mobile applications has generated a significant demand for enhanced computing capacity and real-time processing capabilities. For example, most portable AR devices enhance the real-world experience by using camera data to capture the environment and identify surfaces (Dho et al., 2024). However, their limited computing resources restrict their ability to recognize and distinguish complex objects in the environment. The existing cloud computing model is inadequate for performing real-time, location-aware computing that supports mobility (Liu et al., 2022). To address these challenges, researchers are developing the concept of Mobile Edge Computing (MEC), which implements highly localized cloud services at the network edge, near the UEs (Tran & Pompili, 2018). Mobile operators are adopting a strategy that combines computing, network, and storage resources with the base station (BS) through the implementation of the MEC platform. As a novel and emerging paradigm, MEC leverages the abundant, unused computing power and storage distributed at the network's edges. MEC servers provide access to radio network information and user traffic, enabling service providers to dynamically adjust their offerings and improve user satisfaction (Jararweh et al., 2016, Liu et al., 2021, Liu et al., 2021). This approach has the potential to provide extensive capacity for the execution of compute-intensive and latency-critical tasks directly at UEs associated with offloading and thus overcoming key challenges in realizing the 5G vision (Hu et al., 2015, Mao et al., 2017). It is noteworthy to mention, due to the limited computing resources, MEC servers can outsource resource-intensive tasks to Mobile Cloud Computing (MCC) servers via the Internet when overloaded. In addition, several MEC servers can cooperate and transfer tasks to each other via different networks, such as the backhaul network. This improves services for mobile users by maintaining balance between workloads and resource allocation (Tan et al., 2017). With this opening, we delve into recent literature and examine the technology and management models presented. At times, we refer to specific points that serve as a stimulus for elucidating our proposed framework.

In this framework, According to CISCO's annual internet report (Cisco Annual Internet Report - Cisco Annual Internet Report, 2024) White Paper − Cisco, n.d.), mobile devices connections are growing from a value of around 8 billion in 2018 to over 13 billion in 2023, with 1.4 billion (equivalent to 10 % of global values in 2023) of these devices being 5G-enabled. This figure is also expected to multifold over the next five years. This growth in the connection will increase data traffic on the network and leads to access barriers to the cloud edge for the mobile users and benefit from their computing facilities. Hence, the key challenge lies in efficiently scheduling the available computing capacities of MEC to meet the dynamic contradictory demands from various mobile users/UEs.

The efficient utilization of MEC resources is intricately related to solving task/request offloading (TO/RO), resource allocation, and task scheduling problems. The authors in (Alameddine et al., 2019) mutually computed the allocated resources for the hosted task. They framed this problem as a dynamic task offloading (TO) and scheduling problem and solve it by considering the three subproblems mentioned above. Their goal was to fulfil the quality of experience (QoE) for shared resource allocation for IoT applications via comprehensive offloading scheme. However, this study overlooks the limited computational capacity of the edge can lead to additional energy consumption and latency, especially during time intervals of congestion. This can have an immediate influence on the number of tasks delegated to the network. Similarly, the authors in (Chen & Hao, 2018) used employed three distributions to assess the influence of data with different sizes and features size on computing offloading. However, they neglected the computing capacity of the cloud edge. The authors in (Tran & Pompili, 2019) addressed the problem of JTORA for a system that uses Orthogonal Frequency Division Multiple Access (OFDMA) (Dahlman et al., 2014) as a multiple access method on the uplink side. To maintain

orthogonality between users sharing the same base station (BS), the operating frequency band was partitioned into sub-bands and each user was allocated to a specific sub-band. As a result, the mobile users are unable to fully utilize the system-wide bandwidth, which can lead to network gridlock and augmented energy consumption of the devices. In (Mao et al., 2016), the authors developed an effective offloading strategy of JTORA for a MEC system that harnesses renewable energy sources to power mobile devices. This particular system is compact in scale and overseen by a telecom operator that displays a system with a single BS. In (Kiani, 2017), the authors introduced a novel hierarchical model comprising a dual- time scale strategy. In this model, computational resources were allocated based on an auction to maximize profit, followed by the allocation of communication resources to meet users' QoS requirements. However, they neglected the computing association and assumed that the tasks of the mobile users were only offloaded to the corresponding edge cloud.

UDN is a new paradigm whose basic idea is to bring 5G mobile users close to the BSs (Chen et al., 2016). In (Kamel et al., 2016, Yu et al., 2016), the authors have provided a comprehensive review of recent achievements and findings in various research directions such as interference management, propagation, resource allocation, energy efficiency, modeling techniques and performance metrics. In (Chopra et al., 2017) and (Salahdine et al., 2021), the authors respectively perform an overview of security challenges and possible solutions as well as sleep mode techniques tailored for UDN in the 5G era and beyond. In (López-Pérez et al., 2015), the authors aimed to promote a common understanding and analyzing the potential benefits of network densification, the application of higher frequency bands and techniques to improve spectral efficiency. Their focus was on evaluating the impact of these factors on the capabilities of small cells in outdoor scenarios. In (B. Yu et al., 2018), the authors presented a breakthrough mobile edge computing framework designed to minimize energy consumption of the entire network, encompassing IoT devices and small BSs. This framework incorporates the joint problem of TO, user mapping, and small BS sleeping, and considers practical constraints. The dynamic and diverse nature of user requests leads to uncertainties in the traffic demands of individual users in UDN. Given this uncertainty, the authors in (Ma et al., 2018) presented a robust optimization-based energy-efficient user allocation scheme. This scheme not only strengthens the stability of associations, but also improves the energy efficiency of the entire system by addressing the uncertain and varying traffic demands of users. In an UDN, the mobile network operator deploys multiple 5G stations categorized as macro and micro BSs to ensure comprehensive coverage of the 5G network. The macro base station, which is connected to a resource-rich deep cloud via the internet, mainly provides outdoor coverage. The micro base station with lower transmission power, on the other hand, is connected to the edge cloud through a local network, predominantly used for indoor scenarios (Hu and Li, 2020, Shi et al., 2022). Owing to the advantages of UDN in terms of high wireless connection capacity for mobile users, we incorporated the UDN architecture within the MEC network.

In a UDN, mobile devices have the ability to transfer computing requests to the macro-BSs when micro-BSs are not able to handle the entirety of the transferred requests. A major drawback in UDN is the dependency on micro-BS that provide network access services and are each equipped with only one cloud edge. Consequently, their computing power is much more intensive compared to a macro station. Therefore, a huge volume of offload requests can lead to increased response latency due to limited computing resources in the edge cloud (Hu & Li, 2020). It therefore makes sense to investigate a suitable two-way offloading system. The Non-Orthogonal Multiple Access (NOMA) protocol, which is considered a promising radio access technology for 5G systems, enables all mobile devices to exploit the entire system bandwidth for data transmission simultaneously (Dai et al., 2015, Ding et al., 2015). However, it faces challenges related to multiple access interference (MAI) (Islam et al., 2017).

In recent years, the challenge of task offloading in a UDEC network has received a lot of attention. Mobile edge computing in ultra- dense networks is a compelling solution to meet the requirements of low-latency computation- and data-intensive tasks. In (Chen & Hao, 2018), the

authors promoted software-defined networks to address the task offloading problem in UDEC networks with the goal of minimizing delays and preserving battery life of mobile devices. They first formulated the task offloading problem as a mixed-integer nonlinear program and then solved it by decomposing it into subproblems of task and resource allocation. With the introduction of multi-access MEC, the authors in (Guo, Liu, et al., 2018) addressed the challenge posed by computation- intensive applications and resource limitations of IoT mobile devices in UDN and solved the offloading problem by a heuristic greedy offloading scheme. In (Yang et al., 2017), the authors addressed the problems arising from the heterogeneity of servers and the agglomeration of users to achieve fair resource allocation and load balancing among MEC base stations (MEC-BSs). The aim was to minimize the time required for all tasks. The authors of (Guo, Zhang, et al., 2018) identified a gap in current UDEC research on task offloading. They emphasized that the focus is on simple tasks without considering dynamic characteristics such as mobile users' task types and arrival times at edge servers. Consequently, they proposed a solution that addresses joint multi-task offloading and transmission power allocation in the UDEC network using a game-theoretic approach. In this study, we considered the dynamic behavior of mobile devices. Toward this end, we solved the scheduling problem of mobile users' dynamic tasks taking into account joint task offloading, computation resource allocation, and transition power optimization at the edge cloud.

Based on our literature review and identification of specific research gaps, we proposed a framework for dynamic request planning in MEC. This framework is specifically tailored for IoT applications implemented in the context of an UDN. In our study, we assumed that all devices within the UDCE network under the 5G architecture, including smart EVs, smartphones, surveillance systems, and unmanned aerial vehicles, function as mobile users. The macro-BS ensures coverage for all mobile users and micro-BSs. Each station in a specific zone extends coverage to mobile users located in that specific zone. We examined a scenario in which mobile users are allowed to transfer their computing tasks to both the associated micro-BS and the macro-BS. We assume that mobile users can employ MAI to offload to BSs under the NOMA protocol, in the uplink channel. Consequently, we present an algorithm for power allocation between mobile users in the uplink channel that aims to minimize the energy consumption associated with data transmission.

Realizing the full advantage of the planning framework, necessitates careful consideration of the following points. First, the power allocation problem proves challenging since the uplink offloading rate (UOR) is affected by the interference between different mobile users, rendering the problem non-convex. Second, the computation resource allocation strategy of a micro-BS is affected by the mobile user's decision on the offloading destination. Thirdly, the resource allocation strategy is influenced by the diversity of the demand profile of mobile users, along with the offload requirement.

**The main contributions and novelties of this paper can be summarized as follows:**

The study addresses the complexities of task scheduling and resource allocation in an ultra-dense edge cloud (UDEC) network. This network integrates micro and macro base stations to support diverse equipment users (UE) within the framework of 5G technology.

By leveraging the Non-Orthogonal Multiple Access (NOMA) protocol and managing multiple access interference (MAI), the research enhances the efficiency of communication among UEs. This approach optimizes spectral efficiency and capacity utilization in the UDEC environment.

A novel two-level scheduling framework is proposed to manage the dynamic nature of user tasks effectively. At the upper level, the study formulates and solves a non-convex, non-linear power allocation problem using the IW-PSO algorithm. This optimization minimizes transmission energy consumption for mobile users.

At the lower level of the framework, the research addresses the joint task offloading and resource allocation problem. Formulated as a MINLP (Mixed-Integer Nonlinear Programming) problem, this stage employs the binary PSO algorithm. The objective is to optimize task scheduling to maximize response rates and overall system welfare while minimizing energy consumption.

The paper conducts comprehensive sensitivity analyses on various parameters critical to UDEC network performance. Parameters include the number of mobile users, channel power, and workload profiles. These analyses provide insights into the robustness and adaptability of the proposed framework under different operational conditions.

Finally, the paper contributes to advancing the field by introducing a sophisticated framework for optimizing task scheduling and resource allocation in UDEC networks under 5G technology. The integration of NOMA, two-level scheduling, and advanced optimization algorithms underscores its potential to enhance network efficiency, responsiveness, and energy utilization in dynamic and densely populated environments.

The subsequent sections are structured as follows: Section 2 provides a thorough exploration of the model description and problem formulation. We begin with a detailed consideration of the problem and then formulate the planning problem. Following that, in Section 3, we present a two-level hierarchical efficient optimization framework, fostering the Inertia Weight Particle Swarm Optimization (IW-PSO) and Binary Multi-objective PSO (BMO-PSO) and apply it to solve the problem. Section 4 presents the numerical results and discusses the experiments, including a comparison. Finally, in Section 5, we draw conclusions for the paper.

Section snippets

System modeling

We considered a system on base of UDEC network as illustrated in Fig. 1. In this network configuration, each BS is equipped with a MEC server. These MEC servers, whether physical or virtual, contain computing resources provided by the network operator. They provide compute offload services specifically designed for resource-constrained mobile users such as smartphones, tablets and portable devices. The MEC servers communicate wirelessly with the mobile devices via the channels provided by the

**Two-level hierarchical frame work for solving JTORA problem**

In this study, a two-level hierarchical framework is introduced to optimize the scheduling of dynamic requests from mobile users to efficiently allocating computation resources in edge clouds. The proposed model incorporates various factors, including delay time, transmission energy, transmission time, system utility function (QoE for mobile users considering transition delay), node price of the edge, additional cost of transmission to the uplink channel, and welfare function of the system. As

**Comparative analysis and simulation configurations**

To assess the efficacy of the proposed two-stage hierarchical optimization framework based on IW-PSO and BPSO, the numerical results of the simulations were compared to the methods documented in the literature, namely:

- Simulated Annealing (SA), a classical heuristic algorithm known for its efficiency in solving NP-hard problems, was specifically selected as the benchmark method to solve PA problem (SAPA).
- Non-cooperative game model based on sub-gradient (NCGG) which designed by Hu et al.

**Conclusion**

To summarize, our study delves into the intricate challenges of task scheduling and resource allocation in a UDEC network that integrates micro- (macro-) BSs and different UEs in the 5G technology landscape. Employing NOMA and a two-level scheduling framework, we optimize the performance of the system's performance by addressing non-convex, non-linear problems at both the user and system levels. The IW-PSO and BPSO algorithms play a central role in minimizing energy consumption and maximizing

**References**

1. G. Chopra et al. A survey on ultra-dense network and emerging technologies: Security challenges and possible solutions Journal of Network and Computer Applications (2017)

2. J. Liu et al.Deeply learning a discriminative spatial–temporal feature for robot action understanding Future Generation Computer Systems (2021)

3. J. Liu et al. Resource allocation and scheduling in the intelligent edge computing context Future Generation Computer Systems (2021)

4. F. Salahdine et al. A survey on sleep mode techniques for ultra-dense networks in 5G and beyond Computer Networks (2021)

5. S.D. SeyedGarmroudi et al. Improved Pelican optimization algorithm for solving load dispatch problems Energy (2024)

6. Z. Wang et al. Deep Reinforcement Learning-based scheduling for optimizing system load and response time in edge and fog computing environments Future Generation Computer Systems (2024)

7. H.A. Alameddine et al. Dynamic task offloading and scheduling for low-latency IoT services in multi-access edge computing IEEE Journal on Selected Areas in Communications (2019)

8. T. Bahreini et al. An envy-free auction mechanism for resource allocation in edge computing systems Proceedings - 2018 3rd ACM/IEEE Symposium on Edge Computing SEC (2018)

9. M. Chen et al.Task offloading for mobile edge computing in software defined ultra-dense network IEEE Journal on Selected Areas in Communications (2018)

10. S. Chen et al. User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions IEEE Wireless Communications (2016)