

# NOMA-Assisted Multi-BS MEC Networks for Delay-Sensitive and Computation-Intensive IoT Applications

Yuang Chen, *Graduate Student Member*, Fengqian Guo, Chang Wu, Mingyu Peng, *Graduate Student Member*, Hancheng Lu, *Senior Member, IEEE*, and Chang Wen Chen, *Life Fellow, IEEE*

**Abstract**—The burgeoning and ubiquitous deployment of the Internet of Things (IoT) landscape struggles with ultra-low latency demands for computation-intensive tasks in massive connectivity scenarios. In this paper, we propose an innovative uplink non-orthogonal multiple access (NOMA)-assisted multi-base station (BS) mobile edge computing (BS-MEC) network tailored for massive IoT connectivity. To fulfill the quality-of-service (QoS) requirements of delay-sensitive and computation-intensive IoT applications, we formulate a joint task offloading, user grouping, and power allocation optimization problem with the overarching objective of minimizing the system’s total delay, aiming to address issues of unbalanced subchannel access, inter-group interference, computational load disparities, and device heterogeneity. To effectively tackle this problem, we first reformulate task offloading and user grouping into a non-cooperative game model and propose an exact potential game-based joint decision-making (EPG-JDM) algorithm, which dynamically selects optimal task offloading and subchannel access decisions for each IoT device based on its channel conditions, thereby achieving the Nash Equilibrium. Then, we propose a majorization-minimization (MM)-based power allocation algorithm, which transforms the original subproblem into a tractable convex optimization paradigm. Extensive simulation experiments demonstrate that our proposed EPG-JDM algorithm significantly outperforms state-of-the-art decision-making algorithms and classic heuristic algorithms, yielding performance improvements of up to 19.3% and 14.7% in terms of total delay and power consumption, respectively.

**Index Terms**—Non-orthogonal Multiple Access (NOMA); Internet of Things (IoT); Task Offloading; Game Theory; Mobile Edge Computing (MEC).

## I. INTRODUCTION

WITH the ubiquitous deployment of fifth-generation (5G) mobile communication technology worldwide, the Internet-of-Things (IoT) has entered a stage of large-scale commercialization, encompassing emerging applications such as intelligent housing, wearable devices, urban infrastructure, and smartphones [1–5]. Remarkably, over 127 new IoT devices are connected to the Internet every second, and it is estimated that by 2030, the number of connected IoT devices will reach to nearly 125 billion [4]. According to the forecast by

This work was supported in part by the National Science Foundation of China under Grant U21A20452, in part by Hong Kong Research Grants Council (GRF-15213322, GRF-15229423). Yuang Chen, Fengqian Guo, Chang Wu, Mingyu Peng, and Hancheng Lu are with the Laboratory of Future Networks, University of Science and Technology of China (USTC), Hefei, P. R. China (e-mail: {yuangchen21, changwu, mypeng}@mail.ustc.edu.cn; fqguo@ustc.edu.cn; hclu@ustc.edu.cn). Yuang Chen and Hancheng Lu are also with Deep Space Exploration Laboratory, Hefei 230088, China. Chang Wen Chen is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: changwen.chen@polyu.edu.hk).

the International Data Company (IDC), the amount of data generated by IoT devices has exceeded 80ZB in 2025 [6]. However, identifying and orchestrating such a massive number of heterogeneous IoT devices has become an increasingly challenging task [5, 7, 8]. The rapid proliferation of active IoT terminal devices globally has resulted in an exponential increase in data traffic. The generation of such massive data volumes introduces a surge in computing tasks, which are typically delay-sensitive and computation-intensive (DSCI) characteristics in real-world applications [9–11]. However, due to constraints in hardware and battery capacity, most IoT devices pose very limited computing power, making it difficult to effectively meet the demands of users for DSCI tasks [2, 7]. As a result, ensuring the smooth execution of IoT networks with DSCI tasks has become an urgent issue to be addressed.

Generally, IoT devices necessitate wirelessly offloading the DSCI tasks requested by users to resource-rich servers for further processing [12]. Traditional cloud computing paradigms address this by offloading DSCI tasks to the cloud center [7, 13, 14]. However, due to the dispersed distribution of massive IoT access, centralized computing paradigms suffer from low transmission efficiency [12]. Moreover, the centralized deployment of DSCI tasks in cloud centers leads to significantly increased cloud workloads, which not only reduces system resource utilization but also results in higher network delays. In this context, mobile edge computing (MEC) paradigms have gained widespread adoption, aiming to migrate computing and caching resources from remote cloud data centers to the network edge, closer to users and data sources [7, 13, 14]. This physical spatial proximity adaptation has significant advantages over geographically dispersed IoT devices in reducing transmission delay, enhancing network reliability, and reducing network bandwidth pressure. Although offloading DSCI tasks to MEC servers can effectively enhance quality-of-service (QoS) provisioning performance, the surge of large-scale IoT access coupled with high-bandwidth traffic generated by applications like Internet-of-Video Things (IoVT), presents a formidable challenge [15, 16]. These services now constitute a significant portion of network traffic and impose stringent requirements on delay and reliability [15, 16].

## A. Research Motivations and Challenges

The development of IoT-enabled MEC networks with DSCI tasks hinges on tailored effective edge offloading and wireless transmission schemes. While extensive studies have established

a solid foundation for facilitating massive IoT connectivity, some significant practical challenges remain, as follows:

**(1) Existing multi-access technologies typically neglect the interdependency between task offloading and user grouping.** As IoT devices with DSCI tasks connecting to MEC networks, the uplink transmission of high-volume DSCI tasks to servers inevitably encounters spectrum resource bottlenecks [17, 18]. Traditional orthogonal multiple access (OMA) technologies struggle to meet the offloading and delay requirements of DSCI tasks in resource-constrained networks. Non-orthogonal multiple access (NOMA) has presented the potential to enhance system spectral efficiency (SE) and gained significant attention by employing superposition encoding to combine signals to access the same resource block in the power domain or code domain [15, 19, 20]. At the receiver, NOMA utilizes successive interference cancellation (SIC) to decode the superimposed signal for efficient multi-device communications [15, 19, 20].

However, existing studies have numerous limitations. On the one hand, the user grouping schemes of most NOMA systems are based on their channel qualities, where the number of users accessing each sub-channel is typically considered as identical [19, 21–23]. Such user grouping strategies fail to adequately consider the interdependency between task offloading and user grouping, as well as the impacts of interference, power consumption, and delay requirements on strategy selection. On the other hand, NOMA systems relying on SIC technology may cause serious inter-device interference when transmitting data from different devices simultaneously [15, 19–22]. Moreover, the user grouping schemes in existing NOMA systems rarely account for the heterogeneity of IoT devices, such as different delay demands for DSCI tasks [15, 19–22].

**(2) IoT access with DSCI tasks in multi-BS-MEC systems significantly exacerbates the complexity of task offloading and resource allocation.** The MEC networks supporting IoT connectivity with DSCI tasks generate huge volumes of transmission tasks and offloaded data [7, 9], imposing heavy pressure on communication and computation resources [17, 24, 25]. Due to the limited transmission bandwidth and computing capacity, single-BS architectures are usually unable to fulfill the QoS requirements of DSCI tasks [7], which are also closely related to the number of subchannel resources and devices [23]. As a result, single-BS architectures cannot flexibly handle diverse task offloading demands, leading to poor resource utilization and system performance.

Although the multi-BS architectures promise to alleviate single-point congestion by distributing the computational load, they introduce new challenges in dynamic wireless networks. On the one hand, the complexity of task offloading and user grouping becomes significantly aggravated. The multi-BS-MEC networks not only need to determine the optimal user grouping decision among multiple sub-channels for uplink transmission, but also select the optimal offloading target MEC server [7, 15, 23], leading to intractable coupling and complexity among user grouping, task offloading, and resource allocation problems. On the other hand, inter-BS interference and coordination control issues are also prominent under NOMA mechanisms [15, 26], as spectrum resource sharing among

multiple BSs easily causes severe co-channel interference. Moreover, the lack of effective coordinated scheduling among multiple BSs leads to reduced resource utilization and uneven delays. Finally, performance optimization in multi-BS-MEC systems typically involves multi-dimensional variables, including BS selection, power allocation, task offloading, and user grouping, which result in an extensive and computationally complex search space for optimal solutions.

### B. Main Contributions

In order to effectively overcome the aforementioned challenges, this paper proposes an uplink NOMA-assisted multi-BS-MEC (NMBM) network designed to facilitate the DSCI IoT applications. The proposed NMBM network comprehensively accounts for uneven device access across sub-channels, inter-group interference, imbalanced computational loads, and heterogeneous delay constraints among devices. We formulate a joint optimization problem that integrates task offloading, user grouping, and power allocation, with the objective of minimizing the total delay of the proposed network. To efficiently tackle this NP-hard problem that involves the coupling between discrete task offloading and user grouping and continuous power allocation decision variables, we first adopt a variable decomposition strategy to divide the original problem into two subproblems: (1) the joint task offloading and user grouping subproblem, and (2) the power allocation subproblem. Corresponding optimization algorithms are then developed to solve each subproblem efficiently. The primary contributions of this work are summarized as follows:

- To support the IoT applications with DSCI tasks access to MEC networks, we propose an uplink NMBM network, which considers user grouping for IoT devices with non-uniform access on sub-channels, and takes into account inter-group interference, imbalanced computational load, and heterogeneous delay among devices. We formulate a joint optimization problem that accounts for task offloading, user grouping, and power allocation, with the goal of minimizing the system's total delay.
- For the subproblem of joint task offloading and user grouping, inspired by game theory, we reformulate it into a non-cooperative game model and propose an exact potential game-based joint decision-making (EPG-JDM) algorithm. Under the proposed game framework, users dynamically select the optimal task offloading and sub-channel access decisions (i.e., user grouping decisions) based on their own channel conditions, taking into account the strategies already determined by other users.
- For the power allocation subproblem, we leverage the Majorization-Minimization (MM) algorithm to transform the power allocation problem into a tractable convex optimization problem. Furthermore, we propose a joint alternating optimization algorithm to iteratively address these two subproblems until the objective function converges.
- Extensive simulations demonstrate that the proposed algorithms significantly outperform the most representative existing baseline algorithms including Max-Min-based Grouping [27], Gale-Shapley Grouping [28], Nearby BS-

based Offloading [29], and Computing Capacity-based Offloading, and our proposed algorithms can achieve at least 19.3% and 14.7% performance improvements in terms of total delay and power consumption, respectively.

The remainder of this paper is organized as follows. Sec. II reviews related works. Sec. III introduces the proposed system model. Sec. IV presents the problem formulation and analysis. Sec. V provides the algorithm designs and solutions. Sec. VI provides extensive performance evaluations against the state-of-the-art comparison schemes. Finally, Sec. VII concludes the paper and explores future directions.

## II. RELATED WORKS

This section first reviews the current research status of NOMA-assisted MEC networks under IoT access with DSCI tasks [15, 16, 30–35]. Then, we provide a comprehensive survey of task offloading and resource allocation in multi-BS-MEC networks [7, 35–38]. Finally, we discuss key limitations of these related works that are closely aligned with our aforementioned challenges.

### A. Research Status of NOMA-assisted MEC Networks Under DCSI IoT Applications

Benefiting from the advantages of NOMA in enhancing SE and transmission throughput, NOMA-assisted MEC networks that facilitate the IoT access with DSCI tasks have garnered significant attention [15, 16, 30–35]. The authors in [30] proposed a two-way relaying NOMA-assisted IoT network, where the issue of the limited lifetime of IoT access points due to device energy constraints was thoroughly investigated. In [31], the authors introduced a jammer-aided covert reconfigurable intelligent surface (RIS)-NOMA system that exploits the NOMA characteristics to conceal the presence of strong signal transmissions to IoT receivers. To flexibly and accurately identify the usage of target frequencies, the authors in [32] investigated an adaptive NOMA-based spectrum sensing approach for uplink IoT access networks. In [33], RIS technology was introduced to enhance the offloading capability of MEC systems. To further improve task offloading capability, the authors proposed an RIS-NOMA-assisted MEC network that jointly optimizes channel allocation, beam-bandwidth allocation, offloading rates, and power control schemes. IoVT imposes higher demands on transmission and computing capabilities of wireless networks than traditional IoT systems [16]. In [15], we proposed a NOMA-assisted IoVT system combined with MEC and formulated a joint optimization problem for NOMA operations and MEC offloading to minimize the weighted average total system delay. In [34], the authors highlighted that existing downlink NOMA systems lead to redundant transmissions in resource-constrained satellite IoT networks; thus, they introduced the age-of-information (AoI) metric and proposed a content-aware sampling strategy to assess both the freshness and value of status updates. The authors in [35] explored the dynamic task offloading problem in NOMA-based IoT networks, which integrates task scheduling and computing resource allocation decisions to support massive IoT connectivity.

### B. Survey on Transmission and Computing Schemes in Multi-BS-MEC Architectures

IoT access with DSCI tasks led to a massive influx of data requiring both transmission and computational resources. To provide more powerful transmission and computation capacities, the multi-BS-MEC architectures have been extensively investigated [7, 35–38]. In [35], the authors investigated a NOMA-assisted MEC system, where a dynamic task offloading and resource allocation joint optimization problem was formulated to minimize the system's energy consumption. The authors in [36] proposed a RIS-assisted multi-MEC system, which can redirect the IoT devices to less-loaded MEC servers through passive beamforming techniques implemented on RISs when MEC servers are overwhelmed by massive computational tasks. In [7], we proposed an AoI-aware multi-BS-MEC real-time monitoring system to support large-scale industrial IoT access, where a joint task offloading and resource allocation optimization problem was formulated to ensure the information freshness of data packets. In [37], the authors proposed a decentralized offloading algorithm for massive IoT-accessed multi-BS scenarios to ensure the effective and fair distribution of BS resources. The authors in [38] developed a multi-BS collaborative edge computing model with heterogeneous computing capabilities, where a distributed training scenario was introduced to leverage the federated learning framework for collaborative optimization across multiple BSs while ensuring data privacy.

### C. Key Limitations of These Research Works

Although the aforementioned studies on NOMA-assisted multi-MEC networks designed for DSCI IoT access have made significant progress in improving SE, task offloading, and resource allocation [15, 16, 30–35], some key limitations remain. Specifically, these works typically overlook the close relationship between task offloading and user grouping, solely relying on channel quality or fixed device counts for user grouping, and fail to fully consider device heterogeneity (e.g., deadline constraints). This leads to increased interference among devices and makes it difficult to meet the QoS requirements of DSCI tasks. Furthermore, under the multi-BS-MEC architecture [7, 35–38], although distributed optimization [37] and federated learning [38] have been introduced to distribute the load, the system still faces challenges such as strong coupling decision complexity, inter-BS co-channel interference, and insufficient coordinated scheduling, leading to poor resource utilization and higher delay in dynamic wireless environments. These limitations highlight the need for more refined joint optimization frameworks to address the technical challenges posed by DSCI IoT access.

## III. SYSTEM MODEL

As illustrated in Fig. 1, we propose an uplink NMBM network for IoT applications with DSCI tasks, which consists of  $M$  BSs and  $N$  single-antenna equipped IoT devices. IoT devices upload their collected sensor data to the MEC servers for processing through uplink wireless transmission. There is an MEC server deployed near each BS to provide sufficient computing resources, and this combination of BSs

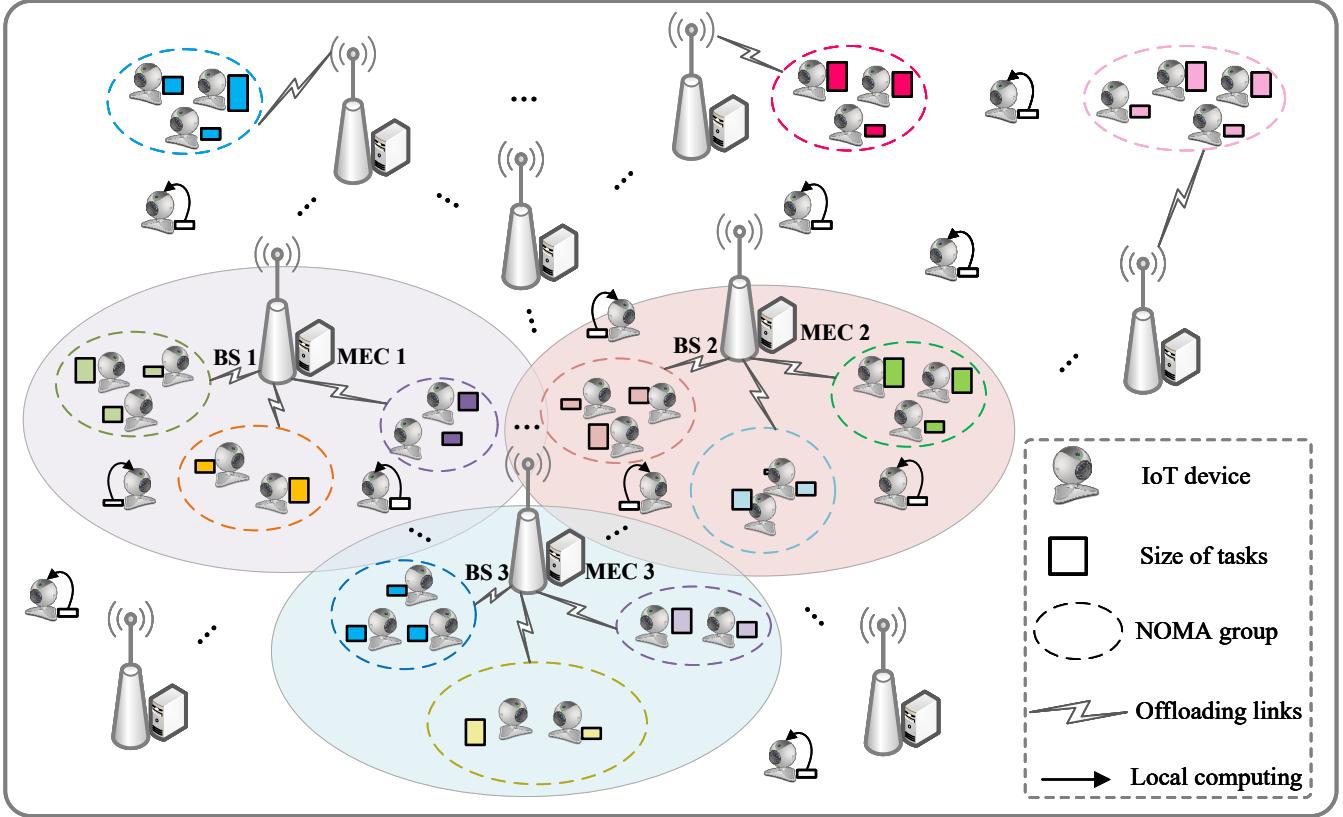


Fig. 1: The proposed Uplink NOMA-assisted multi-BS-MEC network for IoT access with DSCI tasks.

and MECs is referred to as BS-MEC. Each BS can occupy  $G$  subchannels, and orthogonal spectrum resources are allocated among BSs to avoid inter-BS interference [35, 37]. The sets of BS-MECs, IoT devices, and sub-channels can be denoted as  $m \in \mathcal{M} \triangleq \{1, 2, \dots, M\}$ ,  $n \in \mathcal{N} \triangleq \{1, 2, \dots, N\}$ , and  $g \in \mathcal{G} \triangleq \{1, 2, \dots, G\}$ , respectively. We use  $\|\mathcal{U}_{mg}\|$  to denote the set of IoT devices that offload their DSCI tasks to BS-MEC  $m$  through subchannel  $g$ . If the IoT device  $n$  offloads its DSCI tasks to BS-MEC  $m$  via subchannel  $g$ , then the channel coefficient between the IoT device  $n$  and BS-MEC  $m$  can be represented as  $h_n^{mg}$ . For all IoT devices that connect to the identical BS-MEC and jointly occupy the same subchannel, we leverage the NOMA technique to process their data transmission. As shown in Fig. 1, the height of the rectangle next to each IoT device is used to intuitively reflect the amount of their computational demands, also referred to as the amount of data they require to process. Notably, IoT devices in the proposed system model are heterogeneous, which means their requirements for computing resources and delay are different.

Under NOMA mechanisms, the system throughput is not only associated with the channel conditions of individual IoT devices but also significantly influenced by the channel conditions of other devices within the same group. At the receiver end, the decoding process must follow a certain priority order, where the signals of IoT devices with the best channel conditions are decoded first, while the signals of other IoT devices are treated as interference. Based on this principle, decoding can be carried out step by step until the signals from all IoT devices within the group are successfully decoded. This

decoding strategy ensures that the proposed network can fully utilize devices with better channel conditions to enhance the overall throughput. For the IoT device  $n$  that selects to offload its DSCI tasks to the MEC  $m$  via subchannel  $g$ , the throughput of IoT device  $n$  can be expressed as

$$R_n^{mg} = B \log_2 \left\{ 1 + \frac{p_n |h_n^{mg}|^2}{\sum_{k \in \mathcal{U}_{mg}, k \neq n, |h_k^{mg}| < |h_n^{mg}|} p_k |h_k^{mg}|^2 + \sigma^2} \right\}, \quad (1)$$

where  $p_k$  denotes the transmit power of the IoT device  $k \in \mathcal{N}$ ,  $\sigma^2$  indicates the noise power, and  $B$  represents the bandwidth of the occupied subcarrier.

Each IoT device faces two choices for processing its DSCI tasks. One is to utilize its own computation capacity to process tasks locally, and the other is to offload its tasks to candidate MEC servers for processing. To characterize whether IoT device  $n$  chooses to offload its tasks to BS-MEC  $m$ , we introduce an offloading decision variable  $a_{n,m}$ , which can be defined as follows:

$$a_{n,m} = \begin{cases} 1, & \text{offload } n's \text{ tasks to BS-MEC } m; \\ 0, & \text{process tasks locally.} \end{cases} \quad (2)$$

When the IoT device  $n$  processes its tasks locally, the required processing time can be denoted as

$$T_n^{\text{local}} = \frac{W_n}{f_n^{\text{local}}}, \quad (3)$$

where  $W_n$  represents the computational demands, also referred

to as the size of the task needed to be processed, and  $f_n^{\text{local}}$  indicates the computation capacity of the device  $n$  itself.

When the IoT device  $n$  selects offloading its DSCI tasks to the BS-MEC  $m$  via the subchannel  $g$ , then the transmission delay of the device  $n$  can be given by

$$T_{n,m,g}^{\text{trans}} = \frac{W_n}{R_n^{mg}}. \quad (4)$$

Since there may be multiple IoT devices offloading their DSCI tasks to the identical BS-MEC  $m$ , it inevitably leads to the issue of competition for computing resources. Therefore, the computing delay of IoT device  $n$  can be expressed as

$$T_{n,m}^{\text{comp}} = \frac{W_n + \sum_{i \in \mathcal{N}, i \neq n} a_{i,m} W_i}{f_m}, \quad (5)$$

where  $f_m$  denotes the computation capacity of BS-MEC  $m$ , and the term  $\sum_{i \in \mathcal{N}, i \neq n} a_{i,m} W_i$  indicates total data volume of all IoT devices that offload tasks to BS-MEC  $m$  except for device  $n$ . Therefore, the total delay when device  $n$  chooses MEC offloading can be denoted as

$$T_{n,m,g}^{\text{mec}} = T_{n,m,g}^{\text{trans}} + T_{n,m}^{\text{comp}} = \frac{W_n}{R_n^{mg}} + \frac{W_n + \sum_{i \in \mathcal{N}, i \neq n} a_{i,m} W_i}{f_m}. \quad (6)$$

Then, the delay expressions  $T_n$  of IoT device  $n$  under these two offloading strategies can be given by

$$T_n = \begin{cases} \frac{W_n}{R_n^{mg}} + \frac{W_n + \sum_{i \in \mathcal{N}, i \neq n} a_{i,m} W_i}{f_m}, & \text{BS-MEC Offloading;} \\ \frac{W_n}{f_n^{\text{local}}}, & \text{Local Offloading.} \end{cases} \quad (7)$$

Moreover, to describe the user grouping, we introduce the variables  $b_{g,n}$  to indicate whether IoT device  $n$  is connected to sub-channel  $g$ , and  $b_{g,n}$  can be specifically defined as

$$b_{g,n} = \begin{cases} 1, & \text{IoT device } n \text{ access group } g; \\ 0, & \text{IoT device } n \text{ unconnect to group } g. \end{cases} \quad (8)$$

#### IV. PROBLEM FORMULATION AND ANALYSIS

Building on the analysis in Sec. III, this section aims to jointly optimize the task offloading, user grouping, and power allocation schemes under the scenarios of device heterogeneity, to minimize the total delay of the proposed uplink NMBM networks. The detailed optimization problem is formulated as follows:

$$\mathcal{P}1 : \min_{\{\mathcal{U}_{m,g}, p_n\}_{\forall n \in \mathcal{N}, m \in \mathcal{M}, g \in \mathcal{G}}} \sum_{n=1}^N T_n, \quad (9a)$$

$$\text{s.t.} \quad 0 \leq p_n \leq p_{\max}, \quad \forall n \in \mathcal{N}, \quad (9b)$$

$$0 \leq T_n \leq D_n, \quad \forall n \in \mathcal{N}, \quad (9c)$$

$$\sum_{m=1}^M a_{n,m} \leq 1, \quad n \in \mathcal{N}, \quad (9d)$$

$$\sum_{g=1}^G b_{g,n} \leq 1, \quad n \in \mathcal{N}, \quad (9e)$$

$$\mathcal{U}_{m,g} \cap \mathcal{U}_{m,g'} = \emptyset, \quad \forall g, g' \in \mathcal{G}, g \neq g', \quad (9f)$$

$$\mathcal{U}_{m,g} \cap \mathcal{U}_{m',g} = \emptyset, \quad \forall m, m' \in \mathcal{M}, m \neq m', \quad (9g)$$

where constraint (9b) denotes the transmission power limit of the IoT device, in which  $p_{\max}$  indicates the maximum transmit power of each device. Constraint (9c) represents the delay threshold of DSCI tasks, which means that each task must be completed within that time. Constraints (9d) and (9e) indicate that each device can only access a maximum of one BS-MEC and sub-channel during task offloading, respectively. Finally, constraints (9f) and (9g) represent that IoT devices accessing different groups and offloading to different BS-MECs do not overlap with each other.

The original problem  $\mathcal{P}1$  involves a mix of discrete and continuous variables, and extremely complex interdependent coupling between task offloading, user grouping, and power allocation. Thus,  $\mathcal{P}1$  is classified as a Mixed Integer Nonlinear Programming (MINLP) problem, and directly tackling  $\mathcal{P}1$  presents significant challenges. To simplify the solution of  $\mathcal{P}1$ , we decompose the discrete and continuous variables for problem solution. We first fix the power allocation strategy  $\mathbf{p} = \{p_1, \dots, p_N\}$ , and leverages game theory to design efficient algorithms. Based on the obtained task offloading and user grouping strategies, the power allocation strategy  $\mathbf{p} = \{p_1, \dots, p_N\}$  can be effectively solved, at which point the optimization objective in  $\mathbf{T} = \{T_1, \dots, T_N\}$  is solely related to  $\mathbf{p} = \{p_1, \dots, p_N\}$ .

#### V. ALGORITHM DESIGNS AND SOLUTIONS

In this section, we propose the corresponding algorithms for the joint task offloading and user grouping subproblem, as well as the power allocation subproblem. Then, we develop a joint iterative optimization algorithm to effectively address  $\mathcal{P}1$ .

##### A. Potential Game-Based Joint Task Offloading and User Grouping Scheme

Based on the structure of the joint task offloading and user grouping subproblem, we intend to leverage game theory to tackle it. Let  $\pi_n$  denote the device  $n$ 's task offloading and user grouping selection, where  $\pi_n = \{m, g\} \Leftrightarrow n \in \mathcal{U}_{m,g}$ . The strategy profile of all devices can be represented as  $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_N\}$ , and  $\boldsymbol{\pi}_{-n}$  denotes the strategies of all devices except  $n$ . According to the objective function  $T_n$  formulated in (7), if the device's local computing resources can satisfy the delay requirements, it computes locally without uplink transmission. Otherwise, it offloads its DSCI tasks to the BS-MEC  $m$ , requiring a suitable offloading and grouping design. Based on (6), when the device  $n$  offloads its tasks to BS-MEC  $m$ , the total delay is composed of transmission and computation delays, where the former part is impacted by the intra-group interference, and the second part is related to the computing resource competition from other devices offloaded on the same BS-MEC  $m$ . Specifically, greater intra-group interference reduces throughput and increases transmission delay. Moreover, when the competition for computing resources among devices intensifies, the computation delay will also increase accordingly. To model these dynamic interactions among IoT devices, we define an interference function as follows:

$$\nu_n = \sum_{k \in \mathcal{U}_{m,g}, k \neq n, |h_k^{mg}| < |h_n^{mg}|} p_k |h_k^{mg}|^2 + \sum_{i \in \mathcal{N}, i \neq n} W_i a_{i,m}, \quad (10)$$

where the first term denotes the intra-group interference caused by other IoT devices that select the identical sub-channel for task transmission, and the second term reflects the competition for computing resources triggered by other devices that offload their tasks to the identical BS-MEC.

Based on the above analysis, we transform the objective of minimizing the total delay (9a) into minimizing the sum of interference from all devices. Then,  $\mathcal{P}1$  can be reconstructed into a non-cooperative game model, where each IoT device acts as an independent game player, characterized by **absolute rationality** and **self-interest**, aiming to minimize its own interference. In this case, the game problem can be expressed using the tuple  $\Gamma = (\mathcal{N}, \mathcal{M}, \mathcal{G}, \{h_n\}_{n \in \mathcal{N}}, \{\nu_n\}_{n \in \mathcal{N}})$ , where  $\mathcal{N}$  denotes the set of game players,  $\nu_n$  indicates the defined interference function of device  $n$ . Notably,  $\Gamma$  is a deterministic strategy game, which means that the strategies of players are explicit and do not involve probabilistic transitions [39]. In game theory, when all players reach a state where no player can unilaterally change their strategy, we refer to this state as a Nash Equilibrium (NE):

**Definition 1. (Nash Equilibrium)** For the game player  $\forall n \in \mathcal{N}$ , the system reaches the NE state when the strategy  $\pi^* = (\pi_1^*, \pi_2^*, \dots, \pi_N^*)$  satisfies the following conditions, as follows:

$$\nu_n(n \in \mathcal{U}_n^{\pi_n^*}, \pi_{-n}^*) \leq \nu_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*), \quad (11)$$

where  $n \in \mathcal{U}_n^{\pi_n^*}$  represents that the task offloading and grouping scheme belongs to  $\pi_n^*$ .

The system only converges when the game reaches the NE state; however, not all games can achieve the NE state. To guarantee the system's convergence, we transform the original problem  $\mathcal{P}1$  into an Exact Potential Game (EPG) model, which ensures the existence of an NE. The definition of the EPG can be given as follows:

**Definition 2. (Exact Potential Game)** For  $\forall n \in \mathcal{N}$ ,  $\forall g, g' \in \mathcal{G}$ , and  $\forall m, m' \in \mathcal{M}$ , if there exists a function  $\Phi_n$  that satisfies the condition (12), the game  $\Gamma = (\mathcal{N}, \mathcal{M}, \mathcal{G}, \{h_n\}_{n \in \mathcal{N}}, \{\nu_n\}_{n \in \mathcal{N}})$  is a strictly EPG and the function  $\Phi_n$  represents the exact potential function of the EPG  $\Gamma$ , as follows:

$$\begin{aligned} & \Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) - \Phi_n(n \in \mathcal{U}_{m',g'}, \pi_{-n}^*) \\ &= \nu_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) - \nu_n(n \in \mathcal{U}_{m',g'}, \pi_{-n}^*). \end{aligned} \quad (12)$$

The **Definition 2** reveals the change in potential function  $\Phi_n$  during the transfer of device  $n$  from one offloading-group combination  $(m, g)$  to another offloading-group combination  $(m', g')$ , which is equivalent to the change in the original interference function  $\nu_n$ . Although a detailed definition of the EPG's function has been provided in **Definition 2**, there has been no specific proof of whether the game  $\Gamma$  satisfies the conditions of an EPG, which can be proven in **Theorem 1**, as follows:

**Theorem 1.** Based on Definition 2, the closed-form expression

of the  $\Phi_n$  is derived by (14), which makes Eq. (12) established.

*Proof.* According to the general deductive method of game theory [39, 40], we give the expression  $\Phi_n$  that may satisfy (12) based on (10), as follows:

$$\begin{aligned} \Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) &= \frac{1}{2W_n} \sum_{u=1}^N \sum_{i=1, i \neq u}^N \ell_{\{a_{i,m}=a_{u,m}\}} W_i W_u - \sum_{i=1, i \neq n} p_i |h_i^{mg}|^2 \\ &+ 2 \sum_{u=1}^N \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} \ell_{\{a_{i,m}=a_{u,m}\}} \ell_{\{b_{g,i}=b_{g,u}\}} p_i |h_i^{mg}|^2, \end{aligned} \quad (13)$$

where  $\ell_{\text{condition}}$  denotes a binary indicator that takes the value 1 when the indicated condition is true, and takes the value 0 otherwise. The term  $\sum_{u=1}^N \sum_{i=1, i \neq u}^N \ell_{\{a_{i,m}=a_{u,m}\}} W_i W_u$  can be discussed in three cases as follows:

- 1) If  $u = n, i \neq n, \Rightarrow \sum_{i \neq n} \ell_{\{a_{i,m}=a_{n,m}\}} W_i W_n;$
- 2) If  $i = n, u \neq n \Rightarrow \sum_{u \neq n} \ell_{\{a_{u,m}=a_{n,m}\}} W_n W_u;$
- 3) If  $i \neq n, u \neq n \Rightarrow \sum_{u \neq n} \sum_{\substack{i \neq u \\ i \neq n}} \ell_{\{a_{i,m}=a_{u,m}\}} W_i W_u.$

Therefore, we can derive that

$$\begin{aligned} & \frac{1}{2W_n} \sum_{u=1}^N \sum_{i=1, i \neq u}^N \ell_{\{a_{i,m}=a_{u,m}\}} W_i W_u \\ &= \frac{1}{2W_n} \left\{ \sum_{i \neq n} \ell_{\{a_{i,m}=a_{n,m}\}} W_i W_n + \sum_{u \neq n} \ell_{\{a_{u,m}=a_{n,m}\}} W_n W_u \right\} \\ &+ \underbrace{\frac{1}{2W_n} \sum_{u \neq n} \ell_{\{a_{u,m}=a_{n,m}\}} W_n W_u}_{\mathcal{C}(n \in \mathcal{U}_{m,g})} \\ &= \sum_{\substack{i=1 \\ i \neq n}} \ell_{\{a_{i,m}=a_{n,m}\}} W_i + \mathcal{C}(n \in \mathcal{U}_{m,g}). \end{aligned} \quad (14)$$

We define  $H(u)$  as follows:

$$H(u) = \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} \ell_{\{a_{i,m}=a_{u,m}\}} \ell_{\{b_{g,i}=b_{g,u}\}} p_i |h_i^{mg}|^2. \quad (15)$$

Then, we can derive that

$$\begin{aligned} & - \sum_{i=1, i \neq n} p_i |h_i^{mg}|^2 + 2 \sum_{u=1}^N H(u) \\ &= - \sum_{i=1, i \neq n} p_i |h_i^{mg}|^2 + 2H(n) + 2 \sum_{u \neq n} H(u), \end{aligned} \quad (16)$$

where we can further derive that

$$\begin{aligned}
& - \sum_{i=1, i \neq n} p_i |h_i^{mg}|^2 + 2H(n) \\
& = - \sum_{i=1, i \neq n} p_i |h_i^{mg}|^2 + 2 \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2 \\
& = \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2 - \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ i \neq n}} p_i |h_i^{mg}|^2 - \sum_{i \notin \mathcal{U}_{m,g}} p_i |h_i^{mg}|^2 \\
& \quad + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2 \\
& = - \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| > |h_u^{mg}|}} p_i |h_i^{mg}|^2 - \sum_{i \notin \mathcal{U}_{m,g}} p_i |h_i^{mg}|^2 + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2. \tag{17}
\end{aligned}$$

Based on (16) and (17), we can obtain that

$$\begin{aligned}
& - \sum_{i=1, i \neq n} p_i |h_i^{mg}|^2 + 2 \sum_{u=1}^N H(u) \\
& = \left\{ 2 \sum_{\substack{j=1 \\ j \neq n}} \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}| \\ i \neq n, i \neq j}} p_i |h_i^{mg}|^2 - \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| > |h_u^{mg}|}} p_i |h_i^{mg}|^2 \right. \\
& \quad \left. - \sum_{i \notin \mathcal{U}_{m,g}} p_i |h_i^{mg}|^2 \right\} + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2 \\
& = \mathcal{T}(n \in \mathcal{U}_{m,g}) + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2. \tag{18}
\end{aligned}$$

Combining (13), (14), and (18), the potential function  $\Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*)$  can be rewritten as follows:

$$\begin{aligned}
\Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) & = \sum_{\substack{i=1 \\ i \neq n}} \ell_{\{a_{i,m}=a_{n,m}\}} W_i + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_u^{mg}|}} p_i |h_i^{mg}|^2 \\
& \quad + \mathcal{C}(n \in \mathcal{U}_{m,g}) + \mathcal{T}(n \in \mathcal{U}_{m,g}), \tag{19}
\end{aligned}$$

where the expression of  $\mathcal{C}(n \in \mathcal{U}_{m,g})$  and  $\mathcal{T}(n \in \mathcal{U}_{m,g})$  are independent of the offloading and grouping strategies  $(m, g)$  of device  $n$ , and they can be regarded as constant terms. Therefore, when the offloading and grouping strategies of IoT device  $n$  change from  $\mathcal{U}_{m,g}$  to  $\mathcal{U}_{m',g'}$ , the corresponding change in the value of the EPF  $\Phi_n$  can be denoted as follows:

$$\begin{aligned}
& \Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) - \Phi_n(n \in \mathcal{U}_{m',g'}, \pi_{-n}^*) \\
& = \left( \sum_{\substack{i=1 \\ i \neq n}} \ell_{\{a_{i,m}=a_{n,m}\}} W_i + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_n^{mg}|}} p_i |h_i^{mg}|^2 \right) \\
& \quad - \left( \sum_{\substack{i=1 \\ i \neq n}} \ell_{\{a_{i,m'}=a_{n,m'}\}} W_i + \sum_{\substack{i \in \mathcal{U}_{m',g'}, \\ |h_i^{m'g'}| < |h_n^{m'g'}|}} p_i |h_i^{m'g'}|^2 \right) \\
& = \sum_{\substack{i=1 \\ i \neq n}} \ell_{\{a_{i,m}=a_{n,m}\}} W_i + \sum_{\substack{i \in \mathcal{U}_{m,g}, \\ |h_i^{mg}| < |h_n^{mg}|}} p_i |h_i^{mg}|^2 \\
& \quad - \sum_{\substack{i=1 \\ i \neq n}} \ell_{\{a_{i,m'}=a_{n,m'}\}} W_i - \sum_{\substack{i \in \mathcal{U}_{m',g'}, \\ |h_i^{m'g'}| < |h_n^{m'g'}|}} p_i |h_i^{m'g'}|^2
\end{aligned}$$

$$= \nu_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) - \nu_n(n \in \mathcal{U}_{m',g'}, \pi_{-n}^*), \tag{20a}$$

which satisfies (12) in the **Definition 2**.  $\square$

Based on **Definition 2** and (10), the change in the interference function (10) has the same trend with the EPF, thus we can claim the existence of the NE state according to **Theorem 1**, as described in **Corollary 1** as follows:

**Corollary 1.** *The game  $\Gamma = (\mathcal{N}, \mathcal{M}, \mathcal{G}, \{h_n\}_{n \in \mathcal{N}}, \{\nu_n\}_{n \in \mathcal{N}})$  can converge to at least one NE state.*

*Proof.* As defined in **Definition 1**, a NE denotes a state, where each IoT device selects its optimal task offloading and user grouping strategy given the strategies of others, such that no active device can reduce its experienced inter-group interference by unilaterally changing its strategy. As defined in **Definition 2**, every such game has at least one NE, implying that the game is destined to reach a NE within a finite number of steps. In our considered game  $\Gamma = (\mathcal{N}, \mathcal{M}, \mathcal{G}, \{h_n\}_{n \in \mathcal{N}}, \{\nu_n\}_{n \in \mathcal{N}})$ , the set of device  $\mathcal{N}$  is finite, and the strategy evolution towards minimizing the value of interference function. Once the desired objective is achieved, strategies remain unchanged and are repeatedly selected, indicating that a NE has been reached.  $\square$

As mentioned above, in the potential game model, each player adjusts their strategy towards reducing their own interference. This behavior pattern follows the Finite Improvement Property (FIP) that is defined as follows:

**Definition 3. Finite Improvement Property (FIP)** *In potential games, due to the fixed number of strategy spaces, the increasing path length of a game is limited. Therefore, after a limited number of strategy iterations and adjustments, players in a game reach an NE state, enabling the system to achieve optimal performance. This process is known as FIP, which gradually improves system performance through a finite number of iterations and ultimately reaches NE, ensuring the convergence of the system in games involving absolutely rational and selfish players.*

According to **Definition 3**, **Corollary 2** is derived as follows:

**Corollary 2.**  *$\Gamma = (\mathcal{N}, \mathcal{M}, \mathcal{G}, \{h_n\}_{n \in \mathcal{N}}, \{\nu_n\}_{n \in \mathcal{N}})$  meets the conditions of the FIP.*

*Proof.* In the considered  $\Gamma = (\mathcal{N}, \mathcal{M}, \mathcal{G}, \{h_n\}_{n \in \mathcal{N}}, \{\nu_n\}_{n \in \mathcal{N}})$ , due to the finite number of IoT devices, BSs, and subchannels, the corresponding sets of devices  $\mathcal{N}$ , BSs  $\mathcal{M}$ , and subchannels  $\mathcal{G}$  are also limited. This means that the number of optional strategy spaces for offloading and grouping in  $\Gamma$  is fixed. In addition, when players in  $\Gamma$  adjust their strategies, their strategies evolve towards reducing interference and improving individual interests, ensuring that they do not repeatedly choose the same offloading and grouping strategies  $(m, g)$ . Therefore,  $\Gamma$  will stop at an optimal strategy point, reaching a NE state after a limited number of iterative adjustments.  $\square$

In order to guarantee that the game results can achieve optimal performance, it is also necessary to ensure that the game  $\Gamma$  is Pareto optimal, which can be defined as follows:

**Definition 4. Pareto Optimality (PO)** *describes an ideal state of offloading and grouping strategies in which no players' situation can be worsened without improving the utility of at*

least one other player, meaning that the system can benefit one or more players without harming any others. A state is considered Pareto optimal when it is impossible to make any further improvement to the utility of at least one player without negatively impacting the others. In other words, a Pareto optimal state is one where no changes in offloading and grouping strategies can increase the performance of any player without causing harm to others, indicating that the system has reached a global optimal solution. Pareto improvement refers to the process and method of moving from an initial state to a Pareto optimal state.

---

**Algorithm 1:** The Potential Game-based Optimal Joint Offloading and Grouping Algorithm.

---

**Input:** IoT device set  $\mathcal{N}$ , BS set  $\mathcal{M}$ , subchannel set  $\mathcal{G}$ , channel coefficient  $h_n^{mg}$ , delay constraints  $D_n$ ;  
**Output:** Optimal task offloading and grouping strategies  $\pi^*$ .

- 1 **Initialize:**  $\mathcal{U}_{m,g} = \emptyset, \forall m \in \mathcal{M}, g \in \mathcal{G}$ ;
- 2 Randomly initialize the offloading and grouping strategy  $\pi_n$  for each device  $n \in \mathcal{N}$ ;
- 3 **repeat**
- 4   **for**  $t = 1$  to  $T$  **do**
- 5     **for**  $n = 1$  to  $N$  **do**
- 6       Calculate  $T_n^{\text{local}}$  based on Eq. (3);
- 7       **if**  $T_n^{\text{local}} \leq D_n$  **then**
- 8         **for**  $j = 1$  to  $M$  **do**
- 9           |  $a_{n,j} = 0$ ;
- 10          **end**
- 11       **else**
- 12         **for**  $m = 1$  to  $M$  **do**
- 13           | Calculate  $\Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n})$  based on Eq. (13);
- 14          **end**
- 15         Find
- 16          $m_n^* = \arg \min_{m \in \mathcal{M}} \Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n})$ ;
- 17         Update the offloading strategy of device  $n$  to  $m_n^*$ ;
- 18       **end**
- 19     **end**
- 20     **for**  $t = 1$  to  $T$  **do**
- 21       **for**  $n = 1$  to  $N$  **do**
- 22         **for**  $g = 1$  to  $G$  **do**
- 23           | Calculate  $\Phi_n(n \in \mathcal{U}_{m_n^*,g}, \pi_{-n})$  based on Eq. (13);
- 24          **end**
- 25          Find  $g_n^* = \arg \min_{g \in \mathcal{G}} \Phi_n(n \in \mathcal{U}_{m_n^*,g}, \pi_{-n})$ ;
- 26          Update the grouping strategy of device  $n$  to  $g_n^*$ ;
- 27       **end**
- 28     **end**
- 29 **until**  $\pi^* = \{\pi_n^*\}_{n \in \mathcal{N}} = (m_n^*, g_n^*)_{n \in \mathcal{N}}$  remains unchanged;
- 30 **return**  $\pi^*$  and  $T_n, \forall n \in \mathcal{N}$ .

---

In the game  $\Gamma$ , any unilateral change in offloading and grouping strategies  $(m, g)$  may lead to a decrease in the potential function for certain players, thus reducing their interference. However, it could also cause an increase in the potential function value of other players, thereby increasing their interference. According to **Definition 4**, when the potential game  $\Gamma$  reaches a NE state, the system can be considered Pareto optimal. Given that potential games can ensure the achievement of NE, we propose the potential game-based optimal joint offloading and grouping algorithm, as detailed in **Algorithm 1**, which aims to realize the optimal task offloading and user

grouping strategies. **Algorithm 1** utilizes the properties of potential games to update players' offloading and grouping decisions iteratively, ultimately achieving a NE in the system. Thus, based on ensuring the interests of each player, the optimal offloading and grouping strategies of the entire system can be achieved. At the beginning of **Algorithm 1**, it first clears each group and initializes the offloading and grouping strategies for each device (**Line 1-2**). Before the game begins, the system first evaluates whether the device's computing capacity meets the delay requirements  $D_n$ . If the delay limit of  $D_n$  is met, the device chooses local computing, which helps reduce device interference caused by offloading a large number of tasks to the same BS-MEC. If the computation capacity of the device itself is insufficient to meet the delay requirements, it chooses to offload to BS-MEC and then enters the game process corresponding to **Line 11-22**, which aims to determine the optimal offloading and grouping strategies for each device.

### 1) Optimality and Computational Complexity Analysis:

The proposed **Algorithm 1** is based on the framework of dynamic potential games. In each iteration, given the strategies  $\pi_{-n}$  of the other device  $n \in \mathcal{N}$ , autonomously selects its offloading and grouping strategy to minimize its interference, which is equivalently represented by the potential function. Since the formulated game is an EPG, any unilateral deviation by player  $n$  yields an identical change in both its interference function and the global potential function. Formally, for any unilateral deviation we have (12). Consequently, whenever a player strictly improves its interference function, the potential function also strictly decreases, as follows:

$$\begin{aligned} \nu_n(n \in \mathcal{U}_{m',g'}, \pi_{-n}^*) &< \nu_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) \\ \Leftrightarrow \Phi_n(n \in \mathcal{U}_{m',g'}, \pi_{-n}^*) &< \Phi_n(n \in \mathcal{U}_{m,g}, \pi_{-n}^*) . \end{aligned} \quad (21)$$

Since each player's strategy set is finite, the global strategy space  $\Pi$  is also finite. Therefore, the potential function can only take values from a finite set  $\{\Phi(\pi) : \pi \in \Pi\}$ . As a result, an infinite strictly decreasing sequence of potential values cannot exist, which is known as the FIP described in **Definition 3**, guaranteeing that the iterative process terminates in a finite number of steps. Once no player can further improve its interference function through unilateral deviation, the system reaches a NE. By **Definition 4**, such an NE also satisfies the criterion of Pareto optimality, since no further Pareto improvements can be achieved at that point.

The main computational cost of **Algorithm 1** lies in **Steps 11-21**, where each device  $n$  evaluates and updates its offloading and grouping strategies. The complexity of this evaluation is  $\mathcal{O}(M + G)$ . With  $N$  devices in the system, and assuming that on average  $C$  iterations are required for **Algorithm 1** to converge, then the overall computational complexity can be given by  $\mathcal{O}(NC(M + G))$ . In the worst-case scenario, the maximum number of iterations  $C$  is upper-bounded by the size of the joint strategy space, i.e.,  $C \leq |\Pi| - 1 = \prod_{n=1}^N |S_n| - 1$ , where  $S_n$  denotes the strategy set of player  $n$ . However, in practice, the algorithm typically converges with  $C$  much smaller than this bound.

### B. Convex Optimization-Based Power Allocation Scheme

Given the task offloading and user grouping strategies  $\pi^*$  obtained by **Algorithm 1**,  $\mathcal{P}1$  can be degenerated into the power allocation subproblem, as described in  $\mathcal{P}2$ .

$$\mathcal{P}2 : \min_{\{p_n\}_{n \in \mathcal{N}}} \sum_{n=1}^N T_n, \quad (22a)$$

$$\text{s.t.} \quad 0 \leq p_n \leq p_{max}, \quad \forall n \in \mathcal{N}, \quad (22b)$$

$$0 \leq T_n \leq D_n, \quad \forall n \in \mathcal{N}, \quad (22c)$$

$$E_n \leq E_{th}, \quad \forall n \in \mathcal{N}. \quad (22d)$$

$\mathcal{P}2$  is a non-convex function with respect to the power allocation scheme  $\mathbf{p} = \{p_1, \dots, p_N\}$ , as it exists not only in the numerator but also in the denominator during computation. In order to overcome the solving challenges caused by non convexity, we transform the original minimization total delay into maximization of transmission rate, which can find the suboptimal solutions of  $\mathcal{P}2$ . This method utilizes the inverse relationship between maximizing transmission rate and minimizing delay, that is, indirectly reducing the total delay of computation and transmission by increasing transmission rate. After transforming the optimization objective,  $\mathcal{P}2$  can be rewritten as follows:

$$\mathcal{P}3 : \max_{\{p_n\}_{n \in \mathcal{N}}} \sum_{n=1}^N R_n^{mg}, \quad (23a)$$

$$\text{s.t.} \quad (22b) - (22d), \quad (23b)$$

where the expression of  $R_n^{mg}$  can be formulated as follows:

$$R_n^{mg} = B \log_2 \left( \sigma^2 + \sum_{\substack{i \in \mathcal{U}_{m,g} \\ |h_i^{mg}| \leq |h_n^{mg}|}} p_i |h_i^{mg}|^2 \right) - B \log_2 \left( \sigma^2 + \sum_{\substack{i \in \mathcal{U}_{m,g} \\ |h_i^{mg}| < |h_n^{mg}|}} p_i |h_i^{mg}|^2 \right). \quad (24)$$

$\mathcal{P}3$  can be effectively tackled using Majorization-Minimization (MM) algorithm [15, 41], which can simplify the solution of non-convex problems in various ways, such as the Taylor expansion method, convex inequality method, etc. In this paper, we leverage the Taylor expansion method to optimize  $\mathcal{P}3$ , which can effectively transform non-convex functions into convex functions. The Taylor expansion method transforms the original non-convex problem into a convex problem by expanding it using the Taylor series. For example, given a first-order differentiable function  $f$ , its Taylor expansion at point  $x_0$  can be formulated as

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0). \quad (25)$$

Based on the principle of Taylor expansion, the expression of  $R_n^{mg}$  can be expanded as follows:

$$\tilde{R}_n^{mg} = B \log_2 \left( \sigma^2 + \sum_{\substack{i \in \mathcal{U}_{m,g} \\ |h_i^{mg}| \leq |h_n^{mg}|}} p_i^{(t)} |h_i^{mg}|^2 \right)$$

$$- B \log_2 \left( \sigma^2 + \sum_{\substack{i \in \mathcal{U}_{m,g} \\ |h_i^{mg}| < |h_n^{mg}|}} p_i^{(t-1)} |h_i^{mg}|^2 \right) - \frac{B}{\ln 2} \frac{\left( \sum_{\substack{i \in \mathcal{U}_{m,g} \\ |h_i^{mg}| < |h_n^{mg}|}} (p_i^{(t)} - p_i^{(t-1)}) |h_i^{mg}|^2 \right)}{\left( \sigma^2 + \sum_{\substack{i \in \mathcal{U}_{m,g} \\ |h_i^{mg}| < |h_n^{mg}|}} p_i^{(t-1)} |h_i^{mg}|^2 \right)}. \quad (26a)$$

After Taylor expansion,  $\tilde{R}_n^{mg}$  is already a convex function, thus  $\mathcal{P}3$  can be transformed into a convex problem, as described in  $\mathcal{P}4$ , as follows:

$$\mathcal{P}4 : \max_{\{p_n\}_{n \in \mathcal{N}}} \sum_{n=1}^N \tilde{R}_n^{mg}, \quad (27a)$$

$$\text{s.t.} \quad (22b) - (22d). \quad (27b)$$

---

### Algorithm 2: Majorization-Minimization (MM) Based Power Allocation Algorithm.

---

**Input:** Device set  $\mathcal{N}$ ; BS-MEC set  $\mathcal{M}$ ; Subchannel set  $\mathcal{G}$ ; Channel coefficient  $h_n^{mg}$ ; Delay constraints  $D_n$ ; Task offloading and grouping strategies  $\pi$ ; Convergence threshold  $\varepsilon$ ; Maximum iteration numbers  $T_{max}$ ;

**Output:** Optimal power allocation  $\mathbf{p}^* = \{p_1^*, \dots, p_N^*\}$ .

- 1 Under the premise of ensuring constraints (22b), (22c) and (22d), initialize the power allocation scheme of devices;
  - 2 Initialize the number of iteration  $t = 1$ ;
  - 3 **while**  $|\tilde{R}_n^{mg(t)} - \tilde{R}_n^{mg(t-1)}| > \varepsilon$  and  $t < T_{max}$  **do**
  - 4   Using CVX to solve problem  $\mathcal{P}4$  and obtaining the power allocation scheme  $\mathbf{p}^{(t)}$ ;
  - 5    $t = t + 1$ ;
  - 6 **end**
- 

After the transformation of Taylor expansion, the original power allocation problem  $\mathcal{P}2$  is effectively transformed into a convex optimization problem  $\mathcal{P}4$ , which can be solved using the existing convex optimization toolkit, like CVX. Moreover, considering that the objective function  $\tilde{R}_n^{mg}$  is obtained through an iterative process, this paper needs to use an iterative algorithm to gradually approach the optimal power allocation scheme, in order to ensure that the algorithm can converge to the optimal solution while meeting certain accuracy requirements. Specifically, the detailed steps of the power allocation algorithm are detailed in **Algorithm 2**.

### C. AO-Based Joint Task Offloading, User Grouping, and Power Allocation Scheme

In this subsection, we propose the alternating optimization (AO)-based joint task offloading, user grouping, and power allocation algorithm to effectively address  $\mathcal{P}1$ . Firstly, we fix the power allocation scheme  $\mathbf{p} = \{p_n\}_{n \in \mathcal{N}}$ , and then invoke **Algorithm 1** to derive the optimal task offloading and user grouping strategies, i.e.,  $\mathbf{a} = \{a_{m,n}\}_{m \in \mathcal{M}, n \in \mathcal{N}}$  and

$\mathbf{b} = \{b_{g,n}\}_{g \in \mathcal{G}, n \in \mathcal{N}}$ , respectively. Secondly, we use the derived  $[\mathbf{a}, \mathbf{b}]$  as the inputs for Algorithm 2, and then execute it to update the power allocation scheme  $\mathbf{p}$ . Loop iteratively optimize the above process until  $[\mathbf{a}, \mathbf{b}, \mathbf{p}]$  converges to stable solutions or the termination conditions are satisfied. The detailed process mentioned above can be described by **Algorithm 3**.

**Algorithm 3:** AO-Based Joint Task Offloading, User Grouping, and Power Allocation Scheme.

```

Input: IoT device set  $\mathcal{N}$ , BS set  $\mathcal{M}$ , subchannel set  $\mathcal{G}$ , channel coefficient  $h_n^{mg}$ , delay constraints  $D_n$ , convergence threshold  $\varepsilon$ , maximum iteration numbers  $T_{max}$ ;
Output: Optimal task offloading and grouping strategies  $\pi^* = [\mathbf{a}^*, \mathbf{b}^*]$ , and optimal power allocation scheme  $\mathbf{p}^*$ 

1 for  $n = 1$  to  $N$  do
2   Each device  $n$  is connected to the nearest BS;
3   Each device  $n$  randomly selects subchannels;
4   Initializes power allocation  $p_n$ ;
5 end
6 repeat
7   Execute Algorithm 1 to derive the offloading and grouping strategies  $\pi^* = [\mathbf{a}^*, \mathbf{b}^*]$ ;
8   Execute Algorithm 2 to obtain the power allocation scheme  $\mathbf{p}^*$ ;
9 until All devices do not change their offloading and grouping strategies or reach the maximum number of iterations  $T_{max}$ ;
10 return  $[\mathbf{a}^*, \mathbf{b}^*, \mathbf{p}^*]$ .

```

## VI. PERFORMANCE EVALUATION

In this section, we present extensive simulation experiments and discussions to validate the effectiveness of our proposed joint task offloading, user grouping, and power allocation scheme. We compare our algorithms against state-of-the-art methods and classic heuristic approaches, and provide detailed comparative analyses. Moreover, comprehensive ablation studies are conducted to underscore the key contributions of this work and confirm the overall superiority of our schemes.

TABLE I: Experimental Parameter Settings

Parameter	Value
Number of BS-MEC $M$	4
Number of IoT device $N$	80
Number of subchannels $G$	5
Bandwidth of each subchannel $B$	410 kHz
Large-scale fading	$128.1 + 37.6 \log(d_{m,n} [\text{km}])$
Noise power spectral density $N_0$	-174 dBm/Hz
Size of tasks $W_n$	[5, 15] Mbits
Delay Constraints $D_n$	[0.2, 1.2] s
Computing capacity of IoT devices $f_n^{\text{local}}$	[3, 8] Mbits/s
Computing capacity of MEC servers $f_m$	[7, 10] Gbits/s
Maximum transmit power of devices $p_{max}$	27.8 dBm
Energy constraint $E_{th}$	[0.152, 0.910] J

### A. Experimental Parameter Settings

In our considered simulation scenario, IoT devices are randomly distributed within a rectangular area of  $1000 \text{ m} \times 1000 \text{ m}$ , with four BSs located at coordinate points (250 m, 750 m), (250 m, 250 m), (750 m, 750 m), and (750 m, 250 m), respectively. We consider that the small-scale

fading follows a Rayleigh distribution with the mean of 0 and variance of 1, and follows an independent and identically distributed (i.i.d.) Gaussian distribution. The noise power can be calculated according to  $\sigma^2 = BN_0$ . The specific settings of other simulation parameters are shown in Table I.

### B. Baseline Schemes for Comparison

To comprehensively evaluate the effectiveness of the proposed joint task offloading, user grouping, and power allocation scheme (denoted **Proposed**), we compare it against several state-of-the-art methods and classic heuristic approaches. In particular, for the user grouping in NOMA-assisted systems, we select two of the most popular grouping strategies, as follows:

- Gale-Shapley Grouping (**Gale-Shapley**) [28, 42, 43]: In this strategy, devices and groups are regarded as two sets with their own preference lists, with each device tending to access the groups with better channel conditions, and each group also tending to select the devices with higher channel gains.
- Max-Min based Grouping (**Max-Min**) [27, 44]: Each IoT device is grouped into the group with the greatest difference in channel conditions, so that the NOMA can decode the messages of IoT devices by leveraging the differences of channel conditions among IoT devices.

For the comparison of task offloading, we select two of the popular task offloading strategies, as follows:

- Nearby BS-based Offloading (**Nearby**) [29]: Each IoT device will prioritize offloading its tasks to the nearest BS and select the sub-channel with the least interference for access.
- Computing Capacity-based Offloading (**Computing**): In this strategy, each IoT device offloads its task to the BS-MEC with the strongest computing capacity that currently allows access devices (i.e., those that have not yet reached the maximum device access limit).

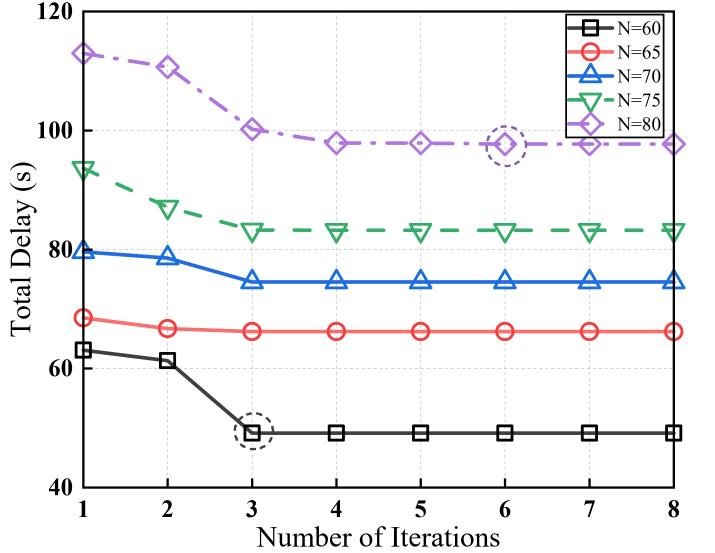


Fig. 2: The convergence performance of the proposed joint task offloading, user grouping, and power allocation scheme under different numbers of IoT devices.

### C. Analysis of Experimental Results

1) *Convergence Verification:* As shown in Fig. 2, we demonstrate the convergence performance of the proposed joint task offloading, user grouping, and power allocation scheme under different numbers of IoT devices. As the number of IoT devices  $N$  increases, the number of iterations required for the algorithm to converge also increases. In particular, when  $N = 60$ , the total delay of the system can converge to a stable value after 3 iterations. However, when  $N = 80$ , the number of iterations required for the algorithm to converge increases to 6. This is mainly because, as the number of devices increases, the decision space available for each device to select from when making offloading and grouping strategies expands, which may require more iterations to achieve convergence. However, the average number of iterations of the proposed algorithm always remains below 6 times when no more than 80 IoT devices are connected, which fully demonstrates that our algorithm can quickly converge to the NE state in a short number of iterations.

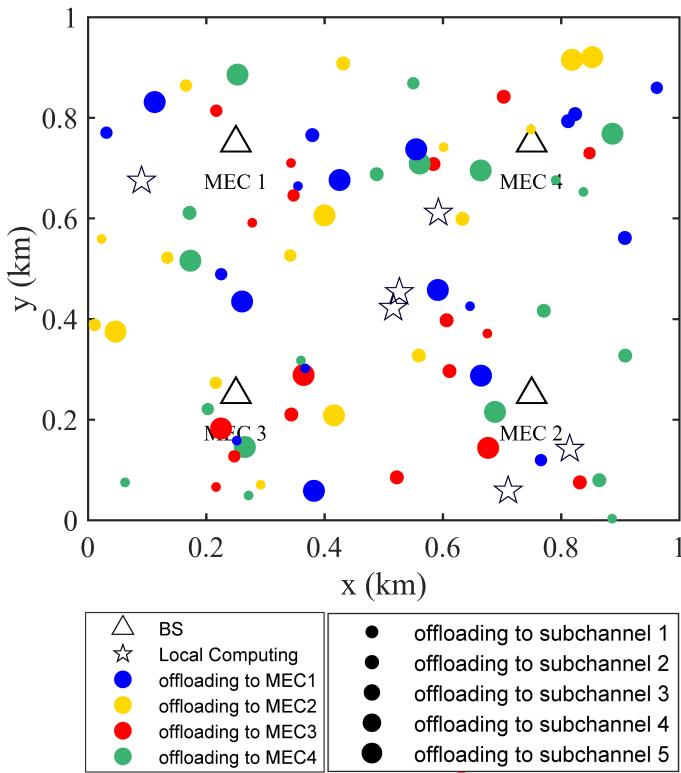


Fig. 3: The distribution of IoT device locations after convergence of the proposed joint task offloading, user grouping, and power allocation scheme.

As shown in Fig. 3, we present the visualization results of the device distribution and the task offloading and user grouping strategies of the algorithm after convergence. The observation results show that in the final offloading decision of each device, the device does not always select to offload its tasks to the nearest BS-MEC. This phenomenon explains that, after considering multiple factors, such as channel conditions, computing requirements, and BS-MEC computing capabilities, an IoT device may make offloading decisions that favor a non-nearest BS in order to optimize overall system performance,

rather than solely determining its own grouping strategy based on the quality of channel conditions.

### D. Performance Comparison

As illustrated in Fig. 4, we present the total delay and average delay performance of different algorithms under various IoT device access scenarios. Under a fixed number of BS-MECs  $M = 4$  and subchannels  $G = 5$ , it can be observed that both the total delay and average delay per device increase with the value of  $N$ . This is mainly because that as the number of IoT devices increases, the number of devices offloaded to the same BS-MEC also increases accordingly, leading to intensified intra-group interference among devices and consequently reducing the uplink transmission throughput. On the other hand, since the computing capacity of the BS-MEC server is also limited, the increase in the number of IoT devices leads to more intense competition for computing resources among devices, which in turn results in higher computational delays.

In addition, Fig. 4 (a) and (b) reveal that the proposed algorithm outperforms all other comparison algorithms, including Gale-Shapley, Max-Min, Nearby, and Computing, in terms of both total delay and average delay performance. In particular, our proposed schemes in this paper reduce the delay by at least 19.3% compared to these comparison algorithms, and up to 41.5%. These experimental results demonstrate that the proposed algorithm has significant advantages in achieving low latency when dealing with the IoT device access with DSCI tasks.

As shown in Fig. 5, we further investigate the total power consumption and average power consumption corresponding to different algorithms under various IoT device access scenarios. Under the set conditions of  $M = 4$  and  $G = 5$ , it can be observed that the system's total power consumption increases with the number of IoT devices. However, our proposed algorithm consistently maintains the lowest power consumption overhead, demonstrating its superior energy efficiency. Specifically, when  $N = 80$ , the proposed algorithm saves approximately 6.7% of power consumption compared to the Gale-Shapley, which has the second-best delay performance in Fig. 5. Furthermore, we study the average power consumption per IoT device under different IoT device access scenarios. It is evident that our proposed algorithm achieves significant improvements in average power consumption per IoT device compared to the Gale-Shapley, Max-Min, Nearby, and Computing. For example, in the scenario of IoT device access with DSCI tasks, the proposed algorithm achieves at least 14.7% reduction in average power consumption per IoT device.

We provide a more detailed analysis of the performance gains in delay and power consumption achieved by our proposed algorithm, as shown in Fig. 4 and Fig. 5. Unlike the comparison algorithms, the proposed algorithm allows for a flexible number of IoT devices in different groups, enabling more dynamic decision-making. In the grouping process, our algorithm takes into account not only the device's channel conditions but also intra-group interference, interference between different BSs, and the power allocation strategy, making the IoT device offloading and grouping decisions more superior than the selected schemes based solely on channel conditions

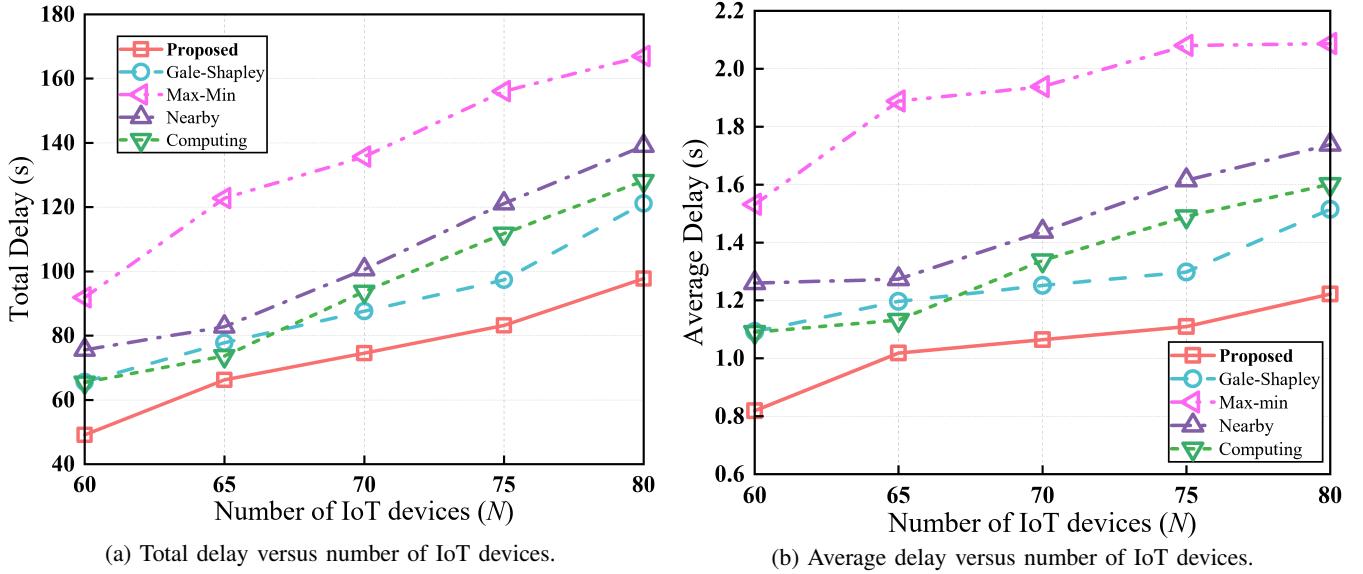


Fig. 4: Total and average delay performance under different IoT device access scenarios.

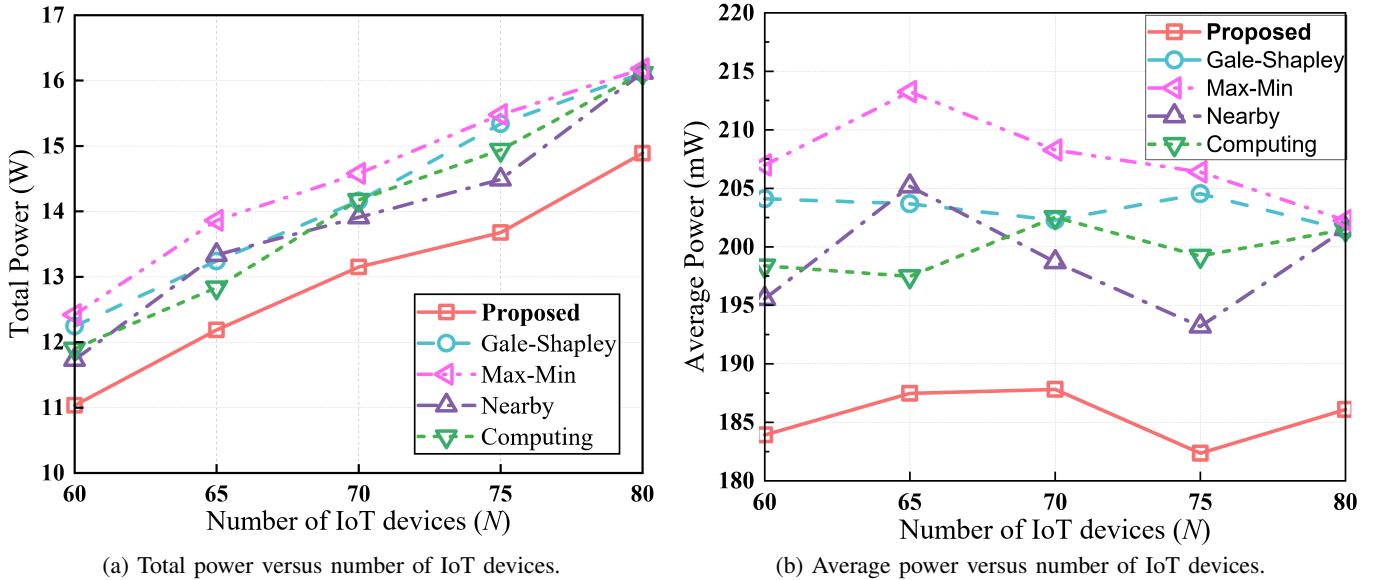


Fig. 5: Total and average power consumption performance under different IoT device access scenarios.

or distance. Moreover, in our algorithm, IoT devices dynamically determine their own strategies, unlike other comparison algorithms that determine all devices' strategies at once. This allows our approach to fully consider the impact of other devices on each device's performance. In contrast, algorithms like Gale-Shapley, Nearby, and Max-Min primarily prioritize based on device channel conditions, neglecting the significant effects of intra-group interference and power distribution on throughput (and thus on delay). The offloading strategies of the selected Nearby may result in too many tasks being offloaded to the same BS-MEC, leading to an imbalanced computational load and a significant increase in intra-group interference, which impacts throughput and increases the system's computational burden. The Computing algorithm, on the other hand, mainly focuses on the computation requirements of device tasks, overlooking the impact of intra-group interference, power allocation, and other key factors. By comprehensively

considering factors such as channel conditions, intra-group and inter-base station interference, and power allocation, our algorithm presents a more refined and efficient task offloading and grouping strategy, significantly enhancing system performance and ensuring lower delays and higher computational efficiency.

As shown in Fig. 6, we studied the performance of the system's total delay and total power consumption under different numbers of groups  $G$  (i.e., number of sub-channels) under large-scale access of IoT devices ( $N = 80$ ). As shown in Fig. 6 (a), with the increase of the number of groups, the delay performance corresponding to different algorithms has been correspondingly improved. However, our proposed algorithm achieves the best latency performance compared to Gale Shapley, Max Min, Nearby, and Computing. This is mainly because, as the number of sub-channels increases, IoT devices offloaded to the same BS-MEC will have more group options, resulting in a corresponding decrease in the number of

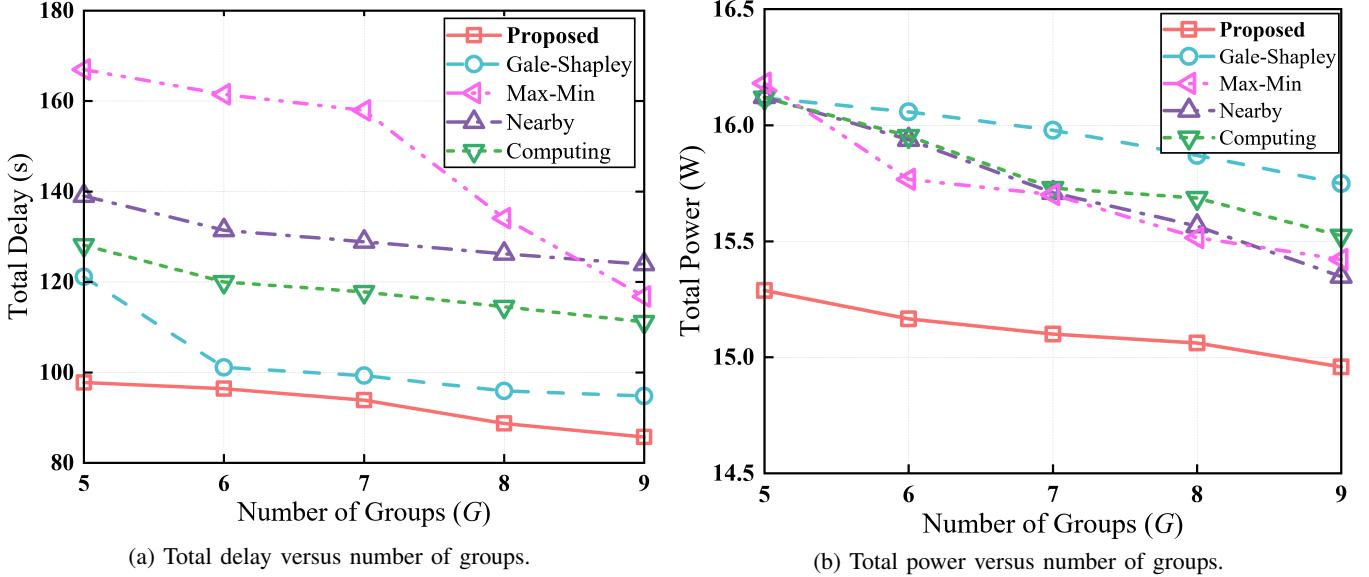


Fig. 6: Performance variations of system's total delay and total power consumption under different number of groups in IoT access scenarios with DSCI tasks. ( $N = 80$ ).

IoT devices within each group. This change means that intra-group interference will be improved to a certain extent, which in turn will lead to an increase in throughput and ultimately improve the latency performance of the system.

As shown in Fig. 6 (b), this paper also investigates the impact of changes in the number of groups on total power consumption. It can be observed that as the number of groups increases, the total power consumption of the system shows a decreasing trend. This is because the increase in the number of groups leads to a decrease in the number of devices in each group, thereby reducing intra-group interference. At this point, lower power can be allocated to meet the delay requirements of each device. This result indicates that the algorithm proposed in this paper can not only effectively reduce system latency compared to the comparative algorithm, but also reduce system power consumption.

As shown in Fig. 7, we demonstrate the impact of varying the number of BS-MECs on the system's total delay performance in IoT access scenarios with DSCI tasks. It can be observed that as the number of BS-MECs increases, the total delay performance corresponding to different algorithms improves accordingly. Notably, the proposed algorithm consistently maintains the best delay performance compared to the benchmark algorithms, including Gale-Shapley, Max-Min, Nearby, and Computing, when the number of BS-MECs increases from  $M = 4$  to  $M = 8$ . However, for the Gale-Shapley algorithm, its delay improvement becomes less significant when  $M \geq 6$ , primarily due to the inherent design of the Gale-Shapley algorithm itself. Initially, as the number of BS-MECs increases, the system benefits from more distributed computing resources and reduced load per BS, thereby improving the system's delay performance. However, given a fixed number of IoT devices, once the number of BS-MECs exceeds a certain threshold, further increasing the number of BSs does not significantly alleviate the load on the BS-MEC nodes, because the number of IoT devices  $N$  and the number of sub-channels  $G$  are fixed. The Gale-Shapley

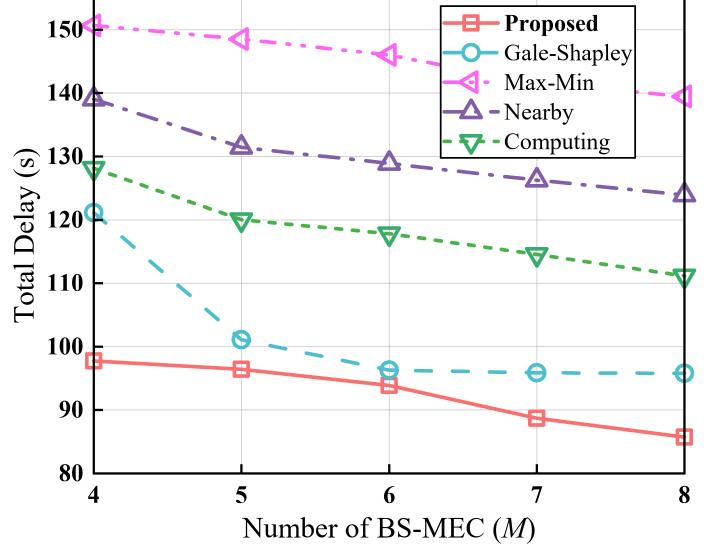


Fig. 7: Performance variations of system's total delay under different number of BS-MECs in IoT access scenarios with DSCI tasks ( $N = 80$ ).

algorithm prioritizes allocation based on channel gains, which may lead to an uneven distribution of devices among BSs, especially when high-gain devices are concentrated in a few BSs with better channel conditions. This concentration results in continuous intra-group interference and causes some BSs to become overloaded, creating computational bottlenecks that offset the potential delay reduction benefits of adding more BSs.

As shown in Fig. 8, we investigated the impact of computing capacity of BS-MEC on the system's total delay performance in IoT access scenarios with DSCI tasks. It can be observed that as the computing capacity of each BS-MEC increases, the system's total delay corresponding to different algorithms improves accordingly. Notably, it can be noticed that when the computing capacity of BS-MECs reaches a certain range,

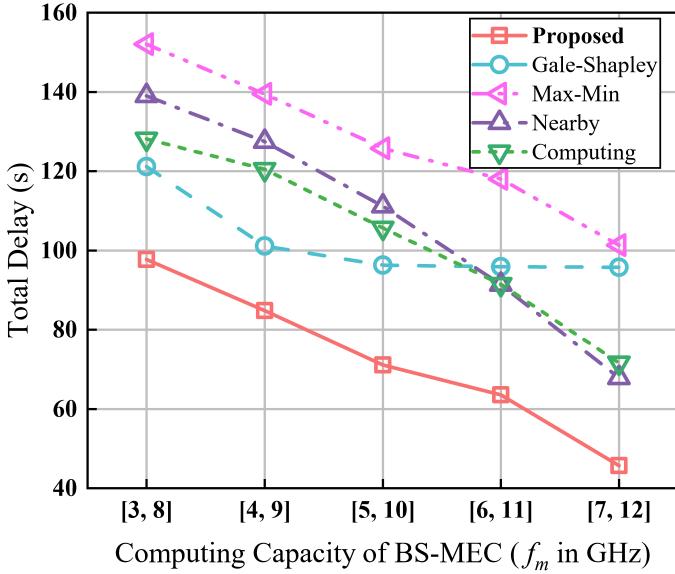


Fig. 8: The impact of BS-MEC's computing capacity on system's total delay performance in IoT access scenarios with DSCI tasks.

the delay performance gain of the Gale-Shapley algorithm similarly reaches a bottleneck. This is mainly because the preference-based matching approach of the Gale-Shapley algorithm leads to an uneven distribution of devices, where high-gain devices tend to concentrate in groups associated with BS-MECs that have favorable channel conditions. Therefore, even with increased computing resources, these groups with those overloaded BS-MEC still experience severe intra-group interference and resource contention, which limits throughput and maintains high communication delay. This imbalance prevents a significant reduction in delay despite enhanced computing capabilities. In addition, it can be observed that under different ranges of BS-MEC computing capacities, the proposed algorithm consistently maintains superior performance gains compared to other benchmark algorithms, which fully demonstrates the effectiveness of our proposed algorithm.

## VII. CONCLUSION AND FUTURE OUTLOOK

This paper proposed an uplink NOMA-assisted multi-BS-MEC network to fulfill the QoS requirements of massive IoT device connectivity with DSCI tasks. To effectively address the challenges of unbalanced subchannel access, inter-group interference, computational load disparities, and device heterogeneity in multi-BS MEC systems, this paper leverages game theory to propose an innovative joint task offloading and user grouping algorithm that dynamically makes optimal offloading-grouping decisions based on the channel states of individual IoT devices. Secondly, this paper introduces an MM-based power allocation algorithm that transforms the original power allocation problem into a tractable convex optimization paradigm, thereby deriving the optimal power allocation strategy. Moreover, an efficient iterative optimization algorithm is further designed to minimize the total system delay. Extensive simulation experiments have verified the effectiveness of our proposed optimization schemes, demonstrating performance improvements of up to 19.3% and 14.7% in terms of the system's total delay and power consumption, respectively.

Although the current work considers an uplink NOMA-assisted multi-MEC framework for large-scale IoT networks with DSCI tasks, each device is restricted to associating with a single BS, which severely limits mobility support. In future work, we will incorporate service migration mechanisms, enabling devices to dynamically offload tasks across multiple BSs and make optimal handover decisions based on channel conditions, MEC loads, and user mobility trajectories. This is expected to significantly enhance task offloading success rates, transmission reliability, and overall system performance in high-mobility scenarios.

## REFERENCES

- [1] C. Saadouni, S. E. Jaouhari, N. Tamani, S. Ziti, L. Mroueh, and K. E. Bouchti, "Identification techniques in the internet of things: Survey, taxonomy and research frontier," *IEEE Commun. Surveys Tuts.*, pp. 1–1, 2025.
- [2] R. R. Chowdhury and P. E. Abas, "A survey on device fingerprinting approach for resource-constraint IoT devices: Comparative study and research challenges," *Internet of Things*, vol. 20, p. 100632, 2022.
- [3] M. El-Hajj, A. Fadlallah, M. Chamoun, and A. Serhrouchni, "A survey of internet of things (IoT) authentication schemes," *Sensors*, vol. 19, no. 5, p. 1141, 2019.
- [4] M. Anjum, M. A. Khan, S. A. Hassan, A. Mahmood, H. K. Qureshi, and M. Gidlund, "RSSI fingerprinting-based localization using machine learning in LoRa networks," *IEEE Internet of Things Mag.*, vol. 3, no. 4, pp. 53–59, 2020.
- [5] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in internet-of-things," *IEEE Internet of Things J.*, vol. 4, no. 5, pp. 1250–1258, 2017.
- [6] ARO, "IoT devices to generate nearly 80 zettabytes of data by 2025," Apr. 2023, accessed: 2025-10-23. [Online]. Available: <https://aro.tech/insights/blog/iot-devices-to-generate-nearly-80-zettabytes-of-data-by-2025/>
- [7] Y. Chen, F. Guo, C. Wu, S. Liu, H. Lu, and C. W. Chen, "AoI-aware task offloading and transmission optimization for industrial IoT networks: A branching deep reinforcement learning approach," *arXiv preprint arXiv:2510.16414*, 2025.
- [8] Y. Chen, H. Lu, C. Wu, L. Qin, and X. Guo, "Performance optimization in RSMA-assisted uplink xURLLC IIoT networks with statistical QoS provisioning," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2025.
- [9] J. Bi, H. Yuan, K. Zhang, and M. Zhou, "Energy-minimized partial computation offloading for delay-sensitive applications in heterogeneous edge networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 4, pp. 1941–1954, 2022.
- [10] G. Sun, J. Wu, and Z. Sun, "Collaborative computation offloading, caching, communication, computation, and trajectory control in UAV-assisted MEC system," in *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, 2024, pp. 203–204.
- [11] B. Qiu, Y. Wang, H. Xiao, and Z. Zhang, "Deep reinforcement learning-based adaptive computation offloading and power allocation in vehicular edge computing networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13 339–13 349, 2024.
- [12] L. Qin, H. Lu, Y. Chen, Z. Gu, D. Zhao, and F. Wu, "Energy-efficient blockchain-enabled user-centric mobile edge computing," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 4, pp. 1452–1466, 2024.
- [13] Y. Chen, C. Lu, Y. Huang, C. Wu, F. Guo, H. Lu, and C. W. Chen, "DMSA: A decentralized microservice architecture for edge networks," *arXiv preprint arXiv:2501.00883*, 2025.
- [14] Y. Chen, C. Wu, F. Zhang, C. Lu, Y. Huang, and H. Lu, "Topology-aware microservice architecture in edge networks: Deployment optimization and implementation," *IEEE Trans. Mobile Comput.*, vol. 24, no. 7, pp. 6090–6105, 2025.
- [15] L. Qin, H. Lu, Y. Chen, B. Chong, and F. Guo, "Joint transmission and resource optimization in NOMA-assisted IoVT with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 9984–9999, 2024.
- [16] C. W. Chen, "Internet of video things: Next-generation IoT with visual sensors," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6676–6685, 2020.
- [17] J. Chen, H. Wu, R. Li, and P. Jiao, "Green parallel online offloading for DSCI-type tasks in IoT-edge systems," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7955–7966, 2022.
- [18] C. Yi, J. Cai, and Z. Su, "A multi-user mobile computation offloading and transmission scheduling mechanism for delay-sensitive applications," *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 29–43, 2020.

- [19] Y. Chen, H. Lu, L. Qin, Y. Deng, and A. Nallanathan, "When xURLLC meets NOMA: A stochastic network calculus perspective," *IEEE Commun. Mag.*, vol. 62, no. 6, pp. 90–96, 2024.
- [20] B. Chong, H. Lu, Y. Chen, L. Qin, and F. Guo, "Achievable sum rate optimization on NOMA-aided cell-free massive MIMO with finite blocklength coding," *IEEE Trans. Veh. Technol.*, vol. 74, no. 7, pp. 10881–10895, 2025.
- [21] Z. Liu, J. Qi, Y. Shen, K. Ma, and X. Guan, "Maximizing energy efficiency in UAV-assisted NOMA-MEC networks," *IEEE Internet of Things J.*, vol. 10, no. 24, pp. 22 208–22 222, 2023.
- [22] Q. Gao, Y. Liu, X. Mu, M. Jia, D. Li, and L. Hanzo, "Joint location and beamforming design for STAR-RIS assisted NOMA systems," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2532–2546, 2023.
- [23] L. Liu, B. Sun, X. Tan, and D. H. K. Tsang, "Energy-efficient resource allocation and subchannel assignment for NOMA-enabled multiaccess edge computing," *IEEE Syst. J.*, vol. 16, no. 1, pp. 1558–1569, 2022.
- [24] M. Hosseinpour and M. H. Yaghmaee Moghaddam, "Quality-of-experience-aware computation offloading in MEC-enabled blockchain-based IoT networks," *IEEE Internet of Things J.*, vol. 11, no. 8, pp. 14483–14493, 2024.
- [25] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE Access*, vol. 6, pp. 6900–6919, 2018.
- [26] F. Du, C. Xiao, D. Wei, X. Wang, Y. Zhang, S. Geng, and X. Zhao, "Adaptive interference hypergraph-based secure resource allocation for multicell multicarrier MISO-NOMA IoT networks with imperfect CSI," *IEEE Internet of Things J.*, vol. 12, no. 19, pp. 41 071–41 086, 2025.
- [27] O. Maraqa, S. Al-Ahmadi, A. S. Rajasekaran, H. U. Sokun, H. Yanikomeroglu, and S. M. Sait, "Energy-efficient optimization of multi-user NOMA-assisted cooperative THz-SIMO MEC systems," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3763–3779, 2023.
- [28] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2771–2784, 2017.
- [29] W. Xu, J. Yu, Y. Wu, and D. H. K. Tsang, "Energy-latency aware intelligent reflecting surface aided multi-cell mobile edge computing," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 1, pp. 362–374, 2024.
- [30] A.-T. Le, D.-H. Tran, C.-B. Le, P. T. Tin, T. N. Nguyen, Z. Ding, H. V. Poor, and M. Voznak, "Power beacon and NOMA-assisted cooperative IoT networks with co-channel interference: Performance analysis and deep learning evaluation," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 7270–7283, 2024.
- [31] Y. Cheng, J. Lu, D. Niyato, B. Lyu, M. Xu, and S. Zhu, "Performance analysis and power allocation for covert mobile edge computing with RIS-aided NOMA," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 4212–4227, 2024.
- [32] J. Wu, T. Xu, T. Zhou, X. Chen, H. Hu, and C. Wu, "Adaptive NOMA-based spectrum sensing for uplink IoT networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 1, pp. 138–149, 2024.
- [33] Z. Chen, J. Tang, M. Wen, Z. Li, J. Yang, X. Y. Zhang, and K.-K. Wong, "Reconfigurable intelligent surface assisted MEC offloading in NOMA-enabled IoT networks," *IEEE Trans. Commun.*, vol. 71, no. 8, pp. 4896–4908, 2023.
- [34] H. Hong, J. Jiao, T. Yang, Y. Wang, R. Lu, and Q. Zhang, "Age of incorrect information minimization for semantic-empowered NOMA system in S-IoT," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 6639–6652, 2024.
- [35] Y. Chen, J. Xu, Y. Wu, J. Gao, and L. Zhao, "Dynamic task offloading and resource allocation for NOMA-aided mobile edge computing: An energy efficient design," *IEEE Trans. Services Comput.*, vol. 17, no. 4, pp. 1492–1503, 2024.
- [36] W. He, D. He, X. Ma, X. Chen, Y. Fang, and W. Zhang, "Joint user association, resource allocation, and beamforming in RIS-assisted multi-server MEC systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 2917–2932, 2024.
- [37] R. Wei and W. Zheng, "NeurDORAs: Neural-aided decentralized offloading based on resource auction," in *Proceedings of the 22nd ACM International Conference on Computing Frontiers*, 2025, pp. 64–67.
- [38] L. Dai, F. Zeng, H. Kong, J. Cai, H. Jiang, and K. Li, "Throughput-aware cooperative task offloading in dynamic mobile edge computing systems," *IEEE Trans. Mobile Comput.*, pp. 1–16, 2025.
- [39] B. Wang, Y. Wu, and K. R. Liu, "Game theory for cognitive radio networks: An overview," *Computer networks*, vol. 54, no. 14, pp. 2537–2561, 2010.
- [40] W. Wu, H. Song, H. Wang, and H. Dong, "Potential game based task offloading in the high-speed railway with reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12 671–12 685, 2023.
- [41] J. Kim, J. Kim, and S.-H. Park, "Joint design of power control and SIC decoding order for max-min fairness optimization in uplink NOMA systems," in *2021 International Conference on Information Networking (ICOIN)*, 2021, pp. 339–342.
- [42] Q. Chen, J. Wu, J. Wang, and H. Jiang, "Coexistence of URLLC and eMBB services in MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 839–851, 2023.
- [43] W. Ji, B. Liang, Y. Wang, R. Qiu, and Z. Yang, "Crowd V-IoE: Visual internet of everything architecture in AI-driven fog computing," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 51–57, 2020.
- [44] Y. Cheng, K. H. Li, K. C. Teh, and S. Luo, "Joint user pairing and subchannel allocation for multisubchannel multiuser nonorthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8238–8248, 2018.



**Yuang Chen** (Graduate Student Member, IEEE) received the B.S. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2021. He is currently pursuing a doctoral degree in Electronic Engineering and Information Science at the University of Science and Technology of China, Hefei, China. In addition, he is currently working as a full-time research assistant in the department of computing at The Hong Kong Polytechnic University, Hong Kong SAR, China. His research interests include 5G/6G wireless network technology, such as next-generation URLLC, next-generation multiple access technology, wireless network resource allocation and performance optimization, microservice deployment and scheduling, etc.



**Fengqian Guo** received the Ph.D. degree in communication and information systems from the University of Science and Technology of China (USTC), Hefei, China, in 2022. He is currently a associate researcher at the University of Science and Technology of China. His research interests include wireless low-latency transmission and wireless resource optimization.



**Chang Wu** received the B.S. degree from the Dalian Maritime University (DLMU), Dalian, China, in 2021. He is currently working toward the PhD degree in communication and information systems with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China. His research interests include 5G/6G wireless network technologies, such as architecture, QoS/QoE provision for business transmission and Deep Reinforcement Learning in performance optimization.



**Mingyu Peng** received the B.S. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2025. He is currently pursuing the master's degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China. His research interests include multimedia communication, wireless edge networks, and deep reinforcement learning.



**Hancheng Lu (Senior Member, IEEE)** received his Ph.D. in communication and information systems from the University of Science and Technology of China, Hefei, China, in 2005. He is currently a tenured professor in the Department of Electronic Engineering and Information Science at the University of Science and Technology of China. He is also working at the Hefei National Comprehensive Science Center Artificial Intelligence Research Institute, Hefei, China. He has rich research experience in multimedia communication, wireless edge networks, future network architecture and protocols, as well as machine learning algorithms for network communication, involving scheduling, resource management, routing, transmission, and other fields. In the past 5 years, more than 80 papers have been published in top journals such as IEEE Trans and flagship conferences such as IEEE INFOCOM, and have won the Best Paper Award of IEEE GLOBECOM 2021 and the Best Paper Award of WCSP 2019 and WCSP 2016 in the field of communication. In addition, he currently serves as an editorial board member for numerous journals, including the IEEE Internet of Things Journal, China Communications, and IET Communications.



**Chang Wen Chen (Life Fellow, IEEE)** received the B.S. degree from the University of Science and Technology of China in 1983, the M.S.E.E. degree from the University of Southern California in 1986, and the Ph.D. degree from the University of Illinois at Urbana-Champaign in 1992. He is currently the Chair Professor in the Department of Visual Computing and the Interim Dean of the School of Computer and Mathematical Sciences at The Hong Kong Polytechnic University. Before his current position, he served as the Dean of the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, from 2017 to 2020. He was an Empire Innovation Professor with the University at Buffalo, The State University of New York, from 2008 to 2021. He was an Allen Henry Endow Chair Professor with Florida Institute of Technology from 2003 to 2007. He was a Faculty Member of electrical and computer engineering at the University of Rochester from 1992 to 1996 and at the University of Missouri-Columbia from 1996 to 2003. His research has been funded by both government agencies and industrial corporations. His research interests include multimedia communication, multimedia systems, mobile video streaming, the Internet of Video Things (IoVT), image/video processing, computer vision, deep learning, multimedia signal processing, and immersive mobile video. He was a SPIE Fellow in 2007 and an Elected Member of Academia Europaea in 2021. He and his students have received ten best paper awards or best student paper awards over the past two decades. He received several research and professional achievement awards, such as the Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, the Alexander von Humboldt Research Award in 2010, the University at Buffalo Exceptional Scholar—Sustained Achievement Award in 2012, the SUNY System Chancellor’s Award for Excellence in Scholarship and Creative Activities in 2016, the University of Illinois ECE Distinguished Alumni Award in 2019, the Outstanding Overseas Contributor of the China Society of Image and Graphics (CSIS) in China MM 2024, and the SIGMM Outstanding Technical Achievement Award in 2024. He is currently an Associate Editor-in-Chief of IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE and a Deputy Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He served as the conference chair for several major IEEE, ACM, and SPIE conferences related to multimedia communications and signal processing. He served as the Editor-in-Chief for IEEE TRANSACTIONS ON MULTIMEDIA from January 2014 to December 2016 and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from January 2006 to December 2009. He has been an Editor of several other major IEEE TRANSACTIONS and journals, including PROCEEDINGS OF THE IEEE, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, and IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.