Delay Minimization for NOMA-mmW Scheme Based MEC Offloading

Jia Shi[1], Yifan Zhou[1], Zan Li[1], Zhongling Zhao[1], Zheng Chu[2] and Pei Xiao[2]

[1] State Key Laboratory of ISN, Xidian University, Xi'an, 710071, China. Emails: jiashi@xidian.edu.cn, fyzhou@stu.xidian.edu.cn, zanli@xidian.edu.cn, zhonglingzhao@stu.xidian.edu.cn.

[2] 5GIC, University of Surrey, Guildford, Surrey, GU1 2XH, UK. Emails: andrew.chuzheng7@gmail.com, p.xiao@surrey.ac.uk.

***Abstract*—By introducing nonorthogonal multiple access (NOMA) based millimeter wave (mmW) communication, it can significantly improve the transmission efficiency of mobile edge computing (MEC) offloading. In this paper, we are motivated to investigate the resource allocation (RA) problem of the NOMA-mmW scheme based MEC offloading system, by jointly optimizing the beamwidth, user equipment (UE) scheduling and transmit power. To tackle the mixed integer nonlinear programming (MINLP) problem of delay minimization, we develop the alternative optimization (AO) approach based RA scheme, namely AO-RA, to obtain the close-optimum solutions. In the AO-RA scheme, we propose the matrix control many-to-one with externality (MC-M2OE) algorithm, to find the best UE scheduling for the NOMA groupings of different types of UEs. Up on the above, we further design the joint beamwidth and transmit power (JBTP) algorithm, which determines the optimal beamwidth and transmit power for the MEC offloading transmissions. Our simulation results show the effectiveness of the proposed AO-RA scheme in minimizing the offloading delay, where our MC-M2OE and JBTP algorithms can significantly outperform the existing approaches. From the simulation results, we may conclude that, it needs to carefully address the trade-off between beam alignment overhead and transmission gain, while properly balancing the loading among different NOMA groups, for the practical consideration of NOMA-mmW MEC technology.**

I. Introduction

IN the B5G/6G era, there will be emerging new types of applications [1], such as augmented/virtual reality online games, autonomous driving, and smart everything. These applications exploit the same features of, computation intensive, delay sensitivity, which definitely overload the capacity of mobile terminals. The emergence of mobile cloud computing (MCC) brings solutions to the above problems [2]. In addition, some scholars also considered the security of user information and proposed related routing protocols and algorithms [2, 3]. However, the high latency of mobile cloud computing still cannot well meet the low latency requirements of users. For this sake, mobile edge computing (MEC) technology will be an indispensable enabler for high-computational and delay sensitivity services, by means of distributing the computing resources at the network edge in the vicinity of end-UEs [4]. Therefore, MEC is bound to become one of the key technologies for low-latency services in the B5G/6G networks [5].

The current studies on MEC technology mainly focus on three aspects: computing, caching, and offloading. In the field of MEC computing, many scholars have addressed offloading decision-making and computing resource management. For instance, [6] investigated the joint problem of offloading decision-making and computing resource allocation in the vehicle network, aiming at improving the system computation time. In [7] and [8], they combined caching and MEC offloading, and proposed a multi-user cooperative offloading strategy based on cache-assisted MEC, which effectively reduced system energy consumption and task executing delay. As for MEC offloading, K. Guo et al. designed the online learning based offloading algorithm, which fully explored the interplay between communication and computation with enriched user experience and reduced energy consumption [9]. In [10], the authors applied the MEC technology to investigate a joint problem of fast charging station selection and EV route planning, where a deep reinforcement learning (DRL) based solution was proposed. In the vehicular edge computing networks [11], Guo et al. discussed the possible networking, communication and computing technologies for 6G vehicular network, narrated many promising hardware devices and facilities with 6G characteristics, and analyzed the pivotal roles of AI. In addition, the authors in [12] used the block coordinate descent and successive convex approximation to solve the problem of UAV trajectory optimization and offloading decision-making, which effectively reduced the energy consumption in the UAV aided MEC system. Apparently, the research on MEC offloading is of great significance, and is playing the most important role in realizing MEC technology.

From the above research, we know that the current studies on MEC offloading were mainly based on UHF communication. By contrast, the use of mmW communication will bring significant benefits to MEC systems [13]. Compared with UHF, mmW communication, with massive spectrum resources, can hugely improve the transmission rate of MEC offloading, which greatly reduces the task delay in the MEC system. Due to the characteristics of short wave length, sensitivity to blockage and high penetration loss, mmW communications require highly directional transmissions in order to overcome the propagation defects, and to provide enhanced capacity. This makes the offloading of mmW MEC systems, especially in resource allocation, much more complicated than that of UHF communication based MEC systems.

Some researches on mmW based MEC technology have been carried out, where the pioneer works [14, 15] showed its feasibility, and proved that the combination of the two

can effectively reduce the task offloading delay. Furthermore, the current studies [16–18] have investigated how to apply the mmW MEC technique to different wireless systems with diverse applications. Specifically, in [16], they developed the joint proactive computing and mmW communication scheme for the wireless virtual reality (VR) system with ultra-reliable and low latency demands. By contrast, the authors of [17] proposed the unmanned aerial vehicle (UAV)-aided wireless edging service system with mmW communication capability. Further, Q. Chen et al. proposed a new architecture of mmW MEC technology for the Industrial Internet of Things (IoT), and effectively reduced the system energy consumption by jointly optimizing communication and computing resources [18].

Due to the characteristics of short wave length, sensitivity to blockage and high penetration loss, mmW communications require highly directional transmissions in order to overcome the propagation defects, and to provide enhanced capacity. This makes the offloading of mmW MEC systems, especially in resource allocation, much more complicated than that of UHF communication based MEC systems. In particular, the authors of [19] designed a distributed joint hybrid beamforming and resource allocation algorithm, which also addressed the implementation issues for the future mmW MEC system. In [20], the authors merged computation offloading techniques for mmW MEC system and showed that, how the joint optimization of computation/communication resources was crucial to design an energy efficient MEC system. A federated learning empowered computation offloading and resource management (FLOR) was proposed in [21], which could outperform the traditional convex solutions in terms of computing complexity. The authors of [22] considered the uplink MEC system, and formulated the problem of minimizing the total energy consumption within the required latency. In [23], they jointly considered mmW MEC offloading under the coexistence of communication-oriented users and computing-oriented users. In addition, the study in [24] paid attention to the downlink of MEC offloading transmission, and discussed the task offloading of mmW MEC by optimizing backhaul bandwidth and edge server resource allocation to reduce the overall delay. Nevertheless, the above works rarely discussed the MEC offloading for multiuser scenarios, which may challenge the accessing efficiency of mmW links with limited beam coverage.

As a promising technology, nonorthogonal multiple access (NOMA) has been recognized as a promising solution to significantly improve the accessing efficiency for crowed networks, by allowing multiple users to share time and spectrum resources in the same spatial layer via power-domain multiplexing. The state-of-the-art showed that the NOMA based MEC offloading can significantly improve the link efficiency and lower the overall latency [25–29]. For instance, [25] proposed the MEC aware NOMA technique, which can exploit the benefits of spectrum efficiency. The studies of [26] and [27] addressed the radio resource management for the NOMA aided MEC system under single-carrier and multi-carrier assumptions, where the game theory based algorithms were proposed to find the Nash equilibrium solution. Furthermore, to investigate the energy efficiency problem, the authors of [28] combined the NOMA MEC technology with wireless power transmission (WPT) to overcome the double distance effect of the far users. In [29], they discussed the trade-off between minimizing offloading delay and energy consumption for the NOMA based MEC system.

A. Motivations and Contributions

Against this background, in this paper, we are motivated to study the delay performance of the NOMA-mmW scheme based MEC system by optimizing the resource allocation for the uplink transmissions of MEC offloading with multiple UEs' tasks happening simultaneously. Due to the poor data processing capability and limited battery power of UAV, the proposed NOMA-mmW MEC scheme can play a great role in UAV Communication [30, 31]. In order to improve the spectral efficiency and energy efficiency of the UAV communication system, our scheme use NOMA and MEC technologies to improve the link efficiency and reduce the communication delays, and use mmW technology to better cope with the UAVs' communication channel dominated by the line-of-sight (LoS) links. For clarify, the contributions of this paper are summarized as follows.

- By blending the concepts of NOMA technology and mmW communication, we conceive the transmission scheme, namely NOMA-mmW, which is used for MEC offloading to improve both the transmission and accessing efficiencies. To jointly consider UE scheduling, beamwidth selection and power allocation, we formulate the RA optimization problem aiming at minimizing the average offloading delay. Upon analyzing, we develop the alternative optimization method based resource allocation (AO-RA) scheme to decouple the mixed integer nonlinear programming (MINLP) problem into series of solvable sub-problems, where the convergency and optimality are proved.
- For solving the sub-problem of UE scheduling, we equivalently transform it to a matching process between two different kinds of UEs (F-UEs and nF-UEs) for NOMA grouping. For this sake, we propose the so-called matrix control aided many-to-one with externality (MC-M2OE) algorithm, where the matrix control method is developed to minimize the number of iterations required for traversing all the possible join-in and swap actions for each UE, as well as to reach the stable matching results.
- We propose the joint beamwidth and transmit power (JBTP) algorithm to solve the sub-problem of optimizing the beamwidth and transmit power of the NOMA-mmW based offloading links. In particular, the non-convex problem is equivalently converted to a series of convex ones, where the fractional logarithm terms are tackled by introducing efficient auxiliary variables and using first-order Taylor expansion. Furthermore, the characteristics of the proposed algorithms are also analyzed.
- We carry out the comprehensive performance analysis for the NOMA-mmW based MEC offloading system. Our simulation results show the effectiveness of the

TABLE I
LIST OF NOTATIONS

Symbol	Definition
$\lceil \cdot \rceil$	Round up operation
K	Total number of UEs
I	Number of nF-UEs
J	Number of F-UEs
C_k	Data amount of the k-th UEs task
D_k	Maximum delay requirement of the k-th UE
θ_{bs}	Beamwidth of base station
θ_k	Beamwidth of the k-th UE
τ_k^{BA}	Time required for beam alignment of the k-th UE
$\tau_j^{\text{DT,F}}$	Data transmission delay of the j-th F-UE
$\tau_i^{\text{DT,nF}}$	Data transmission delay of the i-th nF-UE
T_p	Time required for a pilot transmission
Υ_k	Sector-level beamwidth used by the k-th UE
Δ_0	Time slot of a NOMA group
Q	Maximum number of nF-UEs accessed by a NOMA group
G_{bs}	Main lobe gain for the base station
G_k	Main lobe gain for the k-th UE
z_0	Side lobe gain
L_k	Large-scale fading gain
d_k	Distance between the k-th UE and the base station
α	Pathloss exponent
h_k	Nakagami-m fading channel of the k-th UE
N_0	Noise variance
$s_{i,j}$	Scheduling variable
$\theta^{\min}$	Minimum beamwidth
$\theta^{\max}$	Maximum beamwidth
p_k	Transmit power of UEs
$p^{\max}$	Maximum transmit power of UEs
E_i	Channel condition information of i-th uF-UE
R_i^{nF}	Data rate of i-th uF-UE
R_j^{F}	Data rate of j-th F-UE

proposed AO-RA scheme in minimizing the offloading delay. In particular, the proposed MC-M2OE and JBTP algorithms can significantly outperform the existing algorithms. We may conclude that, in the NOMA-mmW offloading design of practical MEC system, one should properly schedule UE transmission by setting a suitable quota as well as finding the optimal NOMA grouping, rather than simply investing more transmit power causing unnecessary energy waste.

The remainder of this paper is organized as follows. In Section II, we introduce the system model of our NOMA-mmW based MEC. In Section III, we propose a low-complexity UE scheduling scheme. In Section IV, we develop an iterative algorithm for joint optimization of beamwidth and transmit power. Numerical results for evaluating the performance of the proposed schemes are provided in Section V. The conclusions are made in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

In our NOMA-mmW based MEC system conceived, as shown in Fig. 1, the base station (BS) is located in the origin of a sector area, and is equipped with a MEC server. A range of user equipments (UEs) are randomly distributed in the sector area, and their indexes are collected in the set $\mathcal{K} = \{1, 2, \ldots, K\}$. All the K UEs demand computing task offloading, which are supported by mmW communications with NOMA scheme employed. Assume that, each UE's

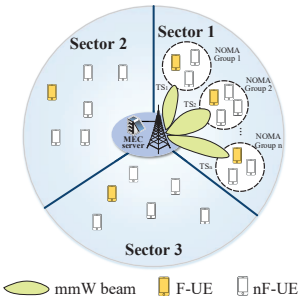

Fig. 1. Schematic for our NOMA-mmW MEC system conceived.

computing task can be represented by $\{C_k, D_k\}$, where C_k is the data amount of UE k's task, and D_k is the maximum delay requirement of the task.

Due to the use of NOMA-mmW scheme, there are two phases in the uplink transmission of MEC offloading: 1) beam alignment, and 2) data transmission. During the first phase, the best beamwidth is searched for one user by another. Let us assume that the beamwidth of BS θ_{bs} is fixed to cover the entire sector area, and the beamwidth of each UE, such as θ_k, needs to be optimized. Hence, the time required for beam alignment of an UE can be given by

$$\tau_k^{\text{BA}} = \lceil \frac{\Upsilon_k}{\theta_k} \rceil T_p, \ \forall k \in \mathcal{K}, \tag{1}$$

where T_p is the time required for a pilot transmission, Υ_k being a fixed value is the sector-level beamwidth used by UE k.

After that, during the second phase, the UEs transmit information of their tasks to the BS. We consider T number of equally divided time slots (TSs), each of which has the value of Δ_0 seconds. There are J UEs, denoted by $\mathcal{J} = \{1, 2, \ldots, J\}$, referred to as fixed UEs (F-UEs). Each of the F-UEs is orthogonally pre-assigned one TS for offloading, where F-UE j is scheduled in TS t, with assuming $J = T$. Further, the NOMA scheme is employed to accommodate the other I UEs, denoted by $\mathcal{I} = \{1, 2, \ldots, I\}$, on the existing T TSs. These I UEs are referred to as non-fixed UEs (nF-UEs). Note that, we have the following relations: $K = J + I$, $\mathcal{K} = \mathcal{J} \cup \mathcal{I}$, $\mathcal{J} \cap \mathcal{I} = \emptyset$. In another word, the nF-UEs in $\mathcal{I}$ need to be clustered into different NOMA groups to access the scheduled TSs, each of which has been assigned to a F-UE. For every NOMA group, the total number of UEs (not including the F-UE) are constrained by Q. The above process is referred to as UE scheduling.

Based on the characteristics of mmW communications, the effective antenna gains for the main lobe of transmitter and receiver can be respectively expressed as

$$G_{\text{bs}} = \frac{2\pi - (2\pi - \theta_{\text{bs}}) z_0}{\theta_{\text{bs}}}, \tag{2}$$

$$G_k = \frac{2\pi - (2\pi - \theta_k) z_0}{\theta_k}, \ \forall k \in \mathcal{K}, \tag{3}$$

where $0 \leq z_0 << 1$ is the side lobe gain, and is assumed to be the same for both the transmitters and the receiver.

Based on our NOMA-mmW scheme, the F-UEs and nF-UEs are clustered into different groups for offloading the task data

to the BS. In our system, the offloading links suffer both path-loss and small-scale fading [32]. In particular, we assume path-loss effect as: $L_k = d_k^{-\alpha}$ for UE k, $k \in \mathcal{K}$, where d_k denotes the distance between the k-th UE and the base station. While, each link also experiences independent Nakagami-m fading, denoted by h_k for UE k, $k \in \mathcal{K}$. Furthermore, the background noise for each link is characterized by a zero mean, complex Gaussian random variable with variance N_0.

Moreover, within each NOMA group, the SIC decoding is employed. Let us assume that, the decoding order of the NOMA UEs always follows the descending order of channel qualities, and the F-UE's information is always decoded at the last stage. For a NOMA group, the nF-UEs with poor channel qualities are free of interference from the nF-UEs with good channel qualities. To achieve a good trade-off between the efficiency of energy and spectrum, we assume that, the nF-UEs with poor channel qualities need longer time interval for offloading, compared to the nF-UEs with good channel qualities.

B. Problem Formulation

For the NOMA-mmW based MEC system conceived, it aims to minimize the average delay of all UEs' offloading, where jointly considering the UE scheduling, beamwidth selection and transmit power allocation. The optimization problem can be expressed as

$$\text{P0:} \quad \min_{\boldsymbol{\Theta},\boldsymbol{S},\boldsymbol{P}} \quad \bar{\tau} = \tau^{\text{BA}} + \frac{\sum_{j\in\mathcal{J}} \tau_j^{\text{DT,F}} + \sum_{i\in\mathcal{I}} \tau_i^{\text{DT,nF}}}{K}, \tag{4}$$

$$\begin{aligned}
\text{s.t. (a) } & s_{i,j} = \{0,1\},\ \forall i \in \mathcal{I}, \forall j \in \mathcal{J},\\
\text{(b) } & \sum_{i\in\mathcal{I}} s_{i,j} \le Q,\ \forall j \in \mathcal{J},\\
\text{(c) } & \sum_{j\in\mathcal{J}} s_{i,j} = 1,\ \forall i \in \mathcal{I},\\
\text{(d) } & \theta^{\min} \le \theta_k \le \theta^{\max},\ \forall k \in \mathcal{K},\\
\text{(e) } & p_k \le p^{\max},\ \forall k \in \mathcal{K},\\
\text{(f) } & C_j / R_j^{\text{F}} = \Delta_0,\ \forall j \in \mathcal{J},\\
\text{(g) } & C_i / R_i^{\text{nF}} \le \Delta_0,\ \forall i \in \mathcal{I},\\
\text{(h) } & p_i \le p_{i'},\ \theta_i \ge \theta_{i'} \text{ for } E_i \le E_{i'},\ i \ne i', \forall i, i' \in \mathcal{I},\\
\text{(i) } & \tau^{\text{BA}} + \tau_i^{\text{DT,nF}} \le D_i,\ \forall i \in \mathcal{I},
\end{aligned}$$

where $s_{i,j}$ presents the scheduling variable, and $s_{i,j} = 1$ means that the i-th uF-UE is assigned to the j-th NOMA group; $\theta^{\min}$ and $\theta^{\max}$ denote the minimum and maximum beamwidth, respectively; $p^{\max}$ denotes the maximum transmit power of UEs; C_j and C_i denote the computing task of j-th F-UE and i-th uF-UE, respectively; Δ_0 denotes the maximum available transmission time for NOMA groups; E_i denotes the channel condition information (CSI) of i-th uF-UE. Above, in the objective function, the total delay comes from two parts: beam alignment and data transmission. First of all, all the K UEs sequentially carry out the beam alignment, hence, the delay can be accumulated as

$$\tau^{\text{BA}} = \sum_{k\in\mathcal{K}} \lceil \frac{\Upsilon_k}{\theta_k} \rceil T_p \approx \sum_{k\in\mathcal{K}} \frac{\Upsilon_k}{\theta_k} T_p. \tag{5}$$

Furthermore, the data transmission delay for an UE, such as F-UE j or nF-UE i, can be expressed as

$$\tau_j^{\text{DT,F}} = j\Delta_0,\ \forall j \in \mathcal{J}, \tag{6}$$

$$\tau_i^{\text{DT,nF}} = \sum_{j\in\mathcal{J}} s_{i,j}(j-1)\Delta_0 + \frac{C_i}{R_i^{\text{nF}}},\ \forall i \in \mathcal{I}. \tag{7}$$

According to NOMA-mmW scheme, the data rate for nF-UE i can be given by (8),

$$R_i^{\text{nF}} = B\log_2\left(1 + \frac{P_i^{nF}}{P_{i'}^{nF} + P_j^{F} + N_0}\right),\ i \in \mathcal{I}, \tag{8}$$

where P_i^{nF} is defined as $\sum_{j\in\mathcal{J}} s_{i,j} G_{\text{bs}} G_i p_i E_i$, which means the received power of nF-UE i; $P_{i'}^{nF}$ is defined as $\sum_{j\in\mathcal{J}} \sum_{i'\in\mathcal{I}, E_{i'}<E_i} s_{i,j} s_{i',j} G_{\text{bs}} G_{i'} p_{i'} E_{i'}$, which represents the interference of other nF-UEs; P_j^F is defined as $\sum_{j\in\mathcal{J}} s_{i,j} G_{\text{bs}} G_j p_j E_j$, which means the interference of F-UEs, and we define $E_x = L_x |h_x|^2,\ x \in \{i, i', j\}$.

By contrast, the data rate for F-user j can be written as

$$R_j^{\text{F}} = B\log_2\left(1 + \frac{G_{\text{bs}} G_j p_j |h_j|^2 L_j}{N_0}\right),\ j \in \mathcal{J}. \tag{9}$$

In problem (P0), the scheduling variables are collected in the vector $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_I]$, where defining $\boldsymbol{s}_i = [s_{i,j}, \forall t]$. Note that, $s_{i,j}$ is a boolean variable: only when $s_{i,j} = 1$, nF-UE j is allocated to the NOMA group of F-UE i scheduled in tth (assuming $t = j$) TS. The constraints (4b) and (4c) indicate that, the maximum number of UEs is Q, while each UE can be assigned to one group only. Furthermore, the beamwidth variables are included in the vector $\boldsymbol{\Theta} = [\theta_1, \theta_2, \ldots, \theta_K]$, and they are limited in the range of (4d). Then, the power allocation variables are collected in the vector $\boldsymbol{P} = [p_1, p_2, \ldots, p_K]$, and constraint (4e) gives the upper-bound for transmit power of each UE. In addition, (4f) and (4g) constrain that each UE's data transmission period should not be longer than Δ_0. While, the constraint in (4h) reflects that the nF-UEs with good channel qualities will be assigned more transmit power and narrower beamwidth. At last, we have the maximum delay requirement for each UE, which is shown in (4i).

C. Problem Analysis

Seen from (8), the rate expression is non-convex, since it contains the fractional terms of power allocation variable and beamwidth variable, as well as including the products of scheduling variables. Hence, in problem (P0), the objective function is obvious not convex, and the constraints (4e)-(4i) are not convex either. Furthermore, (4f) is a strict equality constraint, which is extremely difficult to deal with. At last, the scheduling variables are binary variables, which further complicates the problem. Therefore, problem (P0) cannot be converted to a convex problem, and is extremely difficult to solve. The following proposition is given.

Proposition 1: Problem (P0) for the joint optimization of UE scheduling, beamwidth, power allocation is NP-hard.

Proof: Problem (P0) can be decoupled into different sub-problems. For instance, given the beamwidth and power allocation, the sub-problem becomes a pure UE scheduling

problem (P1). In order to minimize the average delay, problem (P1) selects specific nF-UEs in each time slot, and can be seen as a more restrictive or general traveling salesman problem (TSP). As known, the traveling salesman problem can be reduced to problem (P1). Since the TSP is a classic NP-hard problem, problem (P1) is also a NP-hard problem. In that case, we can prove that the original problem (P0) is NP hard. ■

In problem (P0), we can find that, the UE scheduling variables are independent of beam alignment variables and power allocation variables. Hence, this observation motivates us to decouple the original problem (P0) into the UE scheduling sub-problem (P1) and the joint beam alignment and power allocation sub-problem (P2), which can be solved by the alternating optimization (AO) approach. Specifically, the two sub-problems can be expressed as:

$$\text{P1:} \quad \min_{\boldsymbol{S}^{\text{nF}}} \quad \{\bar{\tau}(\boldsymbol{S}) \mid \boldsymbol{\Theta}^*, \boldsymbol{P}^*\} \tag{10}$$
$$\text{s.t. } (4\text{a}) \ , \ (4\text{b}) \ , \ (4\text{c}),$$

and

$$\text{P2:} \quad \min_{\boldsymbol{\Theta}, \boldsymbol{P}} \quad \{\bar{\tau}(\boldsymbol{\Theta}, \boldsymbol{P}) \mid \boldsymbol{S}^*\} \tag{11}$$
$$\text{s.t. } (4\text{d}) \ , \ (4\text{e}) \ , \ (4\text{f}), \ (4\text{g}), \ (4\text{h}), \ (4\text{i}).$$

From the above, we can derive the following lemma.

Lemma 1: Upon using the AO approach to solve problems (P1), (P2), it is guaranteed to converge to an approximated optimal solution to problem (P0).

Proof: Seen from the problems (P1), (P2) and (P3), the objective function keeps the same, and hence, the feasible regions for the solution of problems (P1) (P2) are belong to problem (P0). During the nth iteration of the AO approach, the optimal solution of problem (P1) is given by $\{\boldsymbol{S}^{\text{nF}*}(n) \mid \widetilde{\boldsymbol{\Theta}}^*(n-1), \widetilde{\boldsymbol{P}}^*(n-1)\}$, which must be a feasible solution of problem (P2). The corresponding objective value is denoted by $\tilde{\tau}_1(n)$. Based on the above, we can find the optimal solution of problem (P2), expressed as $\{\boldsymbol{S}^{\text{nF}*}(n), \boldsymbol{\Theta}^*(n), \boldsymbol{P}^*(n)\}$, which gives the objective value $\tilde{\tau}_2(n)$. In this case, it is easily known that $\tilde{\tau}_2(n) \leq \tilde{\tau}_2(n)$. By repeating the above process, the value of the objective function must be decreasing as the number of iterations gets bigger, which proves the convergency of the solutions obtained by the AO approach. ■

For low-complexity design, we develop the AO-RA scheme, in which the UE scheduling, beamwith selection and power allocation are performed in an iterative approach to obtain a joint promising solution to problem (P0). On the premise that the UE's location information and current CSI are known, the base station can easily obtain the optimal resource allocation by using the proposed AO-RA scheme, which decomposes the original NP-hard problem (P0) into two simpler sub-problems: the UE scheduling problem and the beamwidth and power allocation problem. In particular, upon giving the beamwidth and power allocation solutions, the UE scheduling problem of (P1) can be solved by many-to-one matching method. In turn, upon fixing the UE scheduling solution, the beamwidth and power allocation problem of (P2) can be converted into a convex problem, thereby giving its optimal solution. Note that, the detailed procedure of our AO-RA scheme is provided in Section IV-C.

III. MC-M2OE Algorithm for UE scheduling

In this section, let us solve the UE scheduling problem. Since the variables in $\boldsymbol{S}^{\text{nF}}$ are integers, problem (P1) can not be directly solved by convex optimization approach. As seen, the UE scheduling problem can be interpreted as a many-to-one matching between different UEs, where the nF-UEs are regarded as the many-side, the F-UEs are refered to the one-side. Further, known from (8), as the size of a NOMA group changes, it will affect both the forms of the interferences in its group and those in other related groups. This implies our problem (P1) belongs to a many-to-one matching with externalities being peer effects. To solve the problem, we propose a novel low-complexity algorithm, namely matrix-control aided many-to-one matching with externality (MC-M2OE).

A. Modeling of Matching Algorithm

To comply with problem (P1), in our matching model, there are two disjoint sets, including the F-UE set $\mathcal{J}$, and the nF-UE set $\mathcal{I}$, which are the rational players aiming to maximize their own benefits. Specifically, a matching is an assignment of nF-UEs in $\mathcal{I}$ to F-UEs in $\mathcal{J}$, and can be defined in the following mathematical form.

Definition 1. Given two *disjoint sets*, i.e. the F-UE set $\mathcal{J}$, and the nF-UE set $\mathcal{I}$, a *many-to-one matching* is a mapping from the set $\mathcal{J} \cup \mathcal{I}$ to its all subsets, such that:
1) $j \in \Psi(i) \Leftrightarrow i \in \Psi(j)$;
2) $\Psi(j) \subseteq \mathcal{I}$, $\Psi(i) \subseteq \mathcal{J}$;
3) $|\Psi(i)| \leq Q$, $|\Psi(j)| \leq 1$.
Above, in 1), it indicates that nF-UE j is matched with F-UE i, where the two equivalent expressions are given; In 2), a nF-UE (or a F-UE) is matched with a subset of $\mathcal{I}$ (or $\mathcal{J}$); In 3), it constrains the maximum *quotas* for matching, determined by (4b) and (4c) in problem (P1).

Known from the objective function in problem (P1), the utility of a nF-UE can be defined as the delay of its data transmission, giving that

$$U_j^{\text{nF}} := -(\Psi(j) - 1)\Delta_0 - \frac{C_j}{R_j^{\text{nF}}}, \quad j \subseteq \Psi(i). \tag{12}$$

By contrast, since each F-UE may share its spectrum with multiple nF-UEs based on NOMA scheme, its utility should reflect the delay of all UEs' data transmission in the NOMA group, which can be defined that

$$U_i^{\text{F}} := -j\Delta_0 - |\Psi(i)|(i-1)\Delta_0 - \sum_{l \in \Psi(j)} \frac{C_l}{R_l^{\text{nF}}}, \quad j \subseteq \Psi(i). \tag{13}$$

Furthermore, the total utility for a matching Ψ is defined by collecting all the NOMA groups' utility, given that

$$U_{\sum}(\Psi) := \sum_{i \in \mathcal{I}} U_i^{\text{F}}, \quad j \subseteq \Psi(i). \tag{14}$$

To introduce the stability of a matching, we are motivated to define two kinds of matching actions for our specific model.

Definition 2. **Swap action**: nF-UE j matched with F-UE i and nF-UE j' matched with F-UE i' switch their NOMA groups but not violating their delay requirements, and all other matchings remain the same, such that: $\Psi^{sw}(j) = \{i'\}$, $\Psi^{sw}(j') = \{i\}$, $\Psi^{sw}(i) = \{\Psi(i)\backslash j\} \cup \{j'\}$, $\Psi^{sw}(i') = \{\Psi(i')\backslash j'\} \cup \{j\}$, subject to (4i).

Definition 3. **Join-in action**: nF-UE j leaves its current NOMA group of F-UE i, and joins in another NOMA group of F-UE i' but not violating the delay requirement of nF-UE j, and all other matchings remain the same, such that: $\Psi^{jo}(j) = \{i'\}$, $\Psi^{jo}(i) = \Psi(i)\backslash\{j\}$, $\Psi^{jo}(i') = \Psi(i') \cup \{j\}$, subject to (4i).

During our matching process, each nF-UE needs to find its best NOMA group by trying all possible swap and join-in actions. Nevertheless, the swap and/or join-in actions can be approved only when the involved players' interests are improved. For this sake, we would like to introduce the notion of "*blocking group*", which indicates the condition that approves a certain action.

Definition 4. Given a matching Ψ, including $j \in \Psi(i)$, $j' \in \Psi(i')$, if nF-UEs j and j' switches their NOMA groups such that:
$U_i^{\mathrm{F}}(\Psi) + U_{i'}^{\mathrm{F}}(\Psi) > U_i^{\mathrm{F}}(\tilde{\Psi}^{sw}) + U_{i'}^{\mathrm{F}}(\tilde{\Psi}^{sw})$, subject to (4i),
then the swap action is approved, and is termed as "*swap-blocking action*", where (i, j, i', j') is defined as a blocking group. By contrast, if nF-UE j leaves F-UE i and joins in F-UE i' such that:
$U_i^{\mathrm{F}}(\Psi) + U_{i'}^{\mathrm{F}}(\Psi) > U_i^{\mathrm{F}}(\tilde{\Psi}^{jo}) + U_{i'}^{\mathrm{F}}(\tilde{\Psi}^{jo})$, subject to (4i),
then the join-in action is approved, and is termed as "*join-in-blocking action*", where (i, j, j') is defined as a blocking group.

Based on the above definition, we can evaluate how the UEs evolve their matching behivours with peer effects. For each nF-UE, it can propose a swap action or a join-in action to block the current matching status, the related F-UE will check if it is approved. During our matching process, a swap action or a join-in action is approved as long as the utilities of the NOMA groups involved are increased, i.e. the delays for the corresponding data transmission are decreased towarding minimizing the objective function in (4). Note that, once a blocking action is approved, the dynamic preferences of different nF-UEs and F-UEs associated will be changed based on evaluating the externalities of the interferences within each NOMA group. The UEs keep executing approved swap and join-in actions so as to reach a stable status.

Definition 5. A matching Ψ is "*stable*" if and only if there does not exist a blocking action.

The notion of "*stable*" implies all the blocking actions involved for all the UEs are indifferent, which is similar to [33].

B. Principles of MC-M2OE Algorithm

As the preliminaries definitions given above, we can now introduce the principles of the proposed MC-M2OE algorithm. Different from the traditional matching algorithm, our MC-M2OE algorithm develops a high-efficiency matrix control scheme, so as to obtain the stable matching results with a relative low number of iterations guaranteed.

The detailed principles of the MC-M2OE algorithm are summarized in Algorithm 1. As seen, there are three key phases, including initialization phase, action phase, and update phase. The key idea of our MC-M2OE algorithm is to find the best NOMA group of F-UE for each nF-UE in an iterative process. For this sake, we introduce two control matrices: 1) the control matrix C, for indicating where the swap and join-in actions can take place, and 2) the assignment matrix A, for recording the current matching status. Note that, both matrices C and A have I rows and J columns, which correspond to nF-UE and F-UE indexes. For matrix C, we define that, if $C(i, j) = 1$, nF-UE i and F-UE j are available for taking actions, otherwise, they are not available if $C(i, j) = 0$. While, for matrix A, we define that, if $A(i, j) = 1$, nF-UE i is currently matched with F-UE j, otherwise, they are not matched if $A(i, j) = 0$.

Specifically, in the initialization phase, the initial UE scheduling is obtained by the random approach, and the initial assignment matrix A is written accordingly. Then, the control matrix C is initialized as $C = 1 - A$, where the non-zero elements in C indicate the availability of taking actions. As shown in Table I, the algorithm carries out the matching in an iterative process of one nF-UE by one, and terminates if $C = 0$. During the action phase, for each nF-UE i, it traverses all possible swap or join-in action according to the non-zero elements in the ith row of the control matrix C. Then, the action with the highest utility is always selected. It is worth noting that the nF-UE cannot perform the join-in or swap action within its current NOMA group.

Algorithm 1: MC-M2OE algorithm for UE scheduling

Initialization Phase:
Set the assignment matrix A by random approach; Set the control matrix $C = 1 - A$; Set the temporary control matrix $TempC = C$;
While C has non-zero elements
 For each nF-UE i, **do**
  If $C(i, :)$ has non-zero elements
   $TempC(i, :) \leftarrow C(i, :)$, then $C(i, :) \leftarrow 0$.
   According to $TempC$, calculate $U_{\sum}$ in (14) by trying all possible join-in and swap actions;
   Action Phase: Select the action with the maximum $U_{\sum}$
   If a **join-in** action, **then**, $A(i, j) \leftarrow 0$, $A(i, j') \leftarrow 1$;
   Else a **swap** action, **then** $A(i, j) \leftarrow 0$, $A(i, j') \leftarrow 1$, $A(i', j) \leftarrow 1$, $A(i', j') \leftarrow 0$;
   Update Phase:
   For join-in: $C(i, \hat{j}) \leftarrow TempC(i, \hat{j})$, $\forall \hat{j} \in \{\mathcal{J}\backslash(j, j')\}$;
   $C(\hat{i}, j) \leftarrow 1$, $C(\hat{i}, j') \leftarrow 1$, $\forall \hat{i} \in \{\mathcal{I}\backslash i, A(\hat{i}, j) = 1\}$;
   For swap: $C(i, \hat{j}) \leftarrow TempC(i, \hat{j})$, $\forall \hat{j} \in \{\mathcal{J}\backslash(j, j')\}$;
   $C(i', j) \leftarrow 0$, $C(i', j') \leftarrow 0$;
   $C(\tilde{i}, j) \leftarrow 1$, $C(\tilde{i}, j') \leftarrow 1$, $\forall \tilde{i} \in \{\mathcal{I}\backslash\{i, i'\}, A(\tilde{i}, j) = 1\}$;
  End if
 End for
End

During the update phase, the control matrix is updated

by recording all the actions taken. When a join-in action is performed, e.g., nF-UE i leaves its current NOMA group of F-UE j, and joins in the group of F-UE j', the corresponding elements are disabled, i.e. $C(i,j) = C(i,j') = 0$, except which the other elements become enable for taking actions and are updated to 1. When a swap action is performed, e.g., nF-UE i matched F-UE j and nF-UE i' matched F-UE j' swap their NOMA groups, the corresponding elements are updated to $C(i,j) = C(i,j') = C(i',j) = C(i',j') = 0$. Similarly, the other nF-UEs ($\forall i$ in Algorithm 1) in groups j and j' become available for taking actions. Note further that, $TempC$ in the table is used to record the non-zero elements in C.

C. Characteristics Analysis

Let us analyze the main characteristics of the proposed MC-M2OE algorithm, which include the stability and the complexity required. As shown in Algorithm 1, the proposed MC-M2OE algorithm terminates when there is no blocking group for improving the total utility. Let us assume that, the final matching is denoted by Ψ^*. In other words, for each nF-UE i^*, it can not leave its current NOMA group of F-UE j^* to join-in another NOMA group, or exchange group with another nF-UE. Based on the above fact, we have the following theorem.

Theorem 1: The proposed MC-M2OE algorithm can converge to a stable matching Ψ^* with a finite number of iterations required.

Proof: Within the action phase of the algorithm, each nF-UE needs to test the actions available in the matrix C, which has finite number of rows and columns. Hence, the number of actions for each UE is limited. Furthermore, once the possible actions for all nF-UEs have been executed, C becomes zero-valued, and the algorithm terminates. Therefore, the number of iteration is also finite. At last, since only the join-in and swap actions those strictly improve the total utility can be performed, each nF-UE can be allocated to the best NOMA group. Based on the above facts, the MC-M2OE algorithm can converge to a stable matching Ψ^* in finite iterations. ■

Based on the stability of the MC-M2OE, we are now able to carry out its complexity analysis. Let us first define the notion of complexity involved in this paper. As the number of iterations may be different under different cases, the complexity is computed always for the worst scenarios. During the matching process for each nF-UE, we assume M is the upper-bound for the number of combinations of NOMA groups, where either join in action or swap action is executed. In the worst scenarios, the optimal solution of the NOMA group for a nF-UE always conflicts with other matched group during each loop. Hence, each F-UE needs to change the matched group at most M times. Moreover, each F-UE will update the preference lists with a time complexity of $\mathcal{O}(MlogM)$. Then, the time complexity of the MC-M2OE algorithm is $\mathcal{O}(JM + JMlogM)$.

IV. Joint Optimization of Beamwidth and Transmit Power

To solve the sub-problem (P2), we propose an algorithm, namely JBTP, for joint optimization of beamwidth and transmit power.

A. Design of JBTP

Once having the UE scheduling solutions, the optimization for beamwidth and transmit power in problem (P2) can be re-written as

$$\text{P3:} \quad \min_{\boldsymbol{\Theta},\boldsymbol{P}} \quad \sum_{k\in\mathcal{K}} \frac{\Upsilon_k}{\theta_k} T_p + \frac{1}{K}\sum_{j\in\mathcal{J}} j\Delta_0 + \frac{1}{K}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{F}_j}\left[(j-1)\Delta_0 + \frac{C_i}{R_i^{\text{nF}}}\right], \tag{15}$$
$$\text{s.t. (4d) , (4e) , (4f), (4g), (4h), (4i).}$$

Above, we define that:

$$\mathcal{F}_j = \{i \mid s^*_{i,j} = 1,\ \forall i \in \mathcal{I}\},\ j \in \mathcal{J}, \tag{16}$$

which includes the indexes of the nF-UEs associated with the NOMA group of F-UE j. Note that, $\boldsymbol{S}^* = \{s^*_{i,j}, \forall i, j\}$ are the UE scheduling solutions obtained by the MC-M2OE algorithm. Furthermore, the data rate expression in (8) for the nF-UE becomes

$$R_i^{\text{nF}} = B\log_2\left(1 + \frac{\hat{P}_i^{nF}}{\hat{P}_{i'}^{nF} + hatP_j^F + N_0}\right),\ i \in \mathcal{F}_j. \tag{17}$$

where $\hat{P}_i^{nF}$ is defined as $G_{\text{bs}}G_i p_i E_i$; $\hat{P}_{i'}^{nF}$ is defined as $\sum_{i'\in\mathcal{I}, E_{i'}<E_i} G_{\text{bs}}G_{i'}p_{i'}E_{i'}$; $\hat{P}_j^F$ is defined as $G_{\text{bs}}G_j p_j E_j$.

Apparently, problem (P3) is still difficult to be directly solved, since both the objective function and all the constraints except (4e) are nonconvex. To make the problem solvable, we need to convert the objective function and all the constraints into convex ones, which aim to be equivalent.

1) Tackling Objective Function: In order to tackle the problem, we need to convert the objective function and the constraints in problem (P3) into convex. Seen from (17), the data rate of nF-UE i contains the summation term of $p_{i'}/\theta_{i'}$ in the denominator. This causes the nonconvexity of the objective function in (15) and the constraint in (4i). Let us introduce the auxiliary variables to replace the numerator and denominator of the SINR. In particular, we define the auxiliary variables α_i^0 and α_i^1 for the objective function. In that case, the optimization problem (P3) can be equivalently converted into:

$$\text{P4:} \quad \min_{\boldsymbol{\Theta},\boldsymbol{P}} \quad \sum_{k\in\mathcal{K}} \frac{\Upsilon_k}{\theta_k} T_p + \frac{1}{K}\sum_{j\in\mathcal{J}} j\Delta_0 + \frac{1}{K}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{F}_j}\left[(j-1)\Delta_0 + \frac{C_i}{B(\log_2(\alpha_i^0) - \log_2(\alpha_i^1))}\right], \tag{18}$$

$$\text{s.t. (4d) , (4e) , (4f), (4g), (4h), (4i),}$$

$$\text{(a)}\ G_{\text{bs}}(\frac{Z_0}{\theta_i} + z_0)p_i E_i + \sum_{i'\in\mathcal{F}_j, E_{i'}<E_i} G_{\text{bs}}(\frac{Z_0}{\theta_{i'}} + z_0)p_{i'}E_{i'} + E_j(\frac{Z_0}{\theta_j} + z_0)p_j + N_0 \geq \alpha_i^0,\ ,\ \forall i \in \mathcal{F}_j, j \in \mathcal{J},$$

$$\text{(b)} \sum_{i'\in\mathcal{F}_j, E_{i'}<E_i} G_{\text{bs}}(\frac{Z_0}{\theta_{i'}} + z_0)p_{i'}E_{i'} + E_j(\frac{Z_0}{\theta_j} + z_0)p_j + N_0 \leq \alpha_i^1,\ ,\ \forall i \in \mathcal{F}_j, j \in \mathcal{J}.$$

where defining $Z_0 = 2\pi(1 - z_0)$. Note that, (18a) and (18b) respectively constrain the upper and lower bound of the auxiliary variables.

Seen from the objective function in (18), all the terms except the term of $-\log_2(\alpha_i^1)$ are now convex. In this case, let us apply the first order Taylor expansion to the term, thereby giving the following proposition.

Proposition 2: The logarithmic term $\log_2(\alpha_i^1)$ can be approximated by

$$\log_2(\alpha_i^1) \leq \Omega(\alpha_i^1) = \log_2(\tilde{\alpha}_i^1) + \frac{\alpha_i^1 - \tilde{\alpha}_i^1}{\tilde{\alpha}_i^1 \ln 2}, \forall i \in \mathcal{I}, \quad (19)$$

where $\tilde{\alpha}_i^1$ is a constant. The bound in the above inequality becomes tight for $\alpha_i^1 = \tilde{\alpha}_i^1$.

Proof: This can be found in [34]. ■

By substituting (19) into (18), the objective function becomes a convex one. Let us now consider how to deal with the constraints.

2) Tackling Equality Constraint: The constraint in (4f) means that, each F-UE is scheduled to a specific time slot to complete its MEC offloading. However, (4f) is an equality constraint, and can not be directly transformed into a convex form. In this case, the best option is to transformed the equality constraint in to a pair of inequality constraints. Let us introduce a parameter ε to set up an interval for the offloading period of each F-UE. In this case, the constraint in (4f) becomes

$$\Delta_0 - \varepsilon \leq C_j / R_j^{\mathrm{F}} \leq \Delta_0, \ \forall j \in \mathcal{J}. \quad (20)$$

Remark 1: The inequality constraints in (20) can closely approximate the equality constraint (4f) as long as the parameter ε persists a relatively small value. Furthermore, when the upper-bound holds, the constraint (20) will be equivalent to (4f).

By substituting (9) into (20), it derives

$$\left(\frac{Z_0}{\theta_j} + z_0\right) p_j \geq \frac{A_j^0 N_0}{G_{bs} E_j}, \ \forall j \in \mathcal{J}, \quad (21)$$

$$\left(\frac{Z_0}{\theta_j} + z_0\right) p_j \leq \frac{A_j^1 N_0}{G_{bs} E_j}, \ \forall j \in \mathcal{J}, \quad (22)$$

where defining $A_j^0 = 2^{\frac{C_j}{B(\Delta_0 - \varepsilon)}} - 1$, and $A_j^1 = 2^{\frac{C_j}{B\Delta_0}} - 1$. Seen from above, it readily finds that, the fractional terms of p_j/θ_j determine the non-convexity of the constraints (21) and (22). To deal with this, let us introduce the auxiliary variable ϱ_j to set up an upper bound of the inverse of transmit power variable p_j, giving that

$$1/p_j \leq \varrho_j, \ \forall j \in \mathcal{J}. \quad (23)$$

By replacing the varibale p_j with ϱ_j, the constraints (21) and (22) are converted to

$$\frac{Z_0}{\theta_j \varrho_j} + \frac{z_0}{\varrho_j} \geq \frac{A_j^0 N_0}{G_{bs} E_j}, \ \forall j \in \mathcal{J}, \quad (24)$$

$$\frac{Z_0}{\theta_j \varrho_j} + \frac{z_0}{\varrho_j} \leq \frac{A_j^1 N_0}{G_{bs} E_j}, \ \forall j \in \mathcal{J}. \quad (25)$$

Apparently, the constraint (25) is now convex. Whereas, the constraint (24) is not convex with respect to the specific domain of variables. Hence, we need to further manipulate (24) by applying the approach of the first-order Taylor expansion, in order to guarantee the strict convexity.

Proposition 3: The fractional terms $1/\theta_j \varrho_j$ and $1/\varrho_j$ can be respectively approximated by

$$\frac{1}{\theta_j \varrho_j} \geq \Theta(\theta_j, \varrho_j) = \frac{1}{\tilde{\theta}_j \tilde{\varrho}_j} - \frac{\theta_j - \tilde{\theta}_j}{\tilde{\theta}_j^2 \tilde{\varrho}_j} - \frac{\varrho_j - \tilde{\varrho}_j}{\tilde{\theta}_j \tilde{\varrho}_j^2}, \ \forall j \in \mathcal{J}, \quad (26)$$

$$\frac{1}{\varrho_j} \geq \psi(\varrho_j) = \frac{1}{\tilde{\varrho}_j} - \frac{\varrho_j - \tilde{\varrho}_j}{\tilde{\varrho}_j^2}, \ \forall j \in \mathcal{J}, \quad (27)$$

where $\tilde{\theta}_j$, $\tilde{\varrho}_j$ are constants. The bound in the above inequalities become tight for $\theta_j = \tilde{\theta}_j$, and $\varrho_j = \tilde{\varrho}_j$, respectively.

Proof: This can be found in [35]. ■

At this stage, the optimization problem (P4) can be transformed into:

$$\text{P5:} \quad \min_{\boldsymbol{\Theta}, \boldsymbol{P}} \quad \sum_{k \in \mathcal{K}} \frac{\Upsilon_k}{\theta_k} T_p + \frac{1}{K} \sum_{j \in \mathcal{J}} j \Delta_0 + \frac{1}{K} \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{F}_j} \left[(j-1)\Delta_0 + \frac{C_i}{B(\log_2(\alpha_i^0) - \Omega(\alpha_i^1))} \right], \quad (28)$$

s.t. (4d), (4h), (4i), (18a), (18b), (19), (23), (24), (26), (27), and

(a) $\varrho_j \geq 1/p^{\max}, \ p_i \leq \ p^{\max}, \ \forall i \in \mathcal{F}_j, \ \forall j \in \mathcal{J}$,

(b) $\dfrac{C_i}{R_i^{\mathrm{nF}}} \leq \Delta_0 - \varepsilon, \ \forall i \in \mathcal{F}_j, j \in \mathcal{J}$,

(c) $Z_0 \Theta(\theta_j, \varrho_j) + z_0 \psi(\varrho_j) \geq \dfrac{A_j^0 N_0}{G_{bs} E_j}, \ \forall j \in \mathcal{J}$.

In the above problem, (28a) constrains the maximum transmit power for each F-UE, due to p_j is replaced by the inverse of ϱ_j. As shown by (28a), each nF-UE's delay is now restricted by the upper-bound shrink a tiny interval from the original Δ_0. Last, (28c) is an approximated form of the constraint (24), which can be known from *Proposition* 3.

3) Tackling Nonconvex inequality Constraints: Seen from problem (P5), we find that, the convexities are guaranteed for the objective function, as well as for the constraints (4d), (4h), (19), (23), (24), (26), (27), (28a), and (28c). Hence, let us now tackle the nonconvex inequality constraints remaining in the problem.

Let us first deal with the constraint (18a) and (18b). It is worth noting that, there exists the fractional term p_i/θ_i, which can be tackled by the approach similar to that for p_j/θ_j. Hence, we further introduce the auxiliary variable $\varrho_i \geq 1/p_i$, thereby converting (18b) into:

$$\sum_{i' \in \mathcal{F}_j, E_{i'} < E_i} G_{\mathrm{bs}} \left(\frac{Z_0}{\theta_{i'} \varrho_{i'}} + \frac{z_0}{\varrho_{i'}} \right) E_{i'} + G_{\mathrm{bs}} \left(\frac{Z_0}{\theta_j \varrho_j} + \frac{z_0}{\varrho_j} \right) E_j \leq \alpha_i^1, \quad (29)$$

where $\forall i \in \mathcal{F}_j, j \in \mathcal{J}$, which becomes a convex constraint. For convenience, ξ_0 denotes the left-hand side of the inequality

(29). By contrast, although applying the transformation of (29), (18a) persists non-convexity due to its feasible region defined by being bigger than α_i^0. Therefore, the linearity operation provided in *Proposition* 3 should be employed. As a result, the constraint (18a) can be transformed to

$$\sum_{\hat{ij}\in\{i,j\}} G_{\text{bs}}(Z_0\Theta(\theta_{\hat{ij}},\varrho_{\hat{ij}})+z_0\psi(\varrho_{\hat{ij}}))E_{\hat{ij}}+N_0+ \sum_{i'\in\mathcal{F}_j,E_{i'}<E_i} G_{\text{bs}}(Z_0\Theta(\theta_{i'},\varrho_{i'})+z_0\Psi(\varrho_{i'}))E_{i'}\geq\alpha_i^0, \quad (30)$$

where $\forall i\in\mathcal{F}_j, j\in\mathcal{J}$, which is now a convex constraint. For convenience, we denote the left-hand side of the inequality (30) as ξ_1.

For (28b), we now replace the variables $1/p_i$, $1/p_j$ with the auxiliary variables ϱ_i and ϱ_j, thereby giving

$$\underbrace{G_{\text{bs}}\left(\frac{Z_0}{\theta_i\varrho_i}+\frac{z_0}{\varrho_i}\right)E_i}_{\Xi_0}-\underbrace{A_i^1\xi_0}_{\Xi_1}\geq A_i^1N_0, \quad (31)$$

where the first term Ξ_0 is non-convex. Let us apply the linearity operation in *Proposition* 3, giving the convex form:

$$G_{\text{bs}}(Z_0\Theta(\theta_i,\varrho_i)+z_0\psi(\varrho_i))E_i-\Xi_1\geq A_i^1N_0. \quad (32)$$

At last, the similar approach for tackling the objective function can be applied to the constraint (4i). Let us define the auxiliary variables β_i^0 and β_i^1 to transform the rate expression of the nF-UEs. Then, it needs to bound the SINR of the nF-UEs, where the similar approximation approaches are used. In that case, our optimization problem can be eventually converted to a convex form, expresses as

$$\text{P6:}\quad \min_{\boldsymbol{\Theta},\boldsymbol{P}}\quad \sum_{k\in\mathcal{K}}\frac{\Upsilon_k}{\theta_k}T_p+\frac{1}{K}\sum_{j\in\mathcal{J}}j\Delta_0 + \frac{1}{K}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{F}_j}\left[(j-1)\Delta_0+\frac{C_i}{B(\log_2(\alpha_i^0)-\Omega(\alpha_i^1))}\right], \quad (33)$$

s.t. (4d), (29), (30), (25), (28c), (32), and

(a) $\varrho_k\geq 1/p^{\max},\ \forall k\in\mathcal{K}$,

(b) $\varrho_i\leq\varrho_{i'},\ \theta_i\leq\theta_{i'}$ for $E_i\geq E_{i'},\ i\neq i',\forall i,i'\in\mathcal{I}$,

(c) $\sum_{k\in\mathcal{K}}\frac{\Upsilon_k}{\theta_k}T_p+(j-1)\Delta_0+\frac{C_i}{B(\log_2(\beta_i^0)-\Omega(\beta_i^1))} \leq D_i,\ \forall i\in\mathcal{F}_j,\ \forall j\in\mathcal{J}$,

(d) $\xi_1\geq\beta_i^0,\ \forall i\in\mathcal{F}_j, j\in\mathcal{J}$,

(e) $\xi_0\leq\beta_i^1,\ \forall i\in\mathcal{F}_j, j\in\mathcal{J}$.

Now, (P6) is a convex optimization problem, which can be solved by CVX toolbox.

B. Principles of JBTP algorithm

In this subsection, we introduce the principles of the proposed JBTP algorithm, which are summarized in Algorithm 2, where $\forall i\in\mathcal{I},\forall k\in\mathcal{K}$.

Algorithm 2: JBTP algorithm for solving problem (P2)

Initialization: **1)** Set iteration index $n=0$; **2)** Set maximum iteration number N_{max}; **3)** Set $0<\epsilon<0.001$;
4) Initialize parameters $\tilde{\alpha}_i^{1(0)}$, $\tilde{\beta}_i^{1(0)}$, $\tilde{\theta}_k^{(0)}$, and $\tilde{\varrho}_k^{(0)}$;
5) Initialize variables $\boldsymbol{\alpha}^{\mathbf{0}(0)},\boldsymbol{\beta}^{\mathbf{0}(0)},\boldsymbol{\alpha}^{\mathbf{1}(0)},\boldsymbol{\beta}^{\mathbf{1}(0)},\boldsymbol{\Theta}^{(0)},\boldsymbol{P}^{(0)}$;
While $n<N_{max}$ & $|\tau^n-\tau^{n-1}|>\epsilon$ **do**
  Solve the convex optimization problem (P6) by CVX toolbox;
  Output: the solutions in nth iteration $\boldsymbol{\alpha}^{\mathbf{0}(n)},\boldsymbol{\beta}^{\mathbf{0}(n)},\boldsymbol{\alpha}^{\mathbf{1}(n)},\boldsymbol{\beta}^{\mathbf{1}(n)},\boldsymbol{\Theta}^{(n)},\boldsymbol{P}^{(n)}$;
  Compute the objective value $\tilde{\tau}(\boldsymbol{\alpha}^{\mathbf{0}(n)},\boldsymbol{\alpha}^{\mathbf{1}(n)},\boldsymbol{\Theta}^{(n)})$ according to (33);
  Update the parameters for problem (P6): $\tilde{\alpha}_i^{1(n+1)}\leftarrow\alpha_i^{1(n)}$, $\tilde{\beta}_i^{1(n+1)}\leftarrow\beta_i^{1(n)}$, $\tilde{\theta}_k^{(n+1)}\leftarrow\theta_k^{(n)}$, $\tilde{\varrho}_k^{(n+1)}\leftarrow\varrho_k^{(n)}$, $n\rightarrow n+1$;
End
Output: The optimal solution $\{\boldsymbol{\Theta}^*,\boldsymbol{P}^*\}$.

As shown by Algorithm 2, it is necessary to find the appropriate initial values of the parameters $\tilde{\alpha}_i^{1(0)}$, $\tilde{\beta}_i^{1(0)}$, $\tilde{\theta}_k^{(0)}$, and $\tilde{\varrho}_k^{(0)}$ according to the relevant constraints of problem (P6). During each iteration, we can apply the CVX toolbox to solve the convex problem of (P6), thereby obtaining the optimal solution $\{\boldsymbol{\alpha}^{\mathbf{1}(n)},\boldsymbol{\beta}^{\mathbf{1}(n)},\boldsymbol{\Theta}^{(n)},\boldsymbol{P}^{(n)}\}$. At the end of each iteration, it needs to update the parameters $\tilde{\alpha}_i^{1(n+1)}$, $\tilde{\beta}_i^{1(n+1)}$, $\tilde{\theta}_k^{(n+1)}$, and $\tilde{\varrho}_k^{(n+1)}$, $\forall i\in\mathcal{I},\forall k\in\mathcal{K}$. By repeating solving problem (P6) until the objective function converges, the approximate optimal solution $\{\boldsymbol{\Theta}^*,\boldsymbol{P}^*\}$ of problem (P2) can be finally obtained.

C. Principles of AO-RA scheme

Algorithm 3: AO-RA scheme for solving problem (P0)

Initialization: **1)** Set iteration index $m=0$; **2)** Set $\tilde{\epsilon}>0$; **3)** Set feasible values for $\{\tilde{\boldsymbol{\Theta}}^{(m)},\tilde{\boldsymbol{P}}^{(m)}\}$ in (P0);
While $|\tilde{\tau}^m-\tilde{\tau}^{m-1}|>\tilde{\epsilon}$ **do**
  Step 1: Give $\{\tilde{\boldsymbol{\Theta}}^{(m)},\tilde{\boldsymbol{P}}^{(m)}\}$, run Algorithm 1,
  Output: The assignment matrix $\boldsymbol{S}^{\text{nF}*(m)}$;
  Step 2: Given $\boldsymbol{S}^{\text{nF}*(m)}$, run Algorithm 2,
  Output: The beamwidth and power $\{\boldsymbol{\Theta}^{*(m)},\boldsymbol{P}^{*(m)}\}$, and the objective value $\tilde{\tau}^m$;
  Step 3: Update the beamwidth & transmit power for Algorithm 1, $\tilde{\boldsymbol{\Theta}}^{(m+1)}\leftarrow\boldsymbol{\Theta}^{*(m)}$, $\tilde{\boldsymbol{P}}^{(m+1)}\leftarrow\boldsymbol{P}^{*(m)}$; $m\rightarrow m+1$;
End
Output: the optimal solution: $\boldsymbol{S}^{\text{nF}*},\boldsymbol{\Theta}^*,\boldsymbol{P}^*$.

Let us now introduce the principles of the AO-RA scheme developed for solving the original problem (P0), where the UE scheduling, beamwidth and transmit power are jointly optimized. As shown in Algorithm 3, based on the rationale of AO approach, for iteration n, we need to solve problem (P1) by the MC-M2OE algorithm, giving the solutions $\boldsymbol{A}^*(n)$.

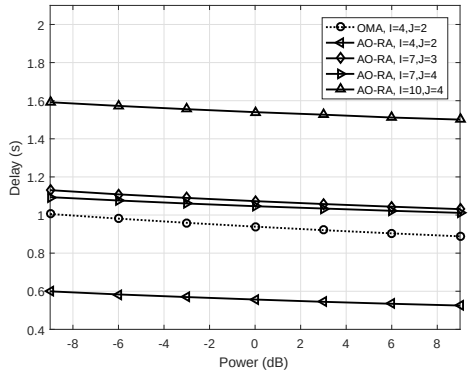

Fig. 2. Average delay performance of the offloading transmission employing the proposed AO-RA scheme.

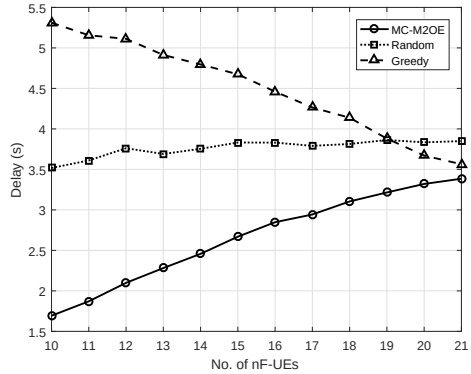

Fig. 3. Performance comparison of different UE scheduling methods for NOMA grouping when fixing $J = 7$.

Based on the above, we in turn find the optimal solution of problem (P6), resulting in the solutions $\boldsymbol{\Theta}^*(n)$ and $\boldsymbol{P}^*(n)$. By repeating the above process, the final solution can be obtained when the objective function of average delay is not improved in demand. After our careful study, we find that, Algorithm 3 only needs a very few number of iterations (less than 10) to converge.

V. Numerical Results

In this section, we provide a range of simulation results for demonstrating the achievable delay performance of the NOMA-mmW MEC systems employing the proposed algorithms. In our simulations, we consider different number of F-UEs and nF-UEs, in order to study the joint impact of UE scheduling, beamwidth selection and transmit power. In addition, the key simulation parameters include: 1) The bandwidth available is $B = 100$MHz; 2) Quota of each NOMA group $Q = 3$; 3) Beamwidth of each mmW link is constrained by $\pi/36 < \theta < \pi/6$; 4) Maximum delay for each nF-UE is $D_i = 4.5\text{s}, \forall i$; 5) All mmW links follow Nakagami-m fading with $m = 5$; 6) Path loss parameter $\alpha = 2.1$. The OMA scheme in this section refers to using a greedy algorithm to group all UEs according to time slots, and the UEs in the same group do not interfere with each other.

Fig. 2 investigate the delay performance of the proposed AO-RA scheme and the OMA scheme when considering different number of F-UEs and nF-UEs, while varying maximum transmit power. As we proved, the proposed AO-RA scheme can always find the promising solutions of UE scheduling, beamwidth and transmit power for the NOMA-mmW based MEC offloading. First of all, we can see that, as the number of UEs gets bigger, the average delay required by the AO-RA scheme is increased, which is due to the ICI becomes severer within the NOMA groups. It can be seen that the performance of the case "$I = 7, J = 3$" is lower than that of the case "$I = 7, J = 4$". This is simply due to that, when there are more F-UEs (i.e. more NOMA groups), our AO-RA scheme will achieve a higher degree of selecting freedom for UE scheduling. In that case, it will in turn increases the probability of reducing the offloading transmission time for more nF-UEs, which therefore results in decreasing the average delay. Nevertheless, if keep increasing the number of F-UEs, the delay time spend on beam alignment will become higher, and would be a dominant factor in minimizing the average delay. From the above, we may conclude that, the NOMA-mmW based MEC offloading needs to carefully balance the selecting freedom of UE scheduling and mmW's beam alignment overhead. Furthermore, the performance gap between "$I = 7, J = 3$" and "$I = 7, J = 4$" is much smaller than the gap between "$I = 7, J = 4$" and "$I = 10, J = 4$". This implies that, the optimal solutions always prefer to assign unbalanced number of nF-UEs to different NOMA groups. The chances are that, the NOMA groups scheduled with higher priorities have less probabilities of being outdated, then are assigned with heavier loading of nF-UEs. Furthermore, we also observe that, when more transmit power can be invested, the delay for offloading transmission can be slightly reduced. From the above observations, we may conclude that, the practical system design for NOMA-mmW based MEC offloading should carefully carry out the UE scheduling, rather than simply investing more transmit power causing unnecessary energy waste. In addition, when "$I = 4, J = 2$", it can be seen that the performance of the OMA scheme is lower than that of our AO-RA scheme, which confirms that the AO-RA scheme can utilize the time-frequency resources more efficiently and achieve lower average delay.

In order to study the impact of UE scheduling, we provide Fig. 3 to compare the proposed M2OE algorithm with the existing algorithms, where fixing the power allocation and beamdwidth solutions. Note that, in the context of the greedy algorithm, each NOMA group selects the nF-UEs with the best channel qualities from the current available options, and it is always fully loaded as long as there are nF-UEs available to join in. By contrast, the random method refers to assign similar number of nF-UEs to each NOMA group, where the nF-UEs are selected randomly. As observed, our M2OE can always achieve the best NOMA grouping results, and significantly outperforms the other two methods. However, we can see that, as the number of nF-UEs gets bigger, the M2OE algorithm's performance is slightly decreased, while that for the greedy algorithm becomes better. This is due to the fact that, the number of NOMA groups and the quota are fixed, more NOMA groups will be fully loaded when there are more nF-UEs, in which case, the ICI will dominate the performance.

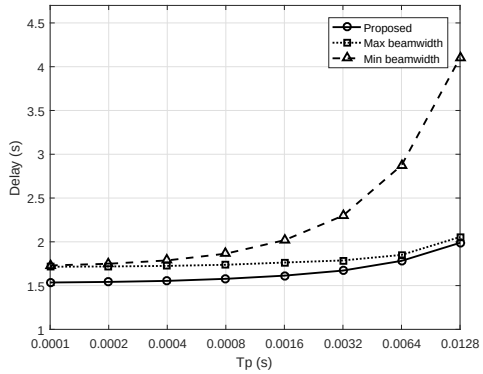

Fig. 4. Performance of mmW beam alignment when considering different pilot transmission T_p.

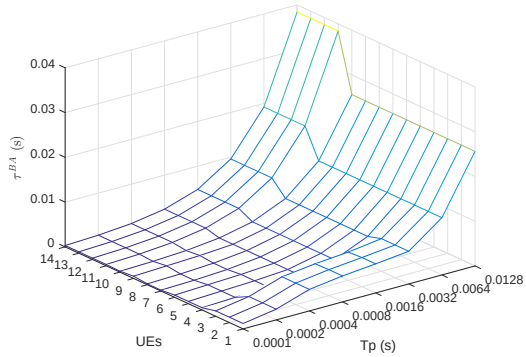

Fig. 5. Delay required by beam alignment when employing the JBTP algorithm.

Furthermore, the greedy algorithm is outperformed by the random method, which implies that the average delay performance may be heavily degraded by fewer nF-UEs with poor channel qualities. From the above, we may conclude that, in the NOMA-mmW offloading design of practical MEC system, one should properly schedule UE transmission by setting a suitable quota as well as finding the optimal for the NOMA grouping.

Furthermore, Fig. 4 shows how the beamwidth optimization affects the delay performance. For comparison, in the figure, we introduce the "max-beamwidth" and "min-beamwidth" methods, which simply use $\{\theta_k = \theta^{\max}, \forall k\}$ and $\{\theta_k = \theta^{\min}, \forall k\}$, respectively. Note that, they all employ the same NOMA grouping and power allocation approaches as the proposed one. As seen, the proposed algorithms can significantly outperform the other two methods, regardless of T_p employed. In the low T_p regions, such as $T_p \leq 0.0004$, the optimal beamwidth prefers to choose a medium value of $[\theta^{\min}, \theta^{\max}]$, and its contribution to reducing the total delay is relatively limited. By contrast, as the value of T_p gets bigger, the optimal beamwidth tends to approaching the maximum beamwidth $\theta^{\max}$. This obervation implies that, when using mmW transmission for MEC offloading, one should carefully balance the benifit of enhanced antenna gain by narrower beamwidth and the increased delay cost by beam alignment.

Fig. 5 shows the delay of beam alignment τ^{BA} by evaluating all the users while considering different the pilot transmission

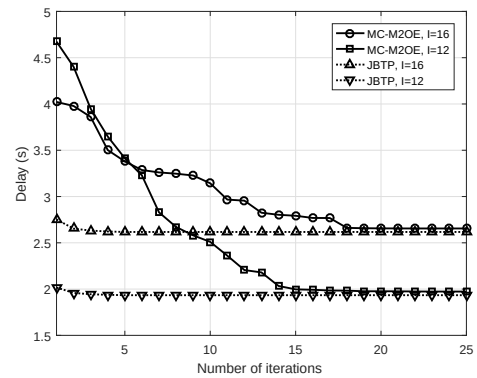

Fig. 6. Convergence performance of the proposed M2OE and JBTP algorithms.

time T_p. Note that, on the x-axis, the indexes from 1 to 10 indicate nF-users, and those from 11 to 14 indicate F-users, where assuming the UEs with smaller indexes have better channel qualities. It can be seen that, for all the UEs, the beam alignment delay τ^{BA} increases as the pilot transmission time T_p gets bigger. Again, this observation confirms that, for mmW communication based MEC offloading, the beam alignment will demand a considerable amount of delay in addition to data transmission, if pilot transmission is not properly designed. Furthermore, we observe that, the F-UEs have higher beamalignment delay than the nF-UEs, since they utilize narrower beamwidth. This is because that, within each NOMA group, the F-UE's signal is always decoded at last, and is free of ICI, which stimulate the F-UE to pursue higher antenna gain for data transmission. Due to the similar reason, the nF-UEs 1, 2, 3 with higher channel qualities need longer delay of beam alignment than the rest of nF-UEs. By contrast, we can see an exception that, when the pilot transmission time is too long, such as $T_p \geq 0.0032$, all the nF-UEs will have almost the same delay for beam alignment. This observation implies that, the JBTP algorithm tends to reduce the delay of data transmission only by optimizing the transmit power, regardless of the beamwidth used.

Finally, in Fig. 6, we evaluate the convergency of the proposed algorithms. As you may aware, in Table 1 and Table 2, they all involve the iterative processes for NOMA grouping, joint beamwidth and transmit power optimization. Seen from the figure, both the M2OE and JBTP algorithms can converge to the optimal solutions within 20 iterations. Further, as the number of nF-UEs increases, slight more iterations are required. We also would like to mention that, the AO process in Table 3 demands the number of iterations being smaller than 5. Overall, the above facts confirm the low-complexity feature of our proposed algorithms.

VI. Conclusion

In this paper, we have studied the delay performance of the NOMA-mmW scheme based MEC offloading system. In particular, we focused on solving the RA problem of jointly optimizing the beamwidth, NOMA grouping and transmit power, aiming at minimizing the average delay of MEC offloading transmissions. To tackle the MINLP problem, we

proposed the AO-RA scheme, where the MC-M2OE algorithm and the JBTP algorithm designed are iteratively operated. We have carried out the comprehensive performance evaluation for the proposed algorithms. Our simulation results showed that, the AO-RA scheme had high-efficiency in minimizing the offloading delay. In particular, the MC-M2OE could always find the best UE scheduling for NOMA grouping, while demanding relatively low complexity. Furthermore, from the delay performance we drew the conclusion that, in the practical MEC system design, one should properly schedule UE transmission by setting a suitable quota as well as finding the optimal NOMA grouping, rather than simply investing more transmit power causing unnecessary energy waste.

REFERENCES

[1] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Enabling massive IoT toward 6G: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11 891–11 915, Aug. 2021.

[2] M. Shabbir *et al.*, "Enhancing security of health information using modular encryption standard in mobile cloud computing," *IEEE Access*, vol. 9, pp. 8820–8834, 2021.

[3] M. Sirajuddin, C. Rupa, C. Iwendi *et al.*, "TBSMR: A trust-based secure multipath routing protocol for enhancing the QoS of the mobile ad hoc network," *Security and Communication Networks*, 2021.

[4] J. Cao, W. Feng, N. Ge, and J. Lu, "Delay characterization of mobile-edge computing for 6G time-sensitive services," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3758–3773, Mar. 2021.

[5] J. Santos, T. Wauters, B. Volckaert, and F. De Turck, "Towards low-latency service delivery in a continuum of virtual resources: State-of-the-art and research directions," *IEEE Commun. Surv. Tutor.*, pp. 1–1, Jul. 2021.

[6] J. Zhao, Q. Li, Y. Gong, and K. Zhang, "Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7944–7956, Jun. 2019.

[7] X. Yang, H. Luo, Y. Sun, and M. S. Obaidat, "Energy-efficient collaborative offloading for multiplayer games with cache-aided MEC," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–7.

[8] H. Feng, S. Guo, L. Yang, and Y. Yang, "Collaborative data caching and computation offloading for multi-service mobile edge computing," *IEEE Trans. Veh. Technol.*, pp. 1–1, Sept. 2021.

[9] K. Guo, R. Gao, W. Xia, and T. Q. S. Quek, "Online learning based computation offloading in MEC systems with communication and computation dynamics," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1147–1162, Nov. 2021.

[10] J. Liu, H. Guo, J. Xiong, N. Kato, J. Zhang, and Y. Zhang, "Smart and resilient EV charging in SDN-enhanced vehicular edge computing networks," *IEEE J. Select. Areas Commun.*, vol. 38, no. 1, pp. 217–228, Jan. 2020.

[11] H. Guo, Z. X., J. Liu *et al.*, "Vehicular intelligence in 6G: Networking, communications, and computing," *Vehicular Communications*, 2021.

[12] H. Guo and J. Liu, "UAV-enhanced intelligent offloading for internet of things at the edge," *IEEE Trans Ind. Informat.*, vol. 16, no. 4, pp. 2737–2746, Apr. 2020.

[13] V. Frascolla *et al.*, "Millimeter-waves, MEC, and network softwarization as enablers of new 5G business opportunities," in *2018 IEEE WCNC*, 2018, pp. 1–5.

[14] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin *et al.*, "Millimeter wave communications for future mobile networks (guest editorial), part i," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1425–1431, Jul. 2017.

[15] J. Kim and W. Lee, "Feasibility study of 60 GHz millimeter-wave technologies for hyperconnected fog computing applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1165–1173, Feb. 2017.

[16] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Edge computing meets millimeter-wave enabled VR: Paving the way to cutting the cord," in *Proc. IEEE WCNC*, 2018, pp. 1–6.

[17] J. Wang, R. Han, L. Bai, T. Zhang, J. Liu, and J. Choi, "Coordinated beamforming for UAV-aided millimeter-wave communications using GPML-based channel estimation," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 100–109, Dec. 2021.

[18] Q. Chen, X. Xu, H. Jiang, and X. Liu, "An energy-aware approach for industrial internet of things in 5G pervasive edge computing environment," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 5087–5097, Jul. 2021.

[19] C. Zhao, Y. Cai, A. Liu, M. Zhao, and L. Hanzo, "Mobile edge computing meets mmwave communications: Joint beamforming and resource allocation for system delay minimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2382–2396, Jan. 2020.

[20] S. Barbarossa, E. Ceci, M. Merluzzi, and E. Calvanese-Strinati, "Enabling effective mobile edge computing using millimeterwave links," in *Proc. Int. Conf. Commun. Workshops (ICC Workshops)*, 2017, pp. 367–372.

[21] S. B. Prathiba, G. Raja, S. Anbalagan, K. Dev, S. Gurumoorthy, and A. P. Sankaran, "Federated learning empowered computation offloading and resource management in 6G-V2X," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–1, Aug. 2021.

[22] Y. Chen, B. Ai, Y. Niu, Z. Zhong, and Z. Han, "Energy efficient resource allocation and computation offloading in millimeter-wave based fog radio access networks," in *Proc. Int. Conf. Commun. (ICC)*, 2020, pp. 1–7.

[23] Z. Zhao, J. Shi, Z. Li, J. Si, P. Xiao, and R. Tafazolli, "Multi-objective resource allocation for mmWave MEC offloading under competition of communication and computing tasks," in *IEEE Internet Things J.*, to be published, doi=10.1109/JIOT.2021.3116718.

[24] K. A. Noghani, H. Ghazzai, and A. Kassler, "A generic framework for task offloading in mmWave MEC backhaul networks," in *2018 IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–7.

[25] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, Jan. 2018.

[26] Q.-V. Pham, H. T. Nguyen, Z. Han, and W.-J. Hwang, "Coalitional games for computation offloading in NOMA-enabled multi-access edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1982–1993, Nov. 2020.

[27] L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "Noma-enabled mobile edge computing for internet of things via joint communication and computation resource allocations," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 718–733, Nov. 2020.

[28] B. Li, F. Si, W. Zhao, and H. Zhang, "Wireless powered mobile edge computing with NOMA and user cooperation," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1957–1961, Feb. 2021.

[29] N. Nouri, J. Abouei, M. Jaseemuddin, and A. Anpalagan, "Joint access and resource allocation in ultradense mmWave NOMA networks with mobile edge computing," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1531–1547, Feb. 2020.

[30] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu *et al.*, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2192–2235, Q4th 2020.

[31] Y. Yu, X. Bu, K. Yang, H. Yang, X. Gao, and Z. Han, "UAV-aided low latency multi-access edge computing," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4955–4967, May 2021.

[32] M. Mezzavilla, M. Zhang, M. Polese *et al.*, "End-to-end simulation of 5G mmWave networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2237–2263, Q3th 2018.

[33] F. Pantisano1, M. Bennis, W. Saad, S. Valentin, and M. Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *2013 IEEE Global Commun. Conf. (GLOBECOM)*, 2013, pp. 1–7.

[34] Y. Li, H. Zhang, and K. Long, "Joint resource, trajectory, and artificial noise optimization in secure driven 3-D UAVs with NOMA and imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 11, pp. 3363–3377, Nov. 2021.

[35] A. Alsharoa and M. Yuksel, "Energy efficient D2D communications using multiple UAV relays," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5337–5351, Aug. 2021.