Resource Allocation for NOMA-MEC Systems in Ultra-Dense Networks: A Learning Aided Mean-Field Game Approach

Lixin Li, *Member, IEEE,* Qianqian Cheng, Xiao Tang, *Member, IEEE,* Tong Bai, *Member, IEEE,* Wei Chen, *Senior Member, IEEE,* Zhiguo Ding, *Fellow, IEEE,* and Zhu Han, *Fellow, IEEE*

***Abstract*—Attracted by the advantages of multi-access edge computing (MEC) and non-orthogonal multiple access (NOMA), this paper studies the resource allocation problem of a NOMA-MEC system in an ultra-dense network (UDN), where each user may opt for offloading tasks to the MEC server when it is computationally intensive. Our optimization goal is to minimize the system computation cost, concerning the energy consumption and task delay of users. In order to tackle the non-convexity issue of the objective function, we decouple this problem into two sub-problems: user clustering as well as jointly power and computation resource allocation. Firstly, we propose a user clustering matching (UCM) algorithm exploiting the differences in channel gains of users. Then, relying on the mean-field game (MFG) framework, we solve the resource allocation problem for intensive user deployment, using the novel deep deterministic policy gradient (DDPG) method, which is termed by a mean-field-deep deterministic policy gradient (MF-DDPG) algorithm. Finally, a jointly iterative optimization algorithm (JIOA) of UCM and MF-DDPG is proposed to minimize the computation cost of users. The simulation results demonstrate that the proposed algorithm exhibits rapid convergence, and is capable of efficiently reducing both the energy consumption and task delay of users.**

***Index Terms*—Multi-access edge computing, non-orthogonal multiple access, deep reinforcement learning, mean-field game, deep deterministic policy gradient**

Manuscript received XXX; revised XXX; accepted XXX. Date of publication XXX; date of current version XXX. This work was supported in part by the Aerospace Science and Technology Innovation Fund of China Aerospace Science and Technology Corporation, in part by the Shanghai Aerospace Science and Technology Innovation Fund under Grant SAST2018045, in part by the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University under Grant CX2020152, in part by National Natural Science Foundation of China under Grant 61901378, Grant 61941119, China Postdoctoral Science Foundation under Grant BX20190287, and in part by Grant NSF EARS-1839818, Grant CNS1717454, Grant CNS-1731424, and Grant CNS-1702850. This article was presented in part at the IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020. The associate editor coordinating the review of this article and approving it for publication was XXX. (Corresponding author: Lixin Li; Xiao Tang.)

L. Li, Q. Cheng, and X. Tang are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (email: lilixin@nwpu.edu.cn; chengqianqian@mail.nwpu.edu.cn; tangxiao@nwpu.edu.cn).

T. Bai is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, U.K. (e-mail: t.bai@outlook.com).

W. Chen is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wchen@tsinghua.edu.cn).

Z. Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

Z. Han is with the Department of ECE, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: zhan2@uh.edu).

I. Introduction

With the rapid development of mobile communication, computationally intensive applications and data traffic have grown explosively. Moreover, the fifth generation (5G) mobile communication technology has developed rapidly in recent years, which facilitates the deployment of ultra-dense networks (UDNs) in the future development of communications [1]. The UDNs can effectively enhance system capacity and data transmission rate while guaranteeing the quality of service (QoS) of users. Specifically, an UDN refers to densely deploying the small base stations (SBSs) and then accessing a large number of users for improving network coverage and for reducing transmission delay [2]. However, it is a huge challenge to solve the computationally intensive tasks in UDNs due to the limited computing power of users .

As an emerging technology, multi-access edge computing (MEC) has been proposed to alleviate computational pressure in UDNs [3]. Specifically, users can offload the entire or a fraction of their computing task to MEC servers through wireless channels for improving the QoS of users, which reduces network delay and energy consumption. Furthermore, due to the deployment of UDNs and the limited spectrum resources, the classical orthogonal multiple access (OMA) technologies can no longer meet the requirement imposed on both low delay and on low energy consumption in MEC. As an emerging technology, non-orthogonal multiple access (NOMA) can effectively improve the system spectral efficiency by allocating the same resources to multiple users compared with OMA [4], [5].

At the time of writing, reinforcement learning (RL) has become a powerful tool in MEC systems to solve diverse problems, such as resource allocation and power control [6]–[8]. In RL, the agent interacts with the environment and selects the actions based on the current state to obtain the optimal policy. However, when the number of agents is very large, the action space and state space of the agent are usually high-dimensional and continuous, which is suitable for applying the deep reinforcement learning (DRL) algorithm based on policy gradients, such as stochastic policy gradient (SDG) and deterministic policy gradient (DPG) [9], [10]. Deep deterministic policy gradient (DDPG) combines the advantage of actor-critic (A2C) and deep Q-network (DQN) based on the DPG algorithm, which solves the problem that reward function is difficult to converge in continuous action spaces

TABLE I: List of Notations.

Notations	Description	Notations	Description
N	Number of SBSs	K_m	Number of users for each NOMA cluster
K	Number of users in each SBS	y_n	Received signals of the nth SBS
M	Number of NOMA clusters in each SBS	τ_{mk}	SINR of the nth user for mth cluster
p_{mk}	Transmit power of the user	I_{NOMA}	Interference introduced by NOMA
h_{mk}	Channel gain of the user	I_{TDMA}	Interference introduced by TDMA
d_{mk}	Input data of the computing task	f_{mk}	Computing capability per CPU cycle
c_{mk}	Number of CPU cycles required to complete the calculation task	e^l_{mk}	Time for local computing
$T^{\max}_{mk}$	Maximum delay of the task	δ^l_{mk}	Energy consumption per CPU cycle
t^l_{mk}	Time for local computing	λ^t_{mk}	Weight coefficient of the task delay
λ^e_{mk}	Weight coefficient of energy consumption	W	Bandwidth of system
f_n	Computing capability per CPU cycle of MEC server of the nth SBS	β_{mk}	determines whether the kth user is on the mth NOMA cluster
t^n_{mk}	Transmission time for offloading computing	t^{ex}_{mk}	Calculation time of the calculation process
e^n_{mk}	Transmission energy consumption for offloading computing	e^{ex}_{mk}	Energy consumption of the calculation process
σ^2	Additive white Gaussian noise	δ_{mk}	Energy consumed by each CPU unit

[11]. Specifically, DDPG adds target networks and evaluation networks to the actor-critic network, while the deep neural network (DNN) is used for functional approximation.

On the other hand, mean-field game (MFG) has also been widely applied to solve various resource allocation problems in UDNs [12]–[14]. Unlike the classical games, MFG is a specific game theory used for studying member intensification, which may simplify the large and complex models. Specifically, for an independent individual in UDNs, the effect that all members are faced with can be deemed to be a mean-field. Thus, MFG simplifies the complex problem to the individual problem of each member. In MFG, the Hamilton-Jacobi-Bellman (HJB) equation is defined to characterize the interactions between individuals and the mean-field. The Fokker-Planck-Kolmogorov (FPK) equation is defined to describe the evolution of the mean-field under the individual decision of the game. Then, the equilibrium solution of the MFG is obtained by obtaining the solutions of the HJB equation and the FPK equation. Conventionally, a finite difference method is usually used to solve the mean-field equilibrium (MFE) [15]. However, when the dimensions of the state space and the action space becomes large, the finite difference method requires a large amount of computation, which prohibits the method from realistic implementation. Given that MFG can be attributed to the Markov decision process (MDP) based optimization problem, it is natural to exploit the RL algorithm to solve the MFE [16].

In this paper, a NOMA-MEC system is modeled for the UDN scenario, where each SBS is equipped with an MEC server. The so-called partial offloading policy is selected for our considered system.To elaborate, when a user is incapable of processing a large number of computing tasks in a timely manner, a fraction of the tasks is offloaded to the MEC server. The main works and contributions of this paper can be summarized as follows:

- An ultra-dense NOMA-MEC system is proposed, where each SBS is equipped with one MEC server and serves multiple users. In this system, all users served by each SBS are divided into different clusters, in which the users in each cluster adopt the NOMA transmission scheme. Moreover, the TDMA transmission scheme is used between different clusters to avoid the NOMA inter-cluster interference.
- A novel resource allocation algorithm is designed to reduce the energy consumption and task delay of users in this paper. Since the prime problem is a mixed integer nonlinear programming (MINLP), which is non-convex and difficult to solve, we decompose it into two sub-problems of user clustering as well as of jointly power and computing resource allocation. For the user clustering problem, a user clustering matching (UCM) algorithm based on the differences in channel gains of users is proposed. And for the second sub-problem, it is modeled as a MFG theoretical framework considering the intensive deployment of users, and then proposes the mean-field-deep deterministic policy gradient (MF-DDPG) algorithm to obtain the equilibrium solution of MFG considering the continuity of the action space.
- In order to obtain a high-quality suboptimal solution, a jointly iterative optimization algorithm (JIOA) of UCM and MF-DDPG is proposed to minimize the computation cost of users. Due to the non-convexity of the objective function, the optimal solution cannot be obtained in polynomial time. However, as the number of iterations increases in the proposed algorithm, the objective function is always updated toward the optimal direction, and at least a suboptimal solution close to the global optimum will be obtained. Simulation results show that the proposed algorithm can effectively reduce the computation cost by comparing with the benchmark algorithms.

The rest of the paper is structured as follows. The related works are introduced in Section II. In Section III, the system model is introduced and the formulation of the optimization problem is demonstrated. In Section IV, a joint optimization algorithm is proposed for the resource allocation problem. The simulation results are discussed in Section V. Finally, Section VI concludes the paper.

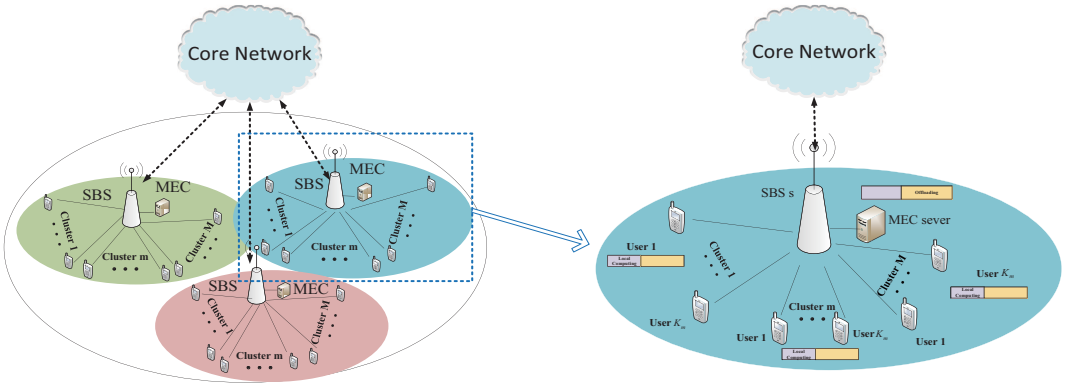


Fig. 1: An illustration of the NOMA-MEC system in the UDN, where a large number of users are covered by N SBSs. Each SBS is equipped with an MEC server. Users may opt for offloading a fraction of their tasks to the MEC server, if their computational task is intensive.

II. Related Works

At present, MEC has made great progresses in the field of communications. In [17], the authors studied the joint task and resource allocation based on MEC in small cell networks, and proposed a novel multi-stack Q-learning method to reduce the total delay of the system. In [18], a heuristic greedy task offloading method was proposed to solve the dynamic task offloading problem for the MEC system in the UDN. The allocation of task and resource were discussed in [19], which aimed for minimizing the energy consumption on the mobile terminals under the constraints of the application delay.

In the past years, OMA technologies such as time division multiple access (TDMA) and frequency division multiple access (FDMA) have been applied to solve the problem of resource allocation with MEC. In [20], the authors proposed an efficient joint optimization algorithm to reduce computational delay with the TDMA transport protocol. In [21], the authors proposed a numerical analysis method under the condition of considering the FDMA scheme to optimize the system energy consumption in the multi-user MEC system. However, with the ultra-dense deployment of devices, the OMA technology can hardly meet the demand for spectrum resources of devices. As an emerging multiple access method, NOMA can effectively improve the spectral efficiency of the system by allocating the same resources to multiple devices compared with OMA [22]. Exploiting this advantages, MEC systems have employed NOMA to reduce energy consumption and task computation delay. More explicitly, in [23], both NOMA uplink and downlink can be applied to MEC. It is further proved that the delay and energy consumption of MEC offloading can be effectively reduced by using the NOMA method. [24] and [25] studied the offloading process of NOMA-MEC, and achieved energy-efficient offloading performance by jointly optimizing transmit power, time, and task offloading policy.

Nowadays, RL methods have been applied to various communication systems to solve the optimization problems. In [26], the authors proposed an effective resource management algorithm based on RL, which learned the optimal offloading policy online. In [27], the authors designed an uncoordinated DRL algorithm to reduce the interference and increase the network throughput in the uplink grant-free NOMA system. In [28], the authors studied a NOMA-MEC system of a multi-user and single MEC server, which employed DQN to select users that are offloaded and minimize the offloading delay of the whole system.

Moreover, the MFG method is proposed to solve the complex collective behavior in UDNs. In [31], the authors proposed a power control method based on the MFG to improve the energy efficiency in a wireless power transmission system. In [29], the authors modeled an ultra-dense D2D network as a MFG framework and proposed a distributed power control algorithm to improve the spectrum efficiency and energy efficiency of the system. In [30], the authors formulated the downlink power problem as a MFG problem to reduce the energy consumption in dense small cells underlying microcells system. Nowadays, extensive research contributions have been devoted to combine RL with MFG for reducing the complexity of solving the MFG. In [32], the author investigated the cell association problem between SBSs and users in dense wireless networks, which described it as a MFG, and then proposed neural Q-learning algorithm to optimize the user's data transmission rate. In [33], the authors used ML and MFG methods to optimize the beamforming and beam-steering respectively, which aimed to increase the transmission rate of users in the MIMO communication system. Different from these works, this paper uses DDPG method to obtain the equilibrium solution of MFG, which solves the problem that DQN cannot handle continuous actions and obtains more accurate equilibrium solution.

III. System Model

As shown in Fig. 1, we consider a NOMA-MEC communication system in an UDN with $N = \{1, 2, \cdots, |N|\}$ SBSs, where each SBS is equipped with an MEC server for the users to compute the offloading computing tasks and serves $K = \{1, 2, \cdots, |K|\}$ users in a certain area. In this network, K users are grouped into M NOMA clusters with K_m users in each cluster. For each user $k \in K$, we assume that it has very limited computing power, and therefore, it may opt for offloading its tasks to the MEC server via the wireless channel according to the designed offloading policy when its computing task is intensive. The channel state information (CSI) is

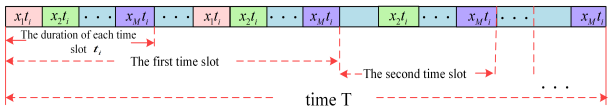


Fig. 2: An illustration of the considered TDMA scheme.

assumed to be perfect. As a benefit, the computing pressure of the user is slowed down by offloading the computing task to the MEC server, and the energy consumption by users for task computing can be reduced.

A. Transmission Model

Let us continue by elaborating on the transmission model of the NOMA-MEC uplink system. For each NOMA cluster m ($m \in M$), each user k ($k \in K$) sends its signal x_{mk} to the SBS n ($n \in N$) with complementary correlation power, so the total received signals y_n of the nth SBS is

$$y_n = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \sqrt{p_{mk}} h_{mk} x_{mk} + z_n, \tag{1}$$

where p_{mk} is the transmit power of the kth user in the mth cluster, and h_{mk} is the channel gain of the kth user in the mth cluster. z_n is the additive noise at the nth SBS.

In each NOMA cluster, the channel gain of the user is sorted in descending order. Therefore, users with large channel gains will be interfered by users with smaller channel gains in the same cluster. Moreover, users with the smallest channel gain will not be interfered by other users in the cluster. Therefore, the interference among users within a cluster can be expressed as:

$$I_{NOMA} = \sum_{f=k+1}^{K_m} p_{mf} |h_{mf}|^2, \tag{2}$$

where p_{mf} is the transmit power of the fth user in the mth cluster, and h_{mf} is the channel gain of the fth user in the mth cluster.

Using only the NOMA transmission scheme will add additional NOAM inter-cluster interference. Therefore, in order to reduce the interference suffered by users, the TDMA transmission scheme is used in different NOMA clusters. As shown in Fig. 2, the TDMA technology divides time T into periodic non-overlapping frames, and then divides each frame into a plurality of mutually non-overlapping slots. And then, these time slots will be allocated to the different NOMA clusters, and users in each cluster take up the total bandwidth in the proportion of x_i to offload the task. Therefore, each user is assigned the time of $x_i T_s$. In the UDN, however, users served by different SBSs transmit tasks in the same time slot, which causes interference each other, which is expressed as:

$$I_{TDMA} = \sum_{j \neq n}^{N} \sum_{k=1}^{K_m} p_{j,mk} |h_{j,mk}|^2, \tag{3}$$

where $p_{j,mk}$ is the transmit power of the kth user in the jth SBS, and $h_{j,mk}$ is the channel gain of the kth user in the jth SBS. Thus, the signal to interference plus noise ratio (SINR)

Algorithm 1 UCM algorithm for NOMA-MEC system

```
Initialize the user's clustering by randomly matching the users
for each SBS n ∈ N do
    Sort the channel gains of users in descending order
    Select the first m users as the first user in each cluster
    for each cluster m = 1 to M, do
        for each user k ∈ K do
            Calculate the difference in channel gain
            Find user k to satisfy (20)
            Add user k to cluster m
        end for
    end for
    if M is not an integer then
        The remaining users are randomly assigned to the NOMA clusters
    end if
end for
```

of the kth user in the mth cluster is

$$\begin{aligned} \tau_{mk} &= \frac{p_{mk}|h_{mk}|^2}{I_{NOMA} + I_{TDMA} + \sigma^2} \\ &= \frac{p_{mk}|h_{mk}|^2}{\sum_{f=k+1}^{K} p_{mf}|h_{mf}|^2 + \sum_{j \neq n}^{N} \sum_{k=1}^{K} p_{j,mk}|h_{j,mk}|^2 + \sigma^2}, \end{aligned} \tag{4}$$

where σ^2 is the additive white Gaussian noise (AWGN). So the data rate of the kth user in the mth NOMA cluster is defined as:

$$r_{mk} = W \log_2(1 + \tau_{mk}), \tag{5}$$

where W is bandwidth of system.

B. Computational Model

For each user's computing task, it can be defined as: $A_{mk} \triangleq (d_{mk}, c_{mk}, T_{mk}^{\max})$. Here d_{mk} indicates the amount of input data (bits) required by the kth user in the mth NOMA cluster to transfer the computing task from the user to the MEC server. c_{mk} denotes the number of CPU cycles required by user k to calculate the input data, i.e., the amount of computation required to complete the computing task, $T_{mk}^{\max}$ represents the deadline that the kth user takes to complete the computing task. We consider the data-partitioning based computational tasks, which can be addressed relying on the so-called partial offloading policy [34]. Specifically, each task can be divided into two parts, some of which can be executed locally and the other part offloaded to the MEC server for execution.

1) Local computing model: Upon using $f_{mk} > 0$ to represent the computing capacity for local computing of the kth user in the mth cluster in terms of CPU cycles per second, it is readily to formulate the time for local computing as

$$t_{mk}^{l} = \frac{c_{mk}}{f_{mk}}. \tag{6}$$

According to the dynamic voltage and frequency scaling (DVFS) techniques [36], the energy consumption of local consumption can be expressed as $\delta_{mk} = \kappa f_{mk}^2$, where κ is the energy coefficient which depends on the integrated chip structure. Hence, the energy consumption for local computing

is given by

$$e^l_{mk} = c_{mk}\delta_{mk}, \tag{7}$$

where δ_{mk} is the energy consumed by each CPU unit.

According to (6) and (7), the overall cost of the kth user in the mth NOMA cluster for selecting the local computing policy can be written as

$$\Phi^l_{mk} = \lambda^t_{mk} t^l_{mk} + \lambda^e_{mk} e^l_{mk}, \tag{8}$$

where $\lambda^t_{mk} + \lambda^e_{mk} = 1$. λ^t_{mk} and λ^e_{mk} represent the weight coefficients of the task delay and energy consumption for each user, respectively. Obviously, we have $\lambda^t_{mk} > 0$ and $\lambda^e_{mk} > 0$. Note that $\lambda^t_{mk} > \lambda^e_{mk}$ implies that the user is sensitive to delay and we should pay more attention to its calculation time, while $\lambda^t_{mk} < \lambda^e_{mk}$ means that the users pay more attention to the energy consumed by computing tasks compared to the task delay.

2) Offloading Computing Model: At the edge node, the calculation of tasks is completed in two steps: data transmission and execution calculation. Upon assuming that the kth user offloads the task to the nth SBS, the total calculation offloading time can be divided into the transmission time t^n_{mk} of the calculation task and the calculation time t^{ex}_{mk} of the calculation performed on the MEC server, which can be respectively written as

$$t^n_{mk} = \frac{d_{mk}}{r_{mk}}, \tag{9}$$

and

$$t^{ex}_{mk} = \frac{c_{mk}}{f_n}, \tag{10}$$

where f_n denotes the computing capacity of the MEC server. It is readily to formulate the total time consumed on the edge computing as

$$t^c_{mk} = t^n_{mk} + t^{ex}_{mk}. \tag{11}$$

Similarly, the energy consumption of the offloading process is composed of the energy consumption e^n_{mk} of the transmission task and the energy consumption e^{ex}_{mk} of the calculation process, which can be expressed as

$$e^n_{mk} = p_{mk}\frac{d_{mk}}{r_{mk}}, \tag{12}$$

and

$$e^{ex}_{mk} = c_{mk} f_n. \tag{13}$$

Then, the total energy consumption of the kth user in the mth NOMA cluster during the calculation of the offloading process is readily given by

$$e^c_{mk} = e^n_{mk} + e^{ex}_{mk}. \tag{14}$$

To this end, the overall cost of the computing task of the kth user in the mth NOMA cluster on the MEC server can be written as

$$\Phi^c_{mk} = \lambda^t_{mk} t^c_{mk} + \lambda^e_{mk} e^c_{mk}. \tag{15}$$

C. Problem Formulation

Considering the user's dual needs for low task delay and low energy consumption, the computation cost of the kth user in the mth NOMA cluster is defined

$$\Phi_{mk} = \rho^l_{mk}\Phi^l_{mk} + \rho^c_{mk}\Phi^c_{mk}, \tag{16}$$

where ρ^l_{mk} and ρ^c_{mk} are the weight coefficients of the local computing and offloading computing for each user, respectively.

In this paper, we aim for minimizing the computational cost of the NOMA-MEC system in an UDN, by controlling the user's transmit power, the offloading decision factor and the weight coefficient of delay and energy consumption. The pertinent problem is formulated as

$$\begin{aligned}
\mathcal{P}0: \min\Phi = \min & \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\beta_{mk}(\rho^l_{mk}\Phi^l_{mk} + \rho^c_{mk}\Phi^c_{mk}),\\
s.t. \quad & C1: \rho^l_{mk} t^l_{mk} + \rho^c_{mk} t^c_{mk} \le T^{\max}_{mk},\\
& C2: W\log_2\Big(1 + \frac{p_{mk}h_{mk}}{I_{NOMA} + I_{TDMA} + \sigma^2}\Big) \ge \Upsilon_{mk},\\
& C3: 0 \le \Upsilon_{mk} \le d_{mk},\\
& C4: 0 \le p_{mk} \le P_{mk},\\
& C5: \rho^l_{mk} + \rho^c_{mk} = 1, \quad \rho^l_{mk} > 0, \quad \rho^c_{mk} > 0,\\
& C6: \lambda^l_{mk} + \lambda^c_{mk} = 1, \quad \lambda^l_{mk} > 0, \quad \lambda\rho^c_{mk} > 0,\\
& C7: \sum_{k=1}^{K}\beta_{mk} = K_n, \beta_{mk} = \{0,1\},
\end{aligned} \tag{17}$$

where β_{mk} represents the association between the kth user and the mth NOMA cluster in the nth SBS. To be specific, $\beta_{mk} = 1$ indicates that the kth user in the nth SBS is assigned to the mth NOMA cluster. As for the constraints, **C1** is the maximum task delay allowed by the user. **C2** represents the minimum transmission rate required by the user. **C3** is the user's transmission rate constraint. **C4** restricts the user's transmit power. **C5** gives the bounds of ρ^l_{mk} and ρ^c_{mk}. **C6** imposes the bounds to λ^l_{mk} and λ^c_{mk}. **C7** details the bounds of β_{mk}.

IV. Algorithm Formulation

It can be seen from formula (17) that the C5 and C6 are continuous variables, and the C7 is a binary discrete variable, so the problem $\mathcal{P}0$ can be regarded as a MINLP problem [35]. To solve this non-convex problem, we decompose it into two sub-problems: user clustering and resource allocation. In the first subproblem, K users are clustered with the aid of the UCM algorithm, for reducing the computation complexity. In order to address the issue of the intensive user deployment, the resource allocation problem is modeled relying on the MFG framework, where the MFE is found by our proposed MF-DDPG algorithm. Finally, we may obtain a locally optimal solution, by stopping the iterations of these two algorithms once the objective function converges. To elaborate a little further, the objective function is non-convex and we cannot obtain the optimal solution in polynomial time. However, in each iteration, the objective function is always updated towards a smaller value. When the objective function converges, at least one suboptimal solution close to the global optimal solution can be obtained.

A. Clustering Matching Algorithm

In the NOMA-MEC system, multiple users share the same spectrum resources at the same channel for transmission. However, too many users will increase interference among users, thereby affecting the energy consumption and time delay of the system. Therefore, in the NOMA system, how to use different user channel conditions to cluster users to maximize system benefits for multiple users in the SBS is a realistic and important issue. Given a fixed resource allocation, we may simplify Problem $\mathcal{P}0$ as

$$\begin{aligned} & \min \Phi', \\ s.t. \quad & C1: W\log_2\left(1+\frac{p_{mk}h_{mk}}{I_{NOMA}+I_{TDMA}+\sigma^2}\right) \geq \Upsilon_{mk}, \\ & C2: 0 \leq \Upsilon_{mk} \leq d_{mk}, \\ & C3: \sum_{k=1}^{K} \beta_{mk} = K_m, \beta_{mk} = \{0,1\} \end{aligned} \tag{18}$$

It is readily seen that the objective function of (18) is directly determined by the channel gain of users when the resources are identified, including power of users and weight coefficient variables. Specifically, provided that the channel gains difference among different channels is large, a better performance can be achieved by increasing the transmit power of each user. In other words, the differences in channel gains among different users in the same cluster determines the upper limit of NOMA performance. At present, some studies have shown that dividing users with large channel gain differences into a cluster can effectively improve the performance of NOMA [38]. More explicitly, when the differences in channel gains of users in the cluster is larger, it is easier for the receiving terminal to detect the superposed signals. As such, the computing cost of the system can be reduced, by maximizing the channel gain difference of the users in the NOMA cluster.

Based on this observation, we propose a novel UCM algorithm based on the differences of channel gains as shown in **Algorithm 1**. The basic idea of the UCM algorithm is to select users with different channel qualities to multiplex the same resources, which are detailed as follows.

In the system, U users are randomly distributed in each SBS, and the Rayleigh fading gains of the users are denoted as $h_1^2, h_2^2, \cdots$, and h_K^2, respectively, i.e., $|h_1|_1^2 > |h_1|_2^2 > \cdots > |h_1|_K^2$. It is assumed that there is no users with the same channel gain. Without loss of generality, K users are grouped into M NOMA clusters, in where there are each with K_m users.

The difference in channel gains between adjacent users in the mth cluster can be defined as

$$\Delta|h_m|^2 = \sum_{i=1}^{K} |h_{m,i}|^2 - |h_{m,i+1}|^2, \tag{19}$$

where $h_{m,i}^2$ and $h_{m,i+1}^2$ are the channel gains of two adjacent users in the mth cluster, respectively. The difference of the channel gains for all clusters needs to be considered to optimize overall system performance, so the optimization objective function in (18) of this sub-problem can be rewritten as

$$\begin{aligned} \max \sum_{m=1}^{M} \Delta|h_m|^2 = \max \sum_{m=1}^{M} |h_{m,i+1}|^2 - |h_{m,i}|^2, \\ s.t. \qquad |h_{m,+1}|^2 > |h_{m,i}|^2. \end{aligned} \tag{20}$$

As for the initial settings of the resource allocation, we allocate resources to the user's equally. Specifically, in the matching process, all users in each SBS are sorted according to the size of their channel gains, and then the users with the first M channel gains are sequentially selected as the first user in the cluster. Next, we select the user with the largest sum of channel gain differences in the cluster from the remaining users satisfied (20) to add the cluster.

Note that since there is relatively large co-channel interference among users in the same NOMA cluster, it is not appropriate to reuse too many users in each cluster. Otherwise, if there are more users in each cluster, the users with the previous decoding order are subject to severe co-channel interference, which will cause the computation cost of the user to increase.

B. Jointly Power and Computation Resource Allocation

Given a specific user clustering scheme, the optimization goal can be rewritten as

$$\begin{aligned} & \min \Phi'' = \min \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K_m} \Phi_{mk}, \\ s.t. \quad & \rho_{mk}^l t_{mk}^l + \rho_{mk}^c t_{mk}^c \leq T_{mk}^{\max}, \\ & C1: W\log_2\left(1+\frac{p_{mk}h_{mk}}{I_{NOMA}+I_{TDMA}+\sigma^2}\right) \geq \Upsilon_{mk}, \\ & C2: 0 \leq \Upsilon_{mk} \leq d_{mk}, \\ & C3: 0 \leq p_{mk} \leq P_{mk}, \\ & C4: \rho_{mk}^l + \rho_{mk}^c = 1, \rho_{mk}^l > 0, \rho_{mk}^c > 0, \\ & C5: \lambda_{mk}^l + \lambda_{mk}^c = 1, \lambda_{mk}^l > 0, \lambda\rho_{mk}^c > 0. \end{aligned} \tag{21}$$

When many users transmit data simultaneously in the NOMA-MEC system, the interference among devices becomes very serious. This severely reduces the data transmission rate of the user, and hence increases task delay and power consumption when offloading computing tasks. Each user needs to make corresponding decisions based on its received interference and its state, which is in line with the game theory. However, the classic game theory is mainly to describe the individual-to-individual interactions. Each individual needs to weigh the influence of other individuals to make the optimal decision. Therefore, when the number of individuals is very large, the analysis of the system will be very complicated to cause a huge amount of calculation.

Based on the above analysis, the problem in (21) is reformulated with the aid of the MFG theoretical framework considering the intensive deployment of users in the NOMA-MEC system. The MFG transforms a large amount of interaction information among individuals into interactions with the mass, which greatly reduces the complexity of the system.

Algorithm 2 MF-DDPG algorithm for NOMA-MEC system

for each user $k \in \mathcal{K}$ **do**
  Initialize channel gain for all devices
  Initialize the number of the users
  Initialize the experience replay buffer D
end for
for each episode $1, 2, ...,$ **do**
  Update the simulation parameters from NOMA-MEC system environment
  Obtain the initial state s for each user $k \in \mathcal{K}$
  for each time slot $t =$, **do**
    for each cluster $m \in M$, where M is determined by 1 **do**
      for each user $k \in K$, **do**
        Choose the action $a_k^t = \mu(s_k^t|\theta_k^\mu) + \Delta\mu$ according to the local observation information of user, where $\Delta\mu$ is the exploration noise
        Implement the action a_k^t, and obtain reward r_k^t and next state s_k^{t+1} from the environment
        Save tuple (s_i, a_i, r_i, s_{i+1}) into D
        Randomly sample a mini-batch from D
        Compute the gradient of the critic network to minimize the loss L

$$L = \frac{1}{B}\sum_{i=1}^{B}\left(r_i + \max Q(s_{i+1}, a|\theta_k^{Q'}) - Q(s_i, a_i|\theta_k^Q)\right)^2$$

        Compute policy gradient of the actor network

$$\nabla_{\theta_k^\mu} J \approx \frac{1}{B}\sum_{i=1}^{B}\nabla_a Q(s_i, a|\theta_k^Q)|_{a=a_i}\nabla_{\theta_k^\mu}\mu(s_i|\theta_k^\mu)$$

        Update the target network
      end for
    end for
  end for
end for

Specific to in the framework of the MFG, the behavior of the large-scale agents is described as the mean-field term, which is a statistical function use to characterize the distribution of the mass. In this case, the complexity of the system can be drastically reduced, if the associated large number of information interactions with other agents is converted into the interaction with the mass.

Firstly, we define the essential elements in the MFG:

• **Player set** $\mathcal{K}$: $\mathcal{K} = \{1, 2, \cdots, |\mathcal{K}|\}$ is the set of users in the NOMA-MEC system, where $|\mathcal{K}| = K \times N$. The number of agents $\mathcal{K}$ is arbitrarily large.

• **State Space** $\mathcal{S}_t$: In practice, the overhead of the system is increased to collect global information of all users at the SBS and then distribute the global information to the user, which reduces the computing performance. Therefore, we assume that the state of each user is only derived from its own local observations. In this paper, the state of the ith user $s_i(t) = \{\tau_i(t), h_i(t)\}$, where $\tau_i(t)$ and $h_i(t)$ represent the SINR and channel gain of the ith user, respectively. So the state space can be expressed as

$$\mathcal{S}_t = [\tau_1(t), \cdots, \tau_{|\mathcal{K}|}(t); h_1(t), \cdots, h_{|\mathcal{K}|}(t)]. \quad (22)$$

• **Action space** $\mathcal{A}_t$: Each user chooses the action from the action space based on the current state $s_t \in \mathcal{S}_t$. The action space is composed of the user's allocated power, the offloading decision and the resource allocation decision, which can be defined as

$$\mathcal{A}_t = [p_1(t), \cdots, p_{|\mathcal{K}|}(t); \lambda_1(t), \cdots, \lambda_{|\mathcal{K}|}(t); \rho_1(t), \cdots, \rho_{|\mathcal{K}|}(t)], \quad (23)$$

where $\lambda_i(t) = (\lambda_i^l(t), \lambda_i^c(t))$ is the weight coefficient of resource allocation of the ith user. $\rho_i(t) = (\rho_i^l(t), \rho_i^c(t))$ is the weight coefficient of task offloading of the ith user.

• **Reward function** $R(a_t, s_t)$: In this paper, we consider to minimize the computing cost of the system. Therefore, the reward function of user according to (21) is

$$R(a_t, s_t) = -\min \Phi'', \quad (24)$$

where $a_t \in \mathcal{A}_t$. Note that, according to the mean-field theory, all users are equal and identical, so they use the same policy.

When the user in state s_t chooses the action $(p(t), \lambda(t), \rho(t))$, the evolution of the system, also is the FPK equation, can be expressed as

$$\pi_{t+1}(s_t) = P_{ij}(a_t|s_t)\pi_t(s_t), \quad (25)$$

where $\pi_t(s_t)$ denotes the distribution of the state space at t time. $P_{ij}(a_t|s_t)$ represents the probability that the agent in state i transits to state j, which depends on the actions of the user.

According to (25), the value function of the system on the state s_{mk} in the tth slot when user chooses μ policy, also called the HJB equation, is defined as

$$V_t^\mu = \mathbb{E}_\mu\left[\sum_{t=0}^{T}\gamma^t R(a_t, \pi_t|s_t)\right]. \quad (26)$$

Next, we define the Nash equilibrium solution of the MFG.

***Definition** 1:* (*The solution of the MFG*) For each $k \in \mathcal{K}$, the optimal policy μ^* is defined as

$$\mu^* = \arg\max V_t^\mu(s_t). \quad (27)$$

Based on this policy, for all policies μ:

$$V_t^\mu(s_t) \le V_t^{\mu^*}(s_t). \quad (28)$$

So according to the optimal policy μ^*, the Nash equilibrium solution of the MFG can be obtained to meet (25) and (26).

Finally, we prove the uniqueness of MFE in ***Lemma 1***.

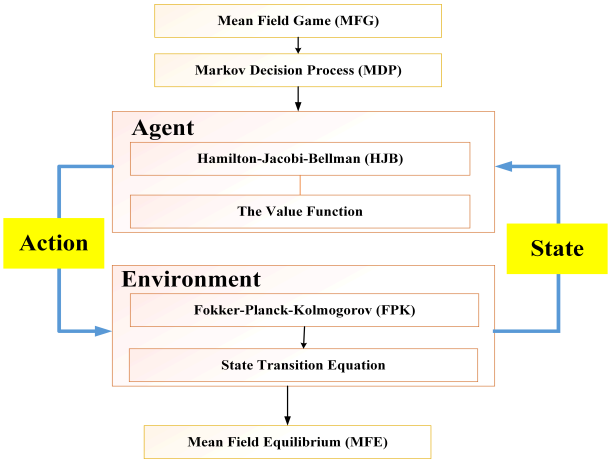


Fig. 3: The interaction process between MFG and RL.

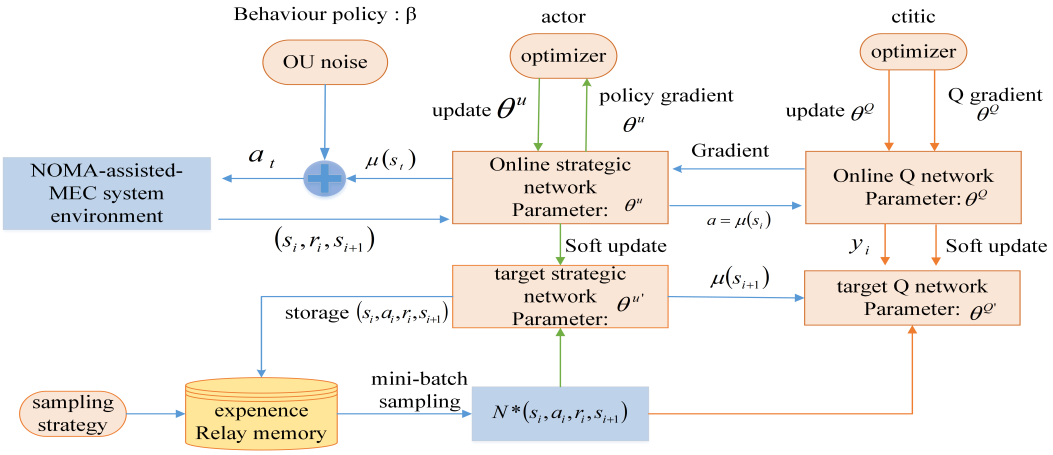


Fig. 4: The schematic framework of DDPG.

***Lemma** 1:* There is a unique equilibrium solution for the MFG.

***Proof** 1:* The value function is a continuous function of reward functions and policy, and assumes that each user has a corresponding optimal policy μ^*. Let the reward function be monotonous about the distribution of the state based on the uniqueness of the optimal policy,

$$\sum_{i=1}^{N}(\pi^2-\pi^1)(r(\mu^*,\pi^2)-r(\mu^*,\pi^1))\geq 0, \tag{29}$$

and then we can obtain the uniqueness of the MFE. It should be noted that proof of the existence of the MFG solution and the uniqueness of the equilibrium have been given in [39]. ■

In usual, the solution of MFG is obtained by the finite difference method, which divides the solution domain into a differential mesh and replaces the continuous solution domain with a finite number of mesh nodes. However, the complexity of this method is very high. The MDP is often used to solve the optimization problems, which the derive between the MDP and the MFG has given in ***Lemma 2***. Therefore, the resource allocation problem for the NOMA-MEC system can be transformed into the optimization problem of the MDP, which means we can reduce the constraints of mean-field solution and simplify the solution process. Fig. 3 also shows the interaction process between MFG and RL.

***Lemma** 2:* The value function of MFG satisfies the dynamic programming equation of the MDP.

***Proof** 2:* The value function in (26) can be rewritten as

$$V_t^{\mu}=E_{\mu}[R(a,\pi,s)+\sum_{t=1}^{T}\gamma^t R(a_t,\pi_t,s_t|s_t,\pi_t]. \tag{30}$$

Separating the two terms inside the expectation and taking outside the expectation, we get

$$V_t^{\mu}=E_{\mu}[R(a,\pi,s)]+\gamma E_{\mu}\left[\sum_{t=1}^{T}\gamma^t R(a_t,\pi_t)|s_t,\pi_t)\right]. \tag{31}$$

Given the current action a_t and state s_t, the system turns to the next state s_{t+1} with probability $P(s_{t+1}|s_t,a_t)$ when following policy μ

$$V_t^{\mu}=R(a,\pi,s)]+\gamma\sum_{s_{t+1}\in S}P(s_{t+1}|s_t,a_t)E_{\mu}\left[\sum_{t=1}^{T}\gamma^t R(a_t,\pi_t|s_t,\pi_t)\right]. \tag{32}$$

From (32), the value function of MFG satisfies the dynamic programming equation of MDP. ■

Different from other works to discretize the action space, such as DQN and Q-learning methods, this paper uses the DDPG algorithm to obtain the equilibrium solution of MFG. Specifically, it selects actions based on the learned policy in a continuous action space, which makes it possible to obtain excellent performance in complex environments. Therefore, considering that the action space is continuous of the NOMA-MEC system, we use the DDPG method to optimize the task delay and energy consumption of the users.

The schematic framework of the resource allocation in the NOMA-MEC system using the DDPG algorithm is shown in Fig. 4. It is well known that the DDPG algorithm is mainly an actor-critic framework, which comprises the actor part and the critic part. Specifically, the actor part is to minimize $Q(s,a)$ and output specific actions a through a deterministic policy μ under the premise of input observation s, the critic part is to output $Q(s,a)$ updated by the Bellman equation on the premise of input observation s and specific actions a. Then, the objective function of the DDPG algorithm can be defined as

$$J(\theta^{\mu})=E_{\theta^{\mu}}\left[\sum_{i=1}^{N}\gamma^{i-1}R_i\right], \tag{33}$$

where θ^{μ} is the parameters of the policy network that generate deterministic action, and θ^{μ} is updated by the policy gradient. In the actor part, there are mainly two networks, namely online policy network and target policy network. The deterministic policy μ is used to directly obtain the determined value of each step of action $a_t=\mu(s_t|\theta^{\mu})$. Similarly, the critic part also has two networks, i.e., the online Q network and target Q network. The Q function (i.e., the action-value function) is defined through the Bellman equation as the reward expectation value of the selection action under the deterministic policy μ, and

the Q network used in the DQN is also used in the DDPG to fit the Q function, i.e.,

$$Q^{\mu}\left(s_t, a_t\right) = E\left[R + \gamma Q\left(s_{t+1}, \mu\left(s_{t+1}\right)\right)\right], \tag{34}$$

where $Q^{\mu}\left(s_t, a_t\right)$ indicates the expected value of the return obtained by selecting the action a_t using the deterministic policy μ in state s_t. To measure the performance of a policy μ, we define the performance objective as

$$\begin{aligned} J_{\beta} &= \int_{S} \rho^{\beta}(s_t) Q^{\mu}(s_t, \mu(s_t)) ds \\ &= E_{S\sim\rho^{\beta}}[Q^{\mu}(s_t, \mu(s_t))], \end{aligned} \tag{35}$$

where β denotes a behavioral policy, ρ^{β} is the probability density function of $\mathcal{S}$. The goal of training is to maximize the performance objective $J_{\beta}\left(\mu\right)$ and minimize the loss function of Q network. In critic part, the mean square error (MSE) [40] is used as the loss function, i.e.,

$$L(\theta^{Q}) = E[R + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) - Q(s_t, a_t|\theta^{Q})]. \tag{36}$$

Therefore, the gradient of L for the θ^{Q} can be obtained based on the standard back-propagation method,

$$\begin{aligned} \frac{\partial L(\theta^{Q})}{\partial \theta^{Q}} =& E[R + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) \\ & - Q(s_t, a_t|\theta^{Q}) \frac{\partial Q(s_t, a_t|\theta^{Q})}{\partial \theta^{Q}}]. \end{aligned} \tag{37}$$

Note that the Adam optimizer [41] is adopted to update θ^{μ} and θ^{Q}, respectively, in the actor part and critic part.

Due to a large number of user and the complicated actions, the network structure is continuously deepened during the learning process. In the MF-DDPG algorithm, users choose the action according to the current state in each step to form the training data set. The large number of users and the sufficient training steps make the samples of the training data very large, which ensures the convergence of the neural network and makes users to effectively learn the optimal policy. In addition, the setting of the learning rate also affects the convergence of the network. A lager learning rate may cause overfitting problems, which leads to convergence divergence of the neural network. Therefore, we set a lower learning rate, which makes users learn more experiences. The process of resource optimization using the DDPG algorithm in the NOMA-MEC system is shown in **Algorithm 2**.

C. Jointly Iterative Optimization Algorithm of UCM and MF-DDPG

According to the introduction of the first two parts in this section, the optimal resource allocation policy of the NOMA-MEC system can be obtained by optimizing one of the variable blocks while keeping the other variables unchanged, and both sub-problems can be effectively solved. However, in order to further obtain a high-quality suboptimal solution, UCM and MF-DDPG are iterated with each other and jointly optimized, as summarized in **Algorithm 3**. During each iteration, update user clustering and resource allocation until convergence. Since the objective function is updated toward the optimal direction in each iteration, a suboptimal solution close to the global optimum can be obtained, once the objective function converges.

Algorithm 3 Jointly Iterative Optimization Algorithm (JIOA) of UCM and MF-DDPG

1: **Initialized:**
2: Equal power distribution for all users
3: $\rho_{mk}^{l} = \rho_{mk}^{c} = 0.5$, $\lambda_{mk}^{l} = \lambda_{mk}^{c} = 0.5$
4: **for** each episode $1, 2, ...,$ **do**
5: According to the initialized resource allocation scheme, the UCM algorithm is executed to obtain the user clustering scheme
6: According to the obtained user clustering scheme, the MF-DDPG is excuted to obtain the resource allocation scheme
7: Obtain the computation cost of users
8: **end for**

TABLE II: Table of symbols

Symbol	Definition
Number of SBSs	20
Number of users in each SBS	64
Maximum transmit power of each user	10dBm
Noise power	-168dBm/Hz
System bandwidth	5MHz
Pass loss model	$126.8 + 36.5\log_{10}d$
Maximum task delay	300ms-1000ms
Computing capability of kth user	0.8GHz
Computing capability of nth MEC server	6GHz
Number of CPU cycles required	600Megacycle
Input data size required	100Kbits

D. Complexity of the proposed algorithm

In this section, the computational complexity of the proposed algorithm is analyzed as follows. For the user clustering, theoretically, all users can be clustered using an exhaustive method, and then a clustering scheme that minimizes the computational cost of system is selected. The computational complexity of the exhaustive method is $O\big(N\frac{K!}{(K_m)^M}\big)$, where as the UCM algorithm proposed in this paper firstly sorts the user's channel gains, so the algorithm complexity is $O(N(MK + K\log_K))$. It can be seen that the computational complexity of the UCM algorithm is much smaller than that of the exhaustive method. Moreover,, the computational complexity is $Q(|A| \times |S|)$ when the factor of the neural network is not considered, where $|A|$ is the number of the actions and $|S|$ is the number of the states. When a deep neural network is added, the computational complexity is closely related to the system environment and parameter settings, and it is often difficult to make specific estimates. Therefore, this paper attempts to find the optimal combination of parameters through a large number of simulations. Finally, the computational complexity of the jointly iterative optimization algorithm of UCM and MF-DDPG is $O(E(N(MK + K\log_K) + |A| \times |S|))$, where E is the number of iterations. By analyzing the complexity of the algorithm, it can be found that the proposed algorithm possesses low computational complexity.

V. Numerical Simulation and Discussion

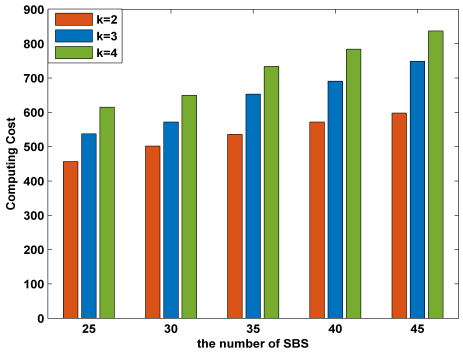
In this section, numerical results are presented to evaluate the performance of the proposed algorithm for NOMA-MEC systems in UDN. We commence with the system setup and network architecture. Following this, the performance of the proposed algorithm is discussed along with other benchmark schemes.

A. Simulation Setup

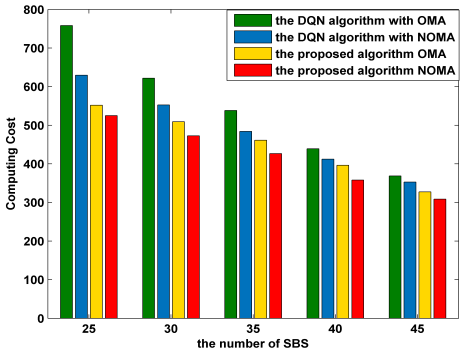
In this paper, we consider an UDN, where 20 SBSs are randomly distributed in a large area of $10\ \mathrm{km} \times 10\ \mathrm{km}$. The coverage radius of each SBS is $20\ \mathrm{m}$. 64 users are randomly distributed around each SBS. Refering to the commonly used simulation parameters in MEC systems [37], we summarize the default parameters in Table II.

To implement the MF-DDPG algorithm, the actor network and critic network use the fully-connected neural network with three hidden layers, where each hidden layer contains 300 neurons. In the actor network, the last output layer uses the Sigmoid activation function to ensure that the probability of the final action output is between 0 and 1, while in the critic network, each layer uses the ReLU activation function. The learning rates of the Actor-network and Critic-network ranges from 0.0001 to 0.001, respectively.

Moreover, in order to better verify the performance of the proposed algorithm, we compare it with DQN, which is one of the classic algorithms in RL. For the fair comparison, the DQN algorithm uses the same environment as our proposed MF-DDPG algorithm. In addition, we compare the proposed NOMA scheme with the pure OMA scheme to verify the improved performance of the NOMA scheme.



(a) Comparison of calculation costs when the number of users in NOMA cluster is different: $k = 2$, $k = 3$, $k = 4$



(b) Comparison results of the overall computation cost with different algorithms.

Fig. 5: The overall computation cost versus the number of SBSs.

B. Numerical Results

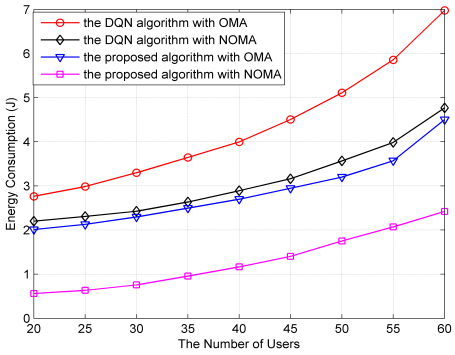
Fig. 5 (a) shows the comparison of computational cost when the number of users in a NOMA cluster is different. It can be observed that the calculation cost increases along with the value of K. This is because when the number of users in the NOMA cluster is larger, the calculation cost is increased due to the increasing in mutual interference of users. Fig. 5 (b) demonstrates the comparison results of the overall computation cost with different numbers of SBSs. We can see that the computational cost of the proposed algorithm is always the smallest under different SBS numbers. Furthermore, the difference between the proposed algorithm and other algorithms is increasing with the increase of the number of SBS, which implies the superior performance of our proposed algorithm.

Fig. 5 presents the comparison results of the overall computation cost versus different numbers of SBSs. As the number of SBS increases, the overall computational cost of the system increases. This is because the interference of the system increases along with the number of SBS increases, and hence users have to consume more energy and time to suppress the interference for offloading tasks.

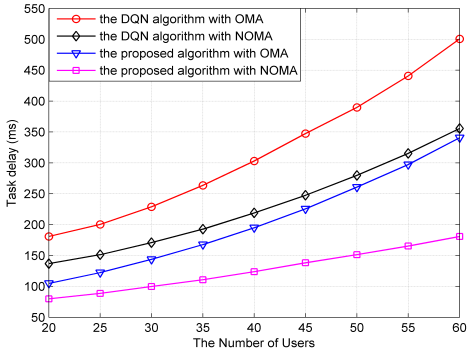
Fig. 6 (a) shows the energy consumption versus the number of users, while Fig. 6 (b) is the trend of the task delay with respect to the number of users. Again, we can see that when the number of users increases, both energy consumption and task delay are on the rise. In addition, the performance of the proposed algorithm is better than the performance of DQN, and the performance of NOMA is better than the performance of OMA. When the same number of SBSs, the energy consumption and task delay of the proposed algorithm are always the smallest.

Fig. 7 shows the trend of the reward function under the settings of different learning rates. As expected, it can be seen that the value of the learning rate imposes a very large impact on the convergence of the reward function. To elaborate, in the training process, higher learning rate may cause overfitting, which makes the reward function to fluctuate greatly or even difficult to converge. However, the reward function converges slowly when the learning rate is too small. As can be observed in Fig. 7, the reward function is optimal when the learning rate is 0.0001.

Fig. 8 portrays the convergence of the reward function using different algorithms and different multiple access methods. It can be seen that upon increasing the number of iterations, the four schemes gradually converge. Meanwhile, the convergence speed of the proposed algorithm is significantly faster than other schemes. This is because DQN discretizes the huge continuous action space, which leads to an increase in the amount of calculation and a slower convergence speed. Moreover, compared with the pure OMA scheme, the system adopts the NOMA scheme to obtain more superior performance.



(a) The energy consumption versus the number of users.



(b) The task delay versus the number of users.

Fig. 6: The energy consumption and task delay variation using different algorithms.

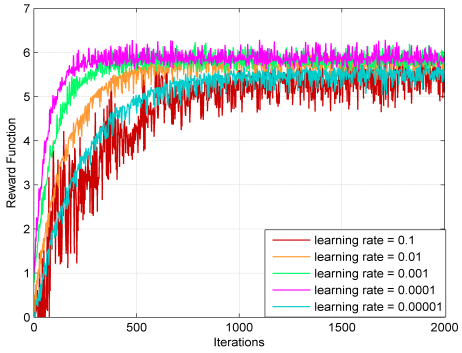


Fig. 7: Convergence performance of MF-DDPG algorithm under different learning rates.

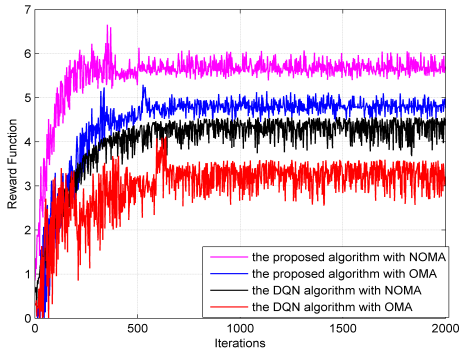
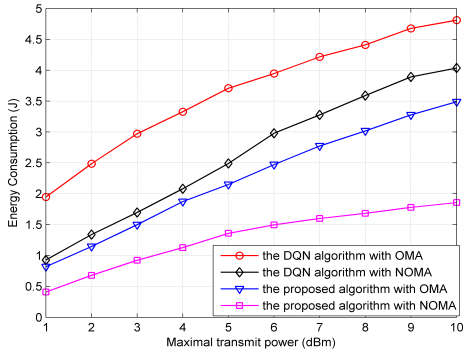
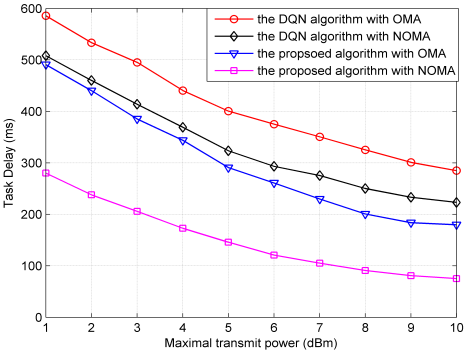


Fig. 8: Convergence performance of reward functions using different algorithms and different multiple access modes.



(a) The energy consumption versus the transmit power.



(b) The task delay versus the transmit power.

Fig. 9: The energy consumption and task delay variation using different algorithms.

Fig. 9 demonstrates the effect of the maximal transmit power under different algorithms and multiple access modes. In Fig. 9 (a), it can be observed that the energy consumption of the system gradually increases along with the maximal transmit power. When the maximal transmit power is fixed, the NOMA scheme can obtain lower energy consumption. This is because users in the NOMA cluster can use the complete spectrum resources to send information at the same time, which may reduce the energy consumption of the system. In addition, as seen from the Fig. 9b, the system delay decreases as the maximal transmit power increases. This is because when the maximal transmit power is of a large value, both the user's calculation speed and the data transmission rate become large, thereby leading to a reduced calculation delay.

Fig. 10 shows the trend of energy consumption of the system versus the CPU capacity in different algorithms and different multiple access schemes. As the CPU capacity increases, the computing ability of the system increases, resulting in a reduced energy consumption of the system. In practice, the CPU performance of MEC may be adjusted in accordance with the user's computing task requirements, for achieving reduced energy consumption.

Fig. 11 demonstrates the energy consumption versus the delay requirement. As the delay requirement increases, energy consumption of the system gradually declines. The reason is that as the delay requirement increases, the system can slow

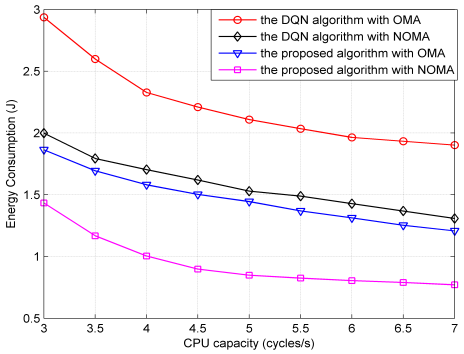


Fig. 10: The CPU capacity versus energy consumption relying on different algorithms and multiple access schemes.

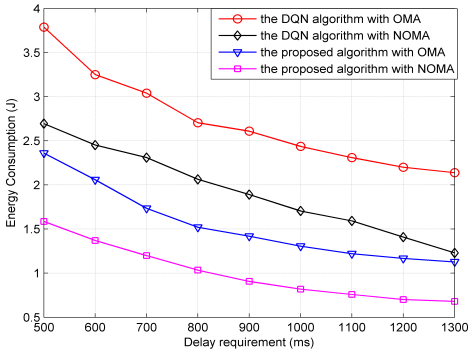


Fig. 11: The delay requirement versus energy consumption using different algorithms and multiple access schemes.

down the task calculations and consume less energy, resulting in reduced energy consumption.

VI. Conclusion

In this paper, the computation cost of the system is minimized for the NOMA-MEC system in an UDN, by optimizing its task and resource allocation. Due to the non-convexity of the original problem, we decouple it into two sub-problems: user clustering and jointly power and computation resource allocation. In order to solve the user clustering problem, a UCM algorithm based on user channel gain difference is proposed. Upon fixing the user clustering scheme, the NOMA-MEC system is modeled as the MFG theoretical framework for the intensive user deployment. Then, a low-complexity MF-DDPG algorithm is conceived for attaining the equilibrium. Finally, JIOA is proposed to jointly iterate UCM and MF-DDPG, for minimizing the computational cost of the system. Compared to the benchmark algorithms, the proposed algorithm can efficiently reduce the energy consumption and task delay of the system, and improve convergence speed. Moreover, the simulation results also show that the number of users in the NOMA cluster directly affects the computation cost of the system. Particular to the ultra-dense network environment, the proposed algorithm has an improved performance along with the increase in the number of SBSs.

Further, we will compare with more algorithms to verify the superior performance of the MF-DDPG algorithm and extend it in various wireless communication scenarios, such as data caching, task migration, and so on. Moreover, the channel estimation model is also one of the research directions to improve the practicality of the proposed algorithm considering that CSI is not perfect in practice.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065-1082, Jun. 2014.

[2] X. Ge, S. Tu, G. Mao, C. Wang, and T. Han, "5G Ultra-Dense Cellular Networks", *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72-79, Feb. 2016.

[3] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," *in IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2961-2991, Jun. 2018.

[4] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[5] M. Mohammadi, X. Shi, B. K. Chalise, Z. Ding, H. A. Suraweera, C. Zhong, and J. S. Thompson, "Full-Duplex Non-Orthogonal Multiple Access for Next Generation Wireless Systems," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 110-116, May 2019.

[6] H. Li, K. Ota, and M. Dong, "Deep Reinforcement Scheduling for Mobile Crowdsensing in Fog Computing," *ACM Transactions on Internet Technology*, vol. 19, no. 21, Apr. 2019.

[7] H. Li, K. Ota, and M. Dong, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," *IEEE Network*, vol. 32, no. 1, pp. 96-101, Feb. 2018.

[8] M. Chen, W. Saad, and C. Yin, "Liquid State Machine Learning for Resource and Cache Management in LTE-U Unmanned Aerial Vehicle (UAV) Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1504-1517, Mar. 2019.

[9] R. Williams, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," *Machine Learning*, vol. 8, pp. 229-256, May 1992.

[10] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms", *International Conference on Machine Learning*, pp. 387-395, Beijing, China, Jun. 2014.

[11] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D.Sliver, and D. Wierstra, "Continuous Control with Deep Reinforcement Learning," *Proc. International Conference Learning Representations (ICLR)*, San Juan, Puerto Rico, May, 2016.

[12] C. Yang, Y. Zhang, J. Li, and Z. Han, "Distributed Interference-Aware Traffic Offloading and Power Control in Ultra-Dense Networks: Mean Field Game With Dominating Player," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8814-8826, Sep. 2019.

[13] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Ultra Dense Small Cell Networks: Turning Density Into Energy Efficiency," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1267-1280, May 2016.

[14] M. de Mari, E. Calvanese Strinati, M. Debbah, and T. Q. S. Quek, "Joint Stochastic Geometry and Mean Field Game Optimization for Energy-Efficient Proactive Scheduling in Ultra Dense Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 766-781, Dec. 2017.

[15] J. M. Schulte, "Adjoint Methods for Hamilton-Jacobi-Bellman Equations," *Diploma Thesis*, University of Munster, Germany, Nov. 2010.

[16] D. Shi, H. Gao, L. Wang, M. Pan, Z. Han, and H. V. Poor, "Mean Field Game Guided Deep Reinforcement Learning for Task Placement in Cooperative Multi-Access Edge Computing," *IEEE Internet of Things Journal*, DOI:10.1109/JIOT.2020.2983741.

[17] S. Wang, M. Chen, X. Liu, and C. Yin, "Task and Resource Allocation in Mobile Edge Computing: An Improved Reinforcement Learning Approach," *IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, Dec. 2019.

[18] H. Guo, J. Liu, and J. Zhang,"Computation Offloading for Multi-Access Mobile Edge Computing in Ultra-Dense Networks," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 14-19, Aug. 2018.

[19] B. Dab, N. Aitsaadi, and R. Langar, "A Novel Joint Offloading and Resource Allocation Scheme for Mobile Edge Computing," *IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, Jan. 2019.

[20] H. Xing, L. Liu, J. Xu, and A. Nallanathan, "Joint Task Assignment and Resource Allocation for D2D-Enabled Mobile-Edge Computing," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4193-4207, Jun. 2019.

[21] F. Wang, H. Xing, and J. Xu, "Optimal Resource Allocation for Wireless Powered Mobile Edge Computing with Dynamic Task Arrivals," *IEEE International Conference on Communications (ICC)*, Shanghai, China, May 2019.

[22] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721-742, May 2017.

[23] Z. Ding, P. Fan, and H. V. Poor, "Impact of Non-Orthogonal Multiple Access on the Offloading of Mobile Edge Computing," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 375-390, Jan. 2019.

[24] H. Yu, Q. Wang, and S. Guo, "Energy-Efficient Task Offloading and Resource Scheduling for Mobile Edge Computing," *IEEE International Conference on Networking, Architecture and Storage (NAS)*, Chongqing, China, Oct. 2018.

[25] J. Zhu, J. Wang, Y. Huang, F. Fang, K. Navaie, and Z. Ding, "Resource Allocation for NOMA MEC Offloading," *IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, Dec. 2019.

[26] J. Xu, L. Chen, and S. Ren, "Online Learning for Offloading and Autoscaling in Energy Harvesting Mobile Edge Computing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 361-373, Sep. 2017.

[27] J. Zhang, X. Tao, H. Wu, N. Zhang and X. Zhang, "Deep Reinforcement Learning for Throughput Improvement of the Uplink Grant-Free NOMA System," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6369-6379, Jul. 2020.

[28] P. Yang, L. Li, W. Liang, H. Zhang, and Z. Ding, "Latency Optimization for Multi-user NOMA-MEC Offloading Using Reinforcement Learning," *2019 28th Wireless and Optical Communications Conference (WOCC)*, Beijing, China, May. 2019.

[29] C. Yang, J. Li, P. Semasinghe, E. Hossain, S. M. Perlaza, and H. Zhu, "Distributed Interference and Energy-Aware Power Control for Ultra-Dense D2D Networks: A Mean Field Game", *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1205-1217, Feb. 2017.

[30] P. Semasinghe and E. Hossain, "Downlink Power Control in Self-Organizing Dense Small Cells Underlaying Macrocells: A Mean Field Game",*IEEE Transactions on Mobile Computing*, vol. 15, no. 2, pp. 350-363, Feb. 2016.

[31] L. Li, Y. Xu, Z. Zhang, J. Yin, W. Chen, and Z. Han, "A Prediction-Based Charging Policy and Interference Mitigation Approach in the Wireless Powered Internet of Things," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 2, pp. 439-451, Feb. 2019.

[32] K. Hamidouche, A. T. Z. Kasgari, W. Saad, M. Bennis, and M. Debbah, "Collaborative Artificial Intelligence (AI) for User-Cell Association in Ultra-Dense Cellular Systems", *IEEE International Conference on Communications Workshops (ICC Workshops)*, Kansas, MO, May 2018.

[33] L. Li, H. Ren, Q. Cheng, K. Xue, W. Chen, M. Debbah, and Z. Han, "Millimeter-Wave Networking in Sky: A Machine Learning and Mean Field Game Approach for Joint Beamforming and Beam-Steering", *IEEE Transactions on Wireless Communications*. vol. 19, no. 10, pp. 6393-6408, Oct. 2020.

[34] T. Bai, J. Wang, Y. Ren, and L. Hanzo, "Energy-Efficient Computation Offloading for Secure UAV-Edge-Computing Systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6074-6087, Jun. 2019.

[35] M. Tawarmalani and N. V. Sahinidis, "Global Optimization of Mixed-Integer Nonlinear Programs: A Theoretical and Computational Study," *MATHEMATICAL PROGRAMMING*, vol. 99, no. 3, pp.536-591, Apr. 2014.

[36] Y. Wen, W. Zhang, and H. Luo, "Energy-Optimal Mobile Application Execution: Taming Resource-poor Mobile Devices with Cloud Clones," *Proceedings IEEE INFOCOM*, Orlando, FL, Mar. 2012, pp. 2716-2720.

[37] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei,"Energy-Efficient NOMA-Based Mobile Edge Computing Offloading," *IEEE Communications Letters*, vol. 23, no. 2, pp. 310-313, Feb. 2019.

[38] J. Kang and I. Kim, "Optimal User Grouping for Downlink NOMA," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 724-727, Oct. 2018.

[39] J. M. L. O. Gueant, and P. L. Lion, "Mean Field Games and Applications", *Paris-Princeton Lectures on Mathematical Finance*, vol. 2, pp. 205-266, Mar. 2011.

[40] Y. Sai, R. Jinxia, and L. Zhongxia, "Learning of Neural Networks Based on Weighted Mean Squares Error Function," *International Symposium on Computational Intelligence and Design*, Changsha, China, Dec. 2009, pp. 241-244.

[41] S. Bock and M. Wei, "A Proof of Local Convergence for the Adam Optimizer," *International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, Sep. 2019.

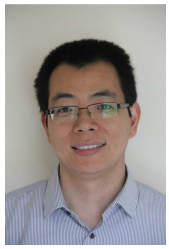
Lixin Li (M'12-) received the B.Sc. and M.Sc. degrees in communication engineering, and the Ph.D. degree in control theory and its applications from Northwestern Polytechnical University (NPU), Xi'an, China, in 2001, 2004, and 2008, respectively. He was a Post-Doctoral Fellow with NPU from 2008 to 2010. In 2017, He was a visiting scholar at the University of Houston, Texas. He is currently an Associate Professor in the School of Electronics and Information, NPU. He has authored or coauthored more than 150 peer-reviewed papers in many prestigious journals and conferences, and he holds 12 patents. His current research interests include wireless communications, game theory, and machine learning. He received the 2016 NPU Outstanding Young Teacher Award, which is the highest research and education honors for young faculties in NPU.

Qianqian Cheng is currently a master student under the supervision of Prof. Lixin Li with the School of Electronics and Information, Northwestern Polytechnical University, Xian, China. Her research interests include unmanned aerial vehicle, mobile edge computing and reinforcement learning in wireless communication networks.

Xiao Tang (S'14–M'18) received his B.S. degree in Information Engineering (Elite Class Named After Tsien Hsue-shen) and Ph.D. degree in Information and Communication Engineering from Xi'an Jiaotong University in 2011 and 2018, respectively. From September 2015 to August 2016, he worked as a visiting student at the Department of Electrical and Computer Engineering in University of Houston. He is now with the Department of Communication Engineering in Northwestern Polytechnical University. His research interests include wireless communications and networking, game theory, and physical layer security.

Tong Bai (S'15M'19) received the B.Sc. degree in telecommunications from Northwestern Polytechnical University, Xian, China, in 2013, and the M.Sc. and Ph.D. degrees in communications and signal processing from the University of Southampton, Southampton, U.K., in 2014 and 2019, respectively. Since 2019, he has been a Postdoctoral Researcher with Queen Mary University of London, London, U.K. His research interests include the performance analysis, transceiver design, and utility optimization for power-line and wireless communications as well as for Internet of Things.

Wei Chen (S05-M07-SM13) received the B.S. and Ph.D. degrees (Hons.) from Tsinghua University in 2002 and 2007, respectively. Since 2007, he has been a Faculty Member with Tsinghua University, where he is currently a Tenured Full Professor, the Director of the Degree Office, and a University Council Member. During 2014-2016, he was the Deputy Head of the Department of Electronic Engineering in Tsinghua University. From 2005 to 2007, he was a Visiting Ph.D. Student with the Hong Kong University of Science and Technology. He visited the University of Southampton in 2010, Telecom Paris Tech in 2014, and Princeton University, Princeton, NJ, USA, in 2016. His research interests are in the areas of communication theory and stochastic optimization. He is a Cheung Kong Young Scholar and a member of the National Program for Special Support of Eminent Professionals, also known as 10,000 talent program. He is a standing committee member of All-China Youth Federation as well as, the secretary-general of its education board. He received the IEEE Marconi Prize Paper Award in 2009 and the IEEE Comsoc Asia Pacific Board Best Young Researcher Award in 2011. He is a recipient of the National May 1st Labor Medal and the China Youth May 4th Medal. He has also been supported by the National 973 Youth Project, the NSFC Excellent Young Investigator Project, the New Century Talent Program of the Ministry of Education, and the Beijing Nova Program. He serves as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS. He has served as a TPC Co-Chair for IEEE VTC-Spring in 2011 and a Symposium Co-Chair for IEEE ICC and Globecom.



Zhiguo Ding (S'03-M'05-F20) received his B.Eng in Electrical Engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree in Electrical Engineering from Imperial College London in 2005.From Jul. 2005 to Apr. 2018, he was working in Queen's University Belfast, Imperial College, Newcastle University and Lancaster University. Since Apr. 2018, he has been with the University of Manchester as a Professor in Communications. From Oct. 2012 to Sept. 2020, he has also been an academic visitor in Princeton University. Dr Ding' research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an Area Editor for the *IEEE Open Journal of the Communications Society*, an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Technology*, and *Journal of Wireless Communications and Mobile Computing*, and was an Editor for *IEEE Wireless Communication Letters*, *IEEE Communication Letters* from 2013 to 2016. He received the best paper award in IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship 2012-2014, the Top IEEE TVT Editor 2017, IEEE Heinrich Hertz Award 2018, IEEE Jack Neubauer Memorial Award 2018, IEEE Best Signal Processing Letter Award 2018 ,and Web of Science Highly Cited Researcher 2019.

Zhu Han (S01M04-SM09-F14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019 and ACM distinguished Member since 2019. Dr. Han is 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."