Joint Design of Access and Backhaul in Densely Deployed mmWave Small Cells

Ziqi Guo, Yong Niu, *Senior Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, Ruisi He, *Senior Member, IEEE*, Ning Wang, *Member, IEEE*, Zhangdui Zhong, *Fellow, IEEE*, and Bo Ai, *Fellow, IEEE*

***Abstract*—With the rapid growth of mobile data traffic, the shortage of radio spectrum resource has become increasingly prominent. Millimeter wave (mmWave) small cells can be densely deployed in macro cells to improve network capacity and spectrum utilization. Such a network architecture is referred to as mmWave heterogeneous cellular networks (HetNets). Compared with the traditional wired backhaul, The integrated access and backhaul (IAB) architecture with wireless backhaul is more flexible and cost-effective for mmWave HetNets. However, the imbalance of throughput between the access and backhaul links will constrain the total system throughput. Consequently, it is necessary to jointly design of radio access and backhaul link. In this article, we study the joint optimization of user association and backhaul resource allocation in mmWave HetNets, where different mmWave bands are adopted by the access and backhaul links. Considering the non-convex and combinatorial characteristics of the optimization problem and the dynamic nature of the mmWave link, we propose a multi-agent deep reinforcement learning (MADRL) based scheme to maximize the long-term total link throughput of the network. The simulation results show that the scheme can not only adjust user association and backhaul resource allocation strategy according to the dynamics in the access link state, but also effectively improve the link throughput under different system configurations.**

***Index Terms*—User association, backhaul bandwidth allocation, joint design of access and backhaul, mmWave heterogeneous cellular network, multi-agent deep reinforcement learning.**

Manuscript received 3 August 2022; revised 16 November 2022, 9 February 2023, and 20 April 2023; accepted 9 June 2023. Date of publication 15 June 2023; date of current version 14 November 2023. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grants 2022JBXT001 and 2022JBQY004, in part by the National Key Research and Development Program of China under Grant 2020YFB1806903, in part by the National Key Research and Development Program of China under Grant 2021YFB2900301, in part by the National Natural Science Foundation of China under Grants 62221001, 62231009, U21A20445 and 61771431, and in part by the Fundamental Research Funds for the Central Universities under Grant 2023JBMC030. The review of this article was coordinated by Prof. Hongzhi Guo. *(Corresponding authors: Yong Niu; Bo Ai.)*

Ziqi Guo is with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China, and also with the Collaborative Innovation Center of Railway Traffic Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: 21120053@bjtu.edu.cn).

Yong Niu is with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China (e-mail: niuy11@163.com).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: smao@ieee.org).

Ruisi He, Zhangdui Zhong, and Bo Ai are with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Engineering Research Center of High-speed Railway Broadband Mobile Communications, Beijing Jiaotong University, Beijing 100044, China (e-mail: ruisi.he@bjtu.edu.cn; zhdzhong@bjtu.edu.cn; aibo@ieee.org).

Ning Wang is with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: ienwang@zzu.edu.cn).

Digital Object Identifier 10.1109/TVT.2023.3286428

I. Introduction

In RECENT years, mobile data traffic has grown rapidly, which leads to escalating demands for wireless spectrum. Heterogeneous cellular networks (HetNets) provide a promising solution to improve network capacity and meet the traffic demand [1]. By deploying small cells underlaying macro cell, network coverage can be greatly enhanced, and user equipment (UE) can be associated with a closer base station (BS) to obtain better quality of service (QoS). However, there is a compelling need to provide enough bandwidth and capacity to meet the growing user demand in HetNets. A promising approach for addressing this demand is spectrum expansion, i.e., exploiting the millimeter wave (mmWave) band from 30 GHz to 300 GHz [2], [3]. The mmWave band offers huge amount of bandwidth, which can be utilized to support bandwidth-intensive applications. However, due to the high carrier frequency, mmWave communications suffer from more severe propagation loss than sub-6 GHz systems. To combat the high channel attenuation, both the transmitter and receiver should adopt high gain directional antennas to achieve directional transmission. Besides, owing to the small wavelength and the directional transmission, mmWave links are sensitive to the random blockage by the obstructions in the environment.

In HetNets, because of the dense deployment of small cells, deploying wired fiber backhaul for each small cell is costly [4]. Therefore, 3GPP advances the integrated access and backhaul (IAB) architecture for 5G cellular network [5], which can provide wireless backhaul connection and support wireless access and backhaul simultaneously with shared time or frequency resources. Two scenarios have been considered for IAB by 3GPP: the in-band backhaul scenario and the out-of-band backhaul scenario [6]. In a mmWave system, the in-band backhaul scenario means that the access links and backhaul links share the same mmWave band to improve spectrum utilization and achieve closer integration. However, this resource sharing makes the available spectrum limited, and the interference between access and backhaul will be introduced into the system, which is more serious in small cells densely deployed [7], [8]. In comparison, an out-of-band backhaul scenario with two different mmWave bands allocated for the access and the backhaul will alleviate the resource pressure resulting from the dense deployment of

small cells and eliminate the backhaul-access link interference. Consequently, the network throughput can be further improved, and the design and optimization of the HetNets can be more flexible. Besides, considering the different characteristics of different frequency bands of mmWave, we can choose the appropriate frequency band for access and backhaul respectively according to the actual propagation environment, which exploits the value of mmWave band to a greater extent [9]. Therefore, this scenario has great potential in future networks with higher data rate requirements and ultra-dense heterogeneous network deployments.

For the mmWave HetNets, the performance of the system is influenced by both the access link and the backhaul link, so it is important to jointly design both of them [10]. Specifically, there are two vital problems to be solved in the design. The first one is the backhaul resource allocation. Because of the randomness and dynamics of wireless networks, the design of resource allocation strategy needs to be robust to deliver satisfactory performance with regard to the network throughput and resource utilization. The other problem is the user association. With a properly designed user association strategy, users can compare the quality of service of different BSs and choose to associate with the best choice, so as to improve the user experience. When a BS serves too many users, some of its users can be handed over to an adjacent BS to reduce the traffic load at this BS. However, in the mmWave HetNets, the user association and the backhaul resource allocation are coupled. The allocation of backhaul resources needs to be decided based on the number of users associated with different BSs, and the backhaul resources of each BS will also affect the decision of the user association. To this end, how to jointly optimize backhaul resource allocation and user association in the HetNets is a key problem to be solved [11], [12], [13], [14].

With the development of artificial intelligence, there is an increasing interest in applying promising learning based algorithms to address wireless communications problems [15], [16], [17], [18], [19]. One of such effective techniques is reinforcement learning (RL) [20]. By interacting with environment and training agents, RL can evaluate policies and adaptively select the optimal policy. The RL based algorithm does not need to know detailed information of the environment, and is adaptive to the changes in the environment. The classical RL, such as Q-learning, has been shown to achieve superior performance in small-sized systems [21]. However, as Q-learning needs to store the Q-value of each state-action pair in a Q-table, such storage and computing cost could be prohibitively high for systems with high-dimensional state and action spaces. To address this problem, a deep neural network (DNN) can be leveraged to approximate the Q-value, thus the Q-table can be replaced. Such a model is usually referred to as deep reinforcement learning (DRL) [22]. Moreover, when there are more than one agents in the system, the state of environment depends on the joint actions of all the agents [23]. Such a model, termed multi-agent reinforcement learning (MARL), has a great potential in solving distributed optimization problems such as resource allocation and user association [24].

In this article, we conduct research on the joint design of the access link and the backhaul link for the mmWave HetNets. The access link and the backhaul link adopt different mmWave frequency bands for data transmission in the two-layer HetNets, which can be considered as out-of-band IAB networks with single-hop backhaul. Small cells are densely deployed in the HetNets and the access link is considered to be affected by the random blockage. We focus on the joint optimization of the user association and the backhaul resource allocation to improve long-term total link throughput in the HetNets. We develop an effective MADRL based method, which allows each UE to select the associated small base station (SBS) and determine the backhaul resource requirements based on its state observations. The method does not require the full channel state information (CSI) and uses a distributed architecture to improve training efficiency. Based on the MADRL algorithm, our scheme can adjust the user association and backhaul resource allocation strategy according to the change of the access link blockage state in time and obtain satisfactory throughput performance. The contributions of this article are summarized as follows:

- We focus on the joint optimization of the backhaul resource allocation and the user association in a HetNet with the dense deployment of mmWave small cells. The access links and the backhaul links in the HetNet adopt two different mmWave frequency bands and the access links are subject to random blockage. We formulate the joint optimization as a mixed integer nonlinear programming (MINLP) problem, with the purpose of maximizing the long-term total link throughput in the HetNet.
- We develop a MADRL based method to solve this joint optimization problem. Specifically, we consider each UE as an agent and define the state, action, and reward for UEs. Then, with the help of the double deep Q-learning algorithm (DDQN), each UE learns an effective joint optimization policy to maximize the long-term total link throughput. Through distributed training, each UE can make decisions independently based only on partial observations of the environment and adjust the user association and backhaul resource allocation policies in time when the link blockage state changes.
- We evaluate the MADRL scheme and show that the proposed learning framework has good convergence performance. When the link blockage state changes, the network throughput performance can be guaranteed by adjusting the user association and backhaul resource allocation strategy. Besides, the scheme contributes to a better balance between access throughput and backhaul throughput, and also achieves higher network link throughput in various network scenarios with different numbers of UEs and SBSs over several baseline schemes.

The rest of the article is organized as follows. Section II reviews related work. The system model and problem formulation are presented in Section III. We present the joint design scheme based on MADRL for both radio access and backhaul network in Section IV. Our simulation results are discussed in Section V. Section VI concludes this article. We list the abbreviations commonly appeared in this article in Table I.

TABLE I
LIST OF ABBREVIATIONS

Notation	Description
BS	base station
DA	distance based user association and equal backhaul bandwidth allocation
DDQN	double deep Q-learning
DNN	deep neural network
DQN	deep Q-learning
DRL	deep reinforcement learning
HL	heuristic based user association and load based backhaul bandwidth allocation
IAB	integrated access and backhaul
LOS	line-of-sight
MADDQN	multi-agent double deep Q-learning
MADRL	multi-agent deep reinforcement learning
MARL	multi-agent reinforcement learning
MBS	macro base station
MDP	Markov decision process
MINLP	mixed integer nonlinear programming
NLOS	non-line-of-sight
ReLU	rectified linear unit
RL	reinforcement learning
SADRL	single-agent deep reinforcement learning
SBS	small base station
SINR	signal-to-interference-plus-noise ratio
SL	SNR based user association and load based backhaul bandwidth allocation
SNR	signal-to-noise ratio
TDMA	time division multiple access
UE	user equipment

II. RELATED WORK

There have been several related works studying joint resource allocation and user association for HetNets. For example, Fooladivanda et al. [25] investigated the performance of different user association policies under three predefined spectrum allocation strategies in HetNets. Lin et al. [26] focused on jointly optimizing user association and spectrum allocation for multi-tier HetNets in both downlink and uplink to maximize network utility. Chen et al. [27] aimed to maximize the system sum rate and design a distributed algorithm for jointly optimizing user association and resource allocation. Zhuang et al. [28] proposed an optimization-based framework to reduce network energy consumption in the HetNets by jointly optimizing user association and spectrum allocation. With huge available bandwidth, mmWave can be adopted in HetNets, significantly increases the network capacity. In [29] and [30], the user association and the resource allocation were jointly optimized in the HetNets with the coexistence of sub-6 GHz BSs and mmWave BSs. Liu et al. [9] investigated the joint user association and resource allocation in mmWave HetNets under two access modes: single-band access and multi-band access.

However, the above works only considered the optimization of the access side. In HetNets with wireless backhaul, the balance of the access and backhaul link throughput needs to be guaranteed for better total throughput. Therefore, it is necessary to consider a joint design of access and backhaul. Many works considered the scenarios where access and backhaul operate at the same frequency band [11], [12], [13], [14]. Liu et al. [11] formulated the joint optimization of user association and resource allocation in in-band full-duplex wireless backhaul HetNets as a MINLP problem. They decomposed the original problem and proposed an iterative algorithm to solve the MINLP problem. Khodmi et al. [12] adopted non-cooperative game theory to solve the joint power allocation and user association problems in heterogeneous ultra-dense networks (UDNs), in order to guarantee throughput balance between the access and backhaul links. Su et al. [13] adopted a distributed optimization algorithm based on primal and dual decomposition to jointly optimize the user association and the backhaul bandwidth allocation. Liu et al. [14] proposed a coalition game based joint user association and bandwidth allocation algorithm for mmWave UDNs to maximize network sum rate. However, the fact that access and backhaul share the same frequency band makes frequency resources limited and introduces backhaul-access interference. The interference is exacerbated by the dense deployment of small cells, which limits the further improvement of network throughput [7]. There has been some works focused on the HetNets with mmWave backhaul and sub-6 GHz access [31], [32], [33]. Despite the elimination of interference between the access and backhaul, the limited bandwidth of the sub-6 GHz band makes it difficult to support higher data rate transmissions. In addition, allocating different mmWave bands to access and backhaul may be another effective solution. However, few works have focused on the joint optimization of access and backhaul under this solution.

In dynamic wireless networks, it is difficult to obtain the accurate and complete information about the environment. Therefore, model-free RL has been widely used to solve optimization problems in wireless communication [24]. Feng et al. [34] applied deep Q-learning (DQN) to find the resource allocation strategy under different system states. The proposed DRL-based approach achieved effective utilization of limited backhaul resources. Wei et al. [35] focused on the user scheduling and resource allocation scheme for HetNets with hybrid energy supply. Considering the stochastic property of wireless channel conditions and renewable energy arrival rates, they proposed a policy-gradient-based actor-critic RL algorithm to obtain the optimal policy. Shen et al. [19] established a DRL framework to optimize the resource allocation and scheduling for time-sensitive traffic in a 5G system subject to mmWave channel variations. However, these works considered centralized optimization, which induces a heavy computational pressure on the central controller. Besides, the communication overhead of collecting global network information cannot be ignored either. Therefore, distributed optimization methods based on MARL are more advantageous in the large-scale dynamic HetNets with the dense deployment of small cells. For example, Zhao et al. [36] regarded each user as an agent and proposed a distributed optimization method based on MADRL to achieve the jointly optimal resource allocation and user association strategies in HetNets. Yang et al. [37] proposed a multi-agent dueling deep-Q network-based algorithm combined with distributed coordinated learning to jointly optimize device association, spectrum allocation, and power allocation in dynamic HetNets. Sana et al. [38] developed a MADRL based user association scheme under the time-varying nature of the mmWave channels for dense mmWave HetNets. However, these works [36], [37], [38] did not consider the joint design of access and backhaul.

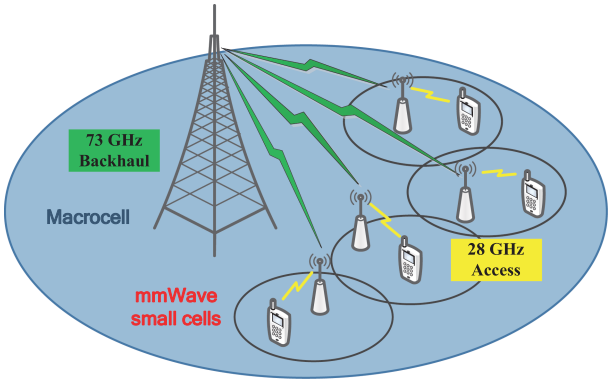


Fig. 1. Dense deployment of small cells underlying the macrocell network with different mmWave bands for access and backhaul.

Motivated by these prior works, we focus on the joint design of access and backhaul networks in densely deployed small cells, where different mmWave bands are used for access and backhaul. We propose a MADRL-based scheme for joint user association and wireless backhaul bandwidth allocation over the access and backhaul networks. The proposed scheme considers the dynamic characteristics of mmWave communications as well as the interaction between backhaul bandwidth allocation and user association in HetNets, aiming to maximize the long-term overall system throughput.

III. System Model and Problem Formulation

A. System Model

Consider a two-tier downlink HetNet composed of a macro base station (MBS), S SBSs, and N randomly located UEs, as shown in Fig. 1. The MBS is connected to the core network through optical fiber. The SBSs are densely deployed within the coverage of the MBS. Data transmission is carried out between the MBS and SBSs through mmWave backhaul links. We assume that the MBS and SBSs can be appropriately deployed to avoid the blockage of the backhaul link [10]. Thus, the stable line-of-sight (LOS) connection can be established between the MBS and each SBS without relaying through other SBSs. Each UE can be associated with a SBS and served by the associated SBS through a mmWave access link. It is worth noting that we do not consider the direct communications between the MBS and UEs. Although this is beneficial to improve the received signal strength for UEs near the MBS, few UEs can benefit from such an improvement due to the large coverage area of the macro cell and the severe path loss and signal blockage suffered by the mmWave signal. In addition, the introduction of direct communications causes more severe interference to UEs served by SBSs, thus degrading their throughput performance.

In this article, we adopt different mmWave bands for the data transmission of access and backhaul. The 28 GHz band and the 73 GHz band are used for the data transmission of access and backhaul, respectively. Both the BSs and UEs are assumed to be equipped with antenna arrays for performing directivity beamforming. Besides, since mmWave is very sensitive to the blockage of environmental obstructions, we further assume that the BSs and the UEs are also equipped with omnidirectional antennas in sub-6 GHz for reliable transmission of the transmission requests and signaling information [39]. Let $U = \{1, 2, \ldots, N\}$ and $B = \{1, 2, \ldots, S\}$ denote the sets of UEs and SBSs, respectively.

1) Access Transmisson Model: The access transmissions are performed in 28 GHz. Channel multiplexing can be adopted in the system for saving channel resources. However, the consequent interference needs to be considered in the network design. In the access network, frequency resources are multiplexed in different small cells. Time is assumed to be partitioned into multiple superframes and each superframe consists of many nonoverlapping time slots. Each SBS uses TDMA to serve its associated UEs, which allows different UEs in the small cell to occupy all the access link bandwidth for data transmission in different time slots. Moreover, each SBS evenly allocates time slots to its associated UEs in the small cell. Suppose that at each time t[1], each SBS can be associated with up to N_s UEs, and each UE can only be associated with one SBS. We use a variable $y_{ij,t}$ to indicate user association, which is defined as

$$y_{ij,t} = \begin{cases} 1, & \text{if UE i is associated with SBS j at time t} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The number of UEs associated with SBS j at time t is given by

$$N_{j,t} = \sum_{i \in U} y_{ij,t}. \tag{2}$$

In the downlink of the access network, we assume that the antenna arrays of each SBS and its served UE at each time slot perform directivity beamforming before transmission. The received power at UE i from BS j at time t, can be written as

$$p^{ac}_{ji,t} = P_S G_s(j,i) G_r(j,i) \mathcal{L}_t\left(d_{ji}\right), \tag{3}$$

where P_S is the transmit power at SBS j, $G_s(j,\ i)$ denotes the transmit antenna gain in the direction of BS $j \rightarrow$ UE i and $G_r(j,\ i)$ denotes receive antenna gain in the opposite direction, d_{ji} denotes the distance from SBS j to UE i, and $\mathcal{L}_t(d_{ji})$ reflects the path loss and shadow fading of the access link. In this article, we adopt the close-in free space reference distance path loss model at 28 GHz in [40], which is based on the real propagation measurements at 28 GHz in downtown Manhattan. Therefore, the model can better reflect the propagation characteristics of 28 GHz signals in the real environment. The path loss is expressed as

$$\mathcal{L}_t\left(d_{ji}\right)[\text{dB}] = 10\log_{10}\left(\frac{4\pi d_0}{\lambda}\right)^2 + 10\bar{n}_{ac}\log_{10}\left(\frac{d_{ji}}{d_0}\right) + X_{ac}, \tag{4}$$

where λ denotes the wavelength, d_0 is the far field reference distance, $\bar{n}_{ac}$ is the path-loss exponent and X_{ac} represents the log-normal shadowing in dB, which is a zero-mean Gaussian random variable with variance σ^2_{ac}. Both $\bar{n}_{ac}$ and σ^2_{ac} are the best fit over all measurements from the particular measurement campaign at 28 GHz in downtown Manhattan [40].

[1]Note that the time t corresponds to the time when the link blockage state changes. The change is a large-scale characteristic of the wireless channel. The time scale of a time t is much larger than the superframe length of the access link.

TABLE II
DIRECTIVITY GAIN OF UNINTENDED SBS-UE PAIR (k, i)

$G_s(k,i)\, G_r(k,i)$	Probability
$G_s^{max} G_r^{max}$	$\frac{\omega_s}{2\pi}\frac{\omega_r}{2\pi}$
$G_s^{max} G_r^{min}$	$\frac{\omega_s}{2\pi}(1-\frac{\omega_r}{2\pi})$
$G_s^{min} G_r^{max}$	$(1-\frac{\omega_s}{2\pi})\frac{\omega_r}{2\pi}$
$G_s^{min} G_r^{min}$	$(1-\frac{\omega_s}{2\pi})(1-\frac{\omega_r}{2\pi})$

Unlike sub-6 GHz frequency band, high gain directional link is always established in mmWave communication to compensate for the high path loss and penetration loss of mmWave. Thus, both UEs and SBSs are equipped with antenna arrays for directional beamforming to direct beams towards each other. In this article, for tractability of the analysis, we utilize a sector antenna model to approximate the actual antenna patterns [41]:

$$G_d(\theta) = \begin{cases} G_d^{\max}, & |\theta| \le \omega_d \\ G_d^{\min}, & |\theta| > \omega_d, \end{cases} \tag{5}$$

where $d \in \{s, r\}$, θ denotes the angle off the boresight direction, $\theta \in [-\pi, \pi)$, ω_d denotes the beamwidth of the main lobe, $G_d^{\max}$ and $G_d^{\min}$ is directivity antenna array gain of the main lobe and the side lobes, respectively. We assume that there are no alignment errors in this system, so each intended SBS-UE pair has the maximum directivity gain $G_{SBS}^{\max} G_{UE}^{\max}$. The beams of other unintended pairs are randomly oriented towards each other and uniformly distributed in $[-\pi, \pi)$. Therefore, the directivity gain of unintended pair is a discrete random variable. The four possible gain values and their corresponding probabilities are given in Table II.

Besides, due to the small wavelength and the directional transmission, the mmWave link is sensitive to the blockage of environmental obstructions like trees, buildings, and human bodies. The most significant difference between the microwave band and the mmWave band is that in many locations, especially when the distance from the transmitter is >200 meters, no mmWave signal with transmit powers between 15 and 30 dBm can be detected. It means that all the paths to the receiver in these locations are blocked by obstructions [42].

Therefore, we consider the three possible states of the access link, including line-of-sight (LOS), non-line-of-sight (NLOS), and outage [40]. The path loss in the outage state is infinite, which means the links between SBS and UE are completely blocked by environmental obstructions. The access link in the NLOS state will experience more serious path loss than in the LOS state, which is reflected in parameters $\bar{n}_{ac}$ and σ_{ac} in (4). Due to the dynamic and random nature of link blockage, three states of access link are assumed to appear randomly over time, and the path loss of the access link will follow the changes of the access link state at different time t [34]. We use $c_{ji,t}$ to indicate the state of the access link between SBS j and UE i at time t, defined as

$$c_{ji,t} = \begin{cases} 0, & \text{if the state of the access link is LOS} \\ 0.5, & \text{if the state of the access link is NLOS} \\ 1, & \text{if the state of the access link is outage,} \end{cases} \tag{6}$$

which can be estimated by BS j through the statistics of UE i signal. The set of the states of all the possible access links between UE i and all the SBSs can be denoted as

$$c_{i,t} = \{c_{1i,t}, c_{2i,t}, \ldots, c_{Si,t}\}. \tag{7}$$

The probability functions of three states is related to the distance between the transmitting and receiving antennas:

$$P_{LOS}(d_{ji}) = (1 - P_{outage}(d_{ji}))\mathrm{e}^{-a_{los} d_{ji}} \tag{8}$$

$$P_{NLOS}(d_{ji}) = 1 - P_{outage}(d_{ji}) - P_{LOS}(d_{ji}) \tag{9}$$

$$P_{outage}(d_{ji}) = \max\left(0, 1 - \mathrm{e}^{-a_{out} d_{ji} + b_{out}}\right). \tag{10}$$

where the parameters a_{los}, a_{out}, and b_{out}, and the other channel parameters corresponding to the three states can be determined by fitting the equations to the measured data in downtown Manhattan via maximum likelihood estimation [40].

In this system, we assume that TDMA is adopted in each small cell and different frequency bands is allocated for access and backhaul, so no intra-cell interference and backhaul interference exist. However, since small cells are densely deployed and different small cells reuse the access spectrum resources, it is necessary to consider the inter-cell interference. Therefore, the interference at UE i associated with SBS j at time t, which is denoted as $I_{ji,t}$, can be expressed as

$$I_{ji,t} = \sum_{k \in B \setminus \{j\}} P_S G_s(k,i) G_r(k,i) \mathcal{L}_t(d_{ki}). \tag{11}$$

The interference comes from signals sent from SBSs other than the SBS j associated with UE i, and the values of antenna gain $G_s(k,i)G_r(k,i)$ are shown in Table II.

Assume that the bandwidth of the access link is W_{ac}, and let N_0 denote the one-sided power spectra density of white Gaussian noise. The signal-to-interference-plus-noise ratio (SINR) at UE i associated with SBS j at time t can be calculated as

$$\gamma_{ji,t}^{ac} = \frac{P_S G_s(j,i) G_r(j,i) \mathcal{L}_t(d_{ji})}{N_0 W_{ac} + \sum_{k \in B \setminus \{j\}} P_S G_s(k,i) G_r(k,i) \mathcal{L}_t(d_{ki})}. \tag{12}$$

Consider all the UEs associated with the same SBS j. Then, the average throughput of one of the UEs, UE i, in the access downlink at time t is given by

$$r_{i,t}^{ac} = \sum_{j \in B} \frac{y_{ij,t} W_{ac}}{N_{j,t}} \log_2\left(1 + \gamma_{ji,t}^{ac}\right). \tag{13}$$

Since we assume that each UE is associated with one SBS, if UE i is associated with SBS j^*, i.e., $y_{ij^*,t} = 1$, the average throughput of UE i in the access link can be written as

$$r_{i,t}^{ac} = \frac{W_{ac}}{N_{j^*,t}} \log_2\left(1 + \gamma_{j^*i,t}^{ac}\right). \tag{14}$$

2) Backhaul Downlink Model: The backhaul transmissions are performed in 73 GHz. We assume that the MBS can establish directional backhaul connections with all SBSs at the same time. Each SBS is assumed to be allocated orthogonal backhaul frequency resources without interference between each other.

In the downlink of the backhaul, the antenna arrays of each MBS and all the SBSs perform directivity beamforming before transmission. The received power at SBS j at time t is

$$p_{Mj,t}^{bk} = P_M G_s(M,j) G_r(M,j) L_t(d_{Mj}), \tag{15}$$

where P_M is the transmit power of the MBS; $G_s(M,j)$ and $G_r(M,j)$ denote the transmit and receive directivity antenna gain, respectively, from the MBS to SBS j, d_{Mj} is the distance from the MBS to SBS j; and $\mathcal{L}_t(d_{Mj})$ reflects the path loss and shadow fading of the backhaul link. We adopt the close-in free space reference distance path loss model at 73 GHz in [40], which is based on the real propagation measurements at 73 GHz in downtown Manhattan. The path loss is expressed as

$$\mathcal{L}_t\left(d_{Mj}\right)[\mathrm{dB}] = 10\log_{10}\left(\frac{4\pi d_0}{\lambda}\right)^2 + 10\bar{n}_{bk}\log_{10}\left(\frac{d_{Mj}}{d_0}\right) + X_{bk}, \tag{16}$$

where $\bar{n}_{bk}$ is the path-loss exponent of the backhaul link, and $X_{bk} \sim N(0, \sigma_{bk}^2)$ is related to shadow fading in the backhaul link. Both $\bar{n}_{bk}$ and σ_{bk}^2 are the best fit over all measurements from the particular measurement campaign at 73 GHz in downtown Manhattan [40]. Considering the high gain directional backhaul link achieved by beamforming at antenna arrays, we adopt the same sector antenna in the backhaul link as in the access link. In addition, as mentioned before, we assume that with the appropriate deployment of BSs, the backhaul links can be regarded as LOS connections.

The total bandwidth of the backhaul network is W_{bk}. We assume that MBS allocates orthogonal backhaul frequency resources to each SBS, and we denote the bandwidth allocated by the MBS to the SBS j at time t as $w_{j,t}$. In this article, in order to allocate the backhaul resources more effectively, we consider the available backhaul resources for each UE. In this way, according to the access link state of each UE, different backhaul resources can be allocated to different UEs, and the allocation of backhaul resources can better adapt to the dynamic changes of access link state. Accordingly, the backhaul resources of each SBS, i.e., the bandwidth of each backhaul link, is the sum of the backhaul resources of all the associated UEs.

If UE i is associated with SBS j, the proportion of the backhaul bandwidth that is allocated to UE i at time t is given by the backhaul bandwidth factor $\beta_{ij,t} \in [0,1]$, and the amount of bandwidth assigned to UE i is $\beta_{ij,t}W_{bk}$. So the total proportion of the backhaul bandwidth allocated to SBS j at time t is

$$\beta_{j,t} = \sum_{i \in U} y_{ij,t}\beta_{ij,t}, \tag{17}$$

and the total amount of bandwidth for SBS j is

$$w_{j,t} = \beta_{j,t}W_{bk}. \tag{18}$$

In the backhaul link, due to the different frequency bands used for access and backhaul and the orthogonal resource allocation, we don't need to consider the access interference and the interference between SBSs. Besides, we consider the single MBS in our system, and assume the sufficient long distances between the SBSs in the HetNet with the neighbouring MBSs. So the signal-to-noise ratio (SNR) at SBS j at time t can be calculated as

$$\gamma_{j,t}^{bk} = \frac{P_M G_s(M,j) G_r(M,j) L_t\left(d_{Mj}\right)}{N_0 w_{j,t}}. \tag{19}$$

Then, the throughput of backhaul link between MBS and SBS j at time t is given by

$$r_{j,t}^{bk} = w_{j,t}\log_2\left(1+\gamma_{j,t}^{bk}\right). \tag{20}$$

And the achieveable throughput of UE i in the backhaul link at time t is given by

$$r_{i,t}^{bk} = \sum_{j \in B} y_{ij,t}\beta_{ij,t}W_{bk}\log_2\left(1+\gamma_{j,t}^{bk}\right). \tag{21}$$

Since we assume that each UE is associated with one SBS, if UE i is associated with SBS j^*, i.e., $y_{ij^*,t} = 1$, the average throughput of UE i in the backhaul link can be written as

$$r_{i,t}^{bk} = \beta_{ij^*,t}W_{bk}\log_2\left(1+\gamma_{j^*,t}^{bk}\right). \tag{22}$$

B. Problem Formulation

We consider both the access link and backhaul link. At time t, in order to serve UE i, the actual achievable link throughput in the HetNet is determined by the small one between the backhaul link throughput and the access link throughput, which is given by

$$R_{i,t} = \min\left(r_{i,t}^{bk}, r_{i,t}^{ac}\right). \tag{23}$$

The total actual link throughput of all the UEs at time t in the HetNet is

$$R_t = \sum_{i \in U} R_{i,t}. \tag{24}$$

In the mmWave HetNet, we aim to maximize the long-term total throughput of all the UEs over a finite period T. The joint user association and backhaul resource allocation problem can be formulated as

$$\max_{y,\beta} \sum_{t=1}^{T} R_t \tag{25}$$

$$s.t. \quad \sum_{j \in B} y_{ij,t} = 1, \ \forall i \in U, \ \forall t \tag{26}$$

$$\sum_{j \in B} \beta_{j,t} \le 1, \ \forall t \tag{27}$$

$$N_{j,t} = \sum_{i \in U} y_{ij,t} \le N_s, \ \forall j \in B, \ \forall t \tag{28}$$

$$y_{ij,t} \in \{0,1\}, \ \forall j \in B, \ \forall i \in U \tag{29}$$

$$\beta_{ij,t} \in [0,1], \ \forall j \in B, \ \forall i \in U. \tag{30}$$

This joint optimization problem can be solved by finding the optimal backhaul bandwidth factor set β and the user association indicator set y. Due to the dynamically changing blockage state of the access links, the user association and backhaul resource allocation strategy need to be adjusted in time in order to ensure good throughput performance. Constraint (26) indicates that each UE can only be associated with one SBS at a time. Constraint (27) indicates that the bandwidth resources shared by all backhaul links does not exceed the total available bandwidth resources. Constraint (28) ensures that the number of UEs served by each SBS does not exceed the upper limit of the SBS.

TABLE III
NOTATION SUMMARY

Notation	Description
S, N	number of SBSs, UEs
B, U	the set of SBSs, UEs
$y_{ij,t}$	user association indicator of UE i and SBS j at time t
$N_{j,t}$	number of UEs associated with SBS j at time t
$\beta_{ij,t}$	proportion of the backhaul bandwidth allocated to UE i at time t
$\beta_{j,t}$	proportion of the backhaul bandwidth allocated to SBS j at time t
$w_{j,t}$	backhaul bandwidth allocated to SBS j at time t
$p^{ac}_{ji,t}$	receive power at UE i from SBS j at time t
$p^{bk}_{Mj,t}$	receive power at SBS j from MBS at time t
P_S, P_M	transmit power of SBS, MBS
d_{ji}, d_{Mj}	distance between SBS j and UE i, MBS and SBS j
$\mathcal{L}_t(d_{ji})$	large scale channel gain between SBS j and UE i
$\mathcal{L}_t(d_{Mj})$	large scale channel gain between MBS and SBS j
$\bar{n}_{ac}$, $\bar{n}_{bk}$	path-loss exponent of access link, backhaul link
X_{ac}, X_{bk}	log-normal shadowing of access link, backhaul link
σ_{ac}, σ_{bk}	variance of the shadow fading X_{ac}, X_{bk}
$G_s(j,i)$	transmit antenna gain in the direction of SBS $j \rightarrow$ UE i
$G_s(M,j)$	transmit antenna gain in the direction of MBS $\rightarrow$ SBS j
G^{max}_s	transmit antenna gain of the main lobe
G^{min}_s	transmit antenna gain of the side lobe
ω_s, ω_r	main lobe beamwidth of transmit antenna, receive antenna
$G_r(j,i)$	receive antenna gain in the direction of SBS $j \rightarrow$ UE i
$G_r(M,j)$	receive antenna gain in the direction of MBS $\rightarrow$ SBS j
G^{max}_r	receive antenna gain of the main lobe
G^{min}_r	receive antenna gain of the side lobe
$c_{ji,t}$	state of the access link between SBS j and UE i
$c_{i,t}$	the set of states of all the possible access links of UE i
W_{ac}	bandwidth of the access link
W_{bk}	total bandwidth of the backhaul network
$I_{ji,t}$	interference at UE i associated with BS j at time t
$\gamma^{ac}_{ji,t}$	SINR at UE i associated with BS j at time t
$r^{ac}_{i,t}$	average access link throughput of UE i at time t
$r^{bk}_{j,t}$	backhaul link throughput of SBS j at time t
$r^{bk}_{i,t}$	achievable backhaul link throughput of UE i at time t
$R_{i,t}$	actual achievable link throughput of UE i at time t
R_t	total actual achievable link throughput at time t

Problem (25) is a MINLP problem, which is difficult to solve for the global optimal solution. Many state-of-the-art approaches are helpful to solve the MINLP problem, including heuristic-based, optimization-based, and game theory-based approaches [24]. However, in our system, the access link state changes over time, such approaches will not be accurate. They need to reconfigure to reflect the new environment. Besides, most of these approaches rely on accurate and complete knowledge of wireless environment or a large amount of information exchanged between network entities (e.g., BSs and UEs). They are difficult to perform in the large-scale HetNets. Therefore, we propose a MADRL-based method to solve these problems. The method allows each agent to learn to adapt to the environment without the knowledge of environment in advance and locally find the optimal policy through effective learning. Table III summarizes the notations adopted in this section.

IV. MULTI-AGENT DEEP REINFORCEMENT LEARNING BASED METHOD

In this article, we propose a MADRL based method to solve problem (25). Considering the difficulty in obtaining full and perfect CSI in HetNets with small cells densely deployed, this method does not require complete CSI information. Besides, through the extensive training, this method can quickly adjust the user association and backhaul resource allocation policy when the access link state changes.

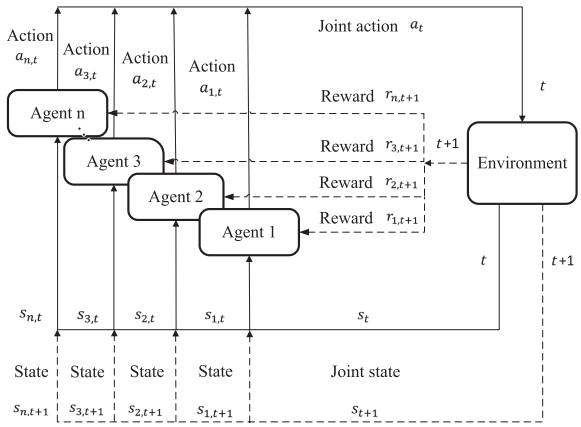


Fig. 2. The Markov game model.

Note that although single-agent deep reinforcement learning (SADRL) method also has the above advantages, it is not practical to use SADRL method for this problem. On the one hand, the state and action space will grow dramatically, which makes the convergence of the algorithm extremely difficult. Specifically, if we denote the size of action space for each agent of the MADRL method as $\mathcal{A}$, the size of action space for the agent of the SADRL method is $\mathcal{A}^N$. Thus, we can see that the growth is even more significant in the HetNets we considered with small cells densely deployed. On the other hand, the SADRL method needs a network controller with agent to collect information from all the network entities, make decisions centrally and then broadcast the decisions to each network entity. The considerable communication overhead will be intolerable in large-scale HetNets. In comparison, the MADRL method we proposed supports that each agent obtain state observations independently and make decision locally, which greatly reduces communication overhead. Besides, the MADRL method we proposed trains all the agents in parallel, thus reducing the training time.

In this section, we first establish a Markov game model corresponding to the joint optimization problem. Then we introduce DDQN algorithm and develop a multi-agent double deep Q-learning (MADDQN) method to obtain the optimal solution for the Markov game model.

A. Markov Game Model

In MADRL, the interaction between agents and the environment is usually described as a Markov game [23], as shown in Fig. 2. There are four key elements in a Markov game: (i) a finite environment state space S, (ii) a finite action space A, (iii) the reward R, and (iv) the state transition probability P. Suppose that there are n agents in the environment. At training time t, each agent i observes the current environment state $s_{i,t}$, and then takes an action $a_{i,t}$. The joint actions of all the agents in the environment are represented by a_t. Afterwards, the environment feeds back the reward for each agent according to the joint actions, while the state transition occurs simultaneously. Each agent i then receives a reward $r_{i,t+1}$ and observes a new

environmental state $s_{i,t+1}$. Then the interaction between each agent and the environment at training time t is completed. The state transition probability and reward at time $t+1$ depend only on the previous state and the previous action at time t, regardless of the earlier states and actions (i.e., satisfying the Markov condition).

Since machine learning algorithms are computationally and memory intensive, some notable efforts have already been made both in hardware design and software acceleration, which makes it possible to move the optimization process at the UE [43], [44], [45]. In this sense, [36] and [38] have treated UEs as agents and proposed distributed MARL algorithms to solve the user association problem. Thus, as for the joint optimization of user association and backhaul bandwidth allocation, we regard each UE i as an agent.

The entire mmWave HetNet system can be regarded as the environment. At each time t, after the user association and backhaul resource allocation strategies of all the UEs are adopted, the environment can generate the link throughput corresponding to each UE i and the total link throughput in the network according to the link models. This can be used as the basis for UEs' decision-making.

To transform the joint optimization problem into a Markov game, which can be solved by MADRL, we design the key elements of the corresponding Markov game below in detail.

1) State: State should fully represent the features of the network environment at different times, when different resource allocation and user association policies lead to different link throughput. Moreover, due to the dynamic characteristics of mmWave links, the access links will assume different states, which change over time. Therefore, the state of the environment should contain the total achieved link throughput in the HetNet, the achieved link throughput of each UE and the states of all the access links.

Besides, considering the constraint (27), since each UE makes decisions independently, the total backhaul resources allocated to UE may exceed the limit. Thus, each UE needs to observe the allocation of backhaul resources in the network, which is determined by actions of all the UEs at previous time. We can use β_t to represent the observation of the backhaul resource allocation at time t, which is given by

$$\beta_t = \begin{cases} \sum_{j\in B} \beta_{j,t}, & \sum_{j\in B} \beta_{j,t} \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

Therefore, with regard to each UE i, the state observation obtained from the environment consists of four parts: (i) the total achieved link throughput, (ii) the observation of the backhaul resource allocation, (iii) the achieved link throughput of UE i and (iv) the states of all the possible access links of UE i. We can write the state observation of UE i at time t as a tuple: $\{\beta_{t-1}, R_t, c_{i,t}, R_{i,t}\}$.

In addition, in MADRL, since all the agents learn to select their actions at the same time, each agent faces a non-stationary environment, which is harmful to the experience replay in DQN. Therefore, we adopt the fingerprint-based method designed in [46] to deal with this problem. The main idea is to add the estimate of other agents' policies to the state space of each agent. However, the policy of each agent includes a high-dimensional DQN, which makes it difficult to act as a part of state. Consequently, a low-dimensional fingerprint should be included in the state of each agent to track the historical trajectory of other agents' policy. As for MADRL, the policy updates of each agent are correlated with the number of training iterations, denoted by e, and the exploration rate ε in the ε-greedy strategy. Therefore, these two variables should be added to the observation space of each agent as low dimensional fingerprints.

As a result, the state observation of each agent i at time t can be designed as

$$s_{i,t} = \{\beta_{t-1}, R_t, c_{i,t}, R_{i,t}, e, \varepsilon\}, \tag{32}$$

and the joint actions of all the agents can be expressed as

$$s_t = \{s_{1,t}, s_{2,t}, \ldots, s_{N,t}\}. \tag{33}$$

2) Action: The action of each UE i consists of two parts: (i) backhaul bandwidth allocation, and (ii) user association. However, the backhaul bandwidth factor indicating the bandwidth allocation, i.e., $\beta_{ij,t}$, is a fraction in [0,1], leading to a continuous action space, which the DQN algorithm is not good at dealing with. Therefore, discretization of the action space should be considered.

We divide the total bandwidth of the backhaul network into L non-overlapping bandwidth resource blocks. Thus the backhaul bandwidth allocation problem is transformed into a bandwidth resource block allocation problem. Suppose that each bandwidth resource block can only be allocated to one UE, while a UE can occupy multiple bandwidth resource blocks. At time t, if UE i is associated with SBS j, the number of backhaul bandwidth resource blocks that can be occupied by UE i is represented by $l_{ij,t}$. Therefore, the backhaul bandwidth factors $\beta_{ij,t}$ can be written as

$$\beta_{ij,t} = \frac{l_{ij,t}}{L}. \tag{34}$$

Accordingly, the constraint (27) is equivalent to $\sum_{j\in B}\sum_{i\in U} l_{ij,t} \le L,\ \forall t$.

Then, we can use $l_{ij,t}$ to represent the allocation of backhaul bandwidth to UE i at time t in the algorithm. In order to reduce the action space and avoid the case when the backhaul bandwidth is allocated to only one or few UEs, we set an upper limit on the bandwidth resource blocks that can be occupied by each UE according to the number of UEs in the HetNet, denoted by $l_{\max}$. As a result, the action of each agent i at time t is designed as

$$a_{i,t} = \{j^*, l_{ij^*,t}\}, \tag{35}$$

where j^* denotes the index of the SBS associated with UE i, and $l_{ij^*,t}$ satisfies $l_{ij^*,t} \in [0, l_{\max}]$. Therefore, the size of the action space, denoted by $\mathcal{A}$, is the product of the size of the range $[0, l_{\max}]$ and the number of SBSs in the HetNet, i.e., $(l_{\max}+1)\times S$. The joint actions of all the agents are expressed as

$$a_t = \{a_{1,t}, a_{2,t}, \ldots, a_{N,t}\}. \tag{36}$$

3) Reward: In the HetNet scenario, all the agents make bandwidth allocation and user association decisions in order to

maximize the long-term total link throughput. However, in the training process, the total allocated backhaul resources may exceed the upper limit. At this time, constraint (27) is not satisfied. In addition to add the observation of bandwidth allocation to the state mentioned before, it is useful to constrain the reward value. If the total allocated backhaul resources are more than the total available bandwidth resources, the reward value is equal to 0. This means that, if the joint actions of all the agents do not meet the constraint (27), all the agents will not be rewarded. Besides, if the number of UEs served by the BS j is more than N_s, which means the constraint (28) is not satisfied, the UEs associated with the BS j will not be rewarded. Therefore, agents can learn to avoid the situation of not meeting the constraint in the training process.

Moreover, it seems like all the agents assume a common goal that maximize the long-term total link throughput. However, this cannot be simply regarded as fully cooperative MARL [47]. Each UE has the selfishness to improve its own throughput. Therefore, in the design of reward value, we need to consider not only the total link throughput in the network, but also the link throughput of each UE. We define δ as a factor representing the degree of selfishness of each UE. The reward of each agent i is written as

$$r_{i,t+1} = \begin{cases} \delta R_{i,t} + (1-\delta)\frac{1}{N} R_t, & \text{if } \beta_t \leq 1 \text{and } N_{j^*,t} \leq N_s \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

The larger the value of δ, the more likely UEs are to make decisions conducive to improving their own throughput. We will determine δ with the best network performance through simulation experiments.

B. Double Deep Q-Learning (DDQN) Algorithm

To solve the Markov game, it is necessary to select appropriate policy-making algorithm for each agent. The policy π refers to the mapping from state s to action a, which defines the behavior of agent. The policy is expressed by the probability of taking action a in the current state s. Considering long-term reward, the goal of the agent is to find an optimal policy that maximizes the cumulative discounted reward G_t at each training step t, which is given by

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{38}$$

where γ is the discount rate, which reflects the importance of future rewards, and r_t denotes the reward value at training step t.

To evaluate the policies, the action-value function, a.k.a. the Q-value, is adopted in Q-learning. It is defined as the expected return of the discounted reward of all possible policy sequences after taking action a in the current state s at training step t according to policy π, which can be written as

$$Q(S_t, A_t) = \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right]. \tag{39}$$

Following the Bellman equation [20], the current action-value function at training step t can be associated with the subsequent action-value function at training step $t+1$. The optimal strategy corresponds to the optimal action-value function in the finite Markov decision process (MDP) [20]. Consequently, the optimal action-value function $Q_{target}(S_t, A_t)$, which is also called the target Q-value, can be written as

$$Q_{target}(S_t, A_t) = r_{t+1} + \gamma \cdot \max_a Q(S_{t+1}, a). \tag{40}$$

The Q-value is updated with the target Q-value at each training step t, denoted as

$$Q(S_t, A_t) := Q(S_t, A_t) + \alpha \cdot [Q_{target}(S_t, A_t) - Q(S_t, A_t)], \tag{41}$$

where α represents the learning rate, which determines the updates of the Q-value.

Q-learning needs to establish a Q-table to store the Q-values of all the state-action pairs. For high-dimensional state and action spaces, the Q-table will be very large, which brings a great burden on storage and computing. To address this issue, DQN leverages a DNN to learn and approximate Q-values, which is referred to as the Q-evaluate network. Nevertheless, DNN needs a large number of labeled data for training, while RL has to generate the training data in the learning process, i.e., it does not provide a labeled dataset in advance. In addition, the training data should be uncorrelated, but RL obtains highly correlated data during its operation, which could cause the training process to be unstable. Therefore, in order to leverage a DNN in Q-learning, *experience replay* and *fixed Q-target network* are used to improve the stability of the training process [22].

Experience replay refers to storing the experience obtained from interaction with the environment at each training step t, i.e., $\{S_t, A_t, r_{t+1}, S_{t+1}\}$, in the replay memory. Then, when the Q-evaluate network needs to be updated, a mini-batch of replay memory D is randomly selected from the replay memory as training data, both the new data and historical data will be included for training, thus breaking the correlation in the data.

In addition, fixed Q-target network means the use of an independent Q-target network to generate the target Q-value. The update of the Q-target network is slower than that of the Q-evaluate network, and is achieved by replacing the parameters of the Q-target network θ^- with the parameter of the Q-evaluate network θ.

Therefore, unlike the update of Q-table in Q-learning, as in (41), DQN updates the parameter θ of the Q-evaluate network using mini-batches randomly selected from the replay memory D to minimize the following loss function.

$$L = \sum_D \left[r_{t+1} + \gamma \cdot \max_a Q(S_{t+1}, a; \theta^-) - Q(S_t, A_t; \theta) \right]^2. \tag{42}$$

Moreover, when calculating the target Q-value, both the selection and the evaluation of actions based on the maximum Q-value are estimated by the Q-target network, which is prone to overestimation. To this end, DDQN decouples the selection and evaluation to solve the above problem [48]. In the calculation of the target Q-value, action selection uses the Q-value estimated by the Q-evaluate network, and the Q-target network is only used to evaluate actions. The loss function of the DDQN is modified

as

$$L = \sum_{D} \left[r_{t+1} + \gamma \cdot Q\left(S_{t+1}, \underset{a}{\text{argmax}}\, Q(S_{t+1}, a; \theta); \theta^{-} \right) - Q(S_t, A_t; \theta) \right]^2. \quad (43)$$

C. Multi-Agent Double Deep Q-Learning (MADDQN) Method

Taking DDQN as the policy-making algorithm of each agent, we design MADDQN method to solve the Markov game model. We assume the Markov game model is episodic. Each episode includes T steps. Each step t corresponds to the time t. The state of the access link between each UE and each SBS changes over training time steps, so that agents can learn the link blockage pattern and adaptly make optimal decisions. Our goal is to maximize the sum of the total link throughput in an episode, corresponds to (25).

In this article, we mainly consider the design of joint optimization scheme that can effectively cope with random and dynamic link blockage. In fact, the network may change in other aspects at the same time, such as the rapid movement of UEs and the increase or decrease of the number of UEs, etc. If we spend too much time for training, these changes in the network during the training stage may affect the results of the training. Thus, we adopt a distributed architecture to design algorithm training, each UE can train its Q network locally. From the perspective of the entire network, training is done in parallel, greatly reducing the training time. Besides, the increase in computing power is rapid, so the time cost of training models will be greatly reduced in the future.

The detailed training procedure is presented in Algorithm 1. The agent at each UE i has two dedicated DQNs: the Q-evaluate network and the Q-target network. At each episode, all the agents' states are first initialized. The UE-SBS association has not been established in the initial state, so the link throughput of all the agents are zero at this time. Then, at each training step t, after each UE agent observes the environment state $s_{i,t}$, the Q-evaluate network uses it as input and outputs the Q-values of all the state-action pairs in the current state. The choice of action is based on the Q-values with the ε-greedy policy, which means that action is randomly chosen with probability ε while the action with the maximum Q-value is chosen with probability $(1-\varepsilon)$.

However, if all the agents randomly select $l_{ij,t}$ between $[0, l_{\max}]$, the number of total backhaul resource blocks allocated to UEs in the network may be all the values between 0 and $Nl_{\max}$. The probability of L resource blocks being fully utilized is only $\frac{1}{Nl_{\max}+1}$. Actually, we hope that L resource blocks can be utilized as much as possible, and L should be the maximum number of resource blocks available. Therefore, we implement ε-greedy strategy centrally. Specifically, all the UEs are consistent in the mode of selecting actions, i.e., randomly or according to the output of Q-evaluate network. The mode is controlled by a network controller located in the MBS.

If the network controller decides to select actions according to the output of the Q-evaluate network, all the UEs can make action selection decisions independently according to their Q-evaluate network. Then, each UE sends the access request and the number of required backhaul resources to the SBS that the UE has chosen. If the number of UEs associated with the SBS satisfies constraint (28), the SBS accepts the access request and sends a feedback signal to the UE to establish an access connection. Otherwise, the SBS rejects the access request and the access throughput of all associated UEs is 0. Then, all the SBSs send the total backhaul resource requirements of the associated UEs to the MBS. The MBS determines whether the total backhaul resource allocation satisfies the constraint (27). If satisfied, the MBS allocates the corresponding backhaul resources to SBSs and feeds back the information of the allocated total backhaul resources to each UE. If not, the backhaul resources will not be allocated. At this time, the throughput of all the UEs is 0.

If actions are decided to be selected randomly, the network controller randomly generates rational action selection for each UEs. Specifically, the user association policy should meet the constraint (26), while the backhaul resource allocation policy

Algorithm 1: The MADDQN Method for Joint Optimization of User Association and Resource Allocation in the HetNet.

```
Initialize the parameters of the Q-evaluate network and the Q-target network of all the UE agents randomly;
Initialize the replay memory at each agent;
for each episode e do
    Initialize s_t;
    for each step t do
        for each agent i do
            Observe environment state s_{i,t};
            The Q-evaluate network uses s_{i,t} to choose action a_{i,t} from A with the centralized ε-greedy policy;
        end for
        All agents take actions and obtain reward r_{i,t+1};
        Update access link states c_{ji,t};
        for each agent i do
            Observe environment state s_{i,t+1};
            Store {s_{i,t}, a_{i,t}, r_{i,t+1}, s_{i,t+1}} in the replay memory D_i;
        end for
    end for
    for each agent i do
        Sample minibatch D_i from the replay memory randomly;
        Calculate the loss function using the Q-target network;
        Update the parameters of the Q-evaluate network using stochastic gradient descent;
        For every C episodes, update the parameters of the Q-target network;
    end for
end for
```

should meet:

$$\sum_{j\in B}\sum_{i\in U} l_{ij,t} = L, \tag{44}$$

so as to make sure all the backhaul resources can be fully utilized. The MBS can directly determine the allocation of the backhaul resource based on the action selection and send the action selection scheme to each SBS. Then, each SBS sends the action selection scheme to the previous associated UEs. After receiving the new scheme, each UE sends an access request to the corresponding SBS. The SBS directly accepts the request and sends a feedback signal to the UE to establish an access connection.

After all the agents take actions at each training step t, the achievable link throughput of each UE can be observed locally. Each UE can send its achievable link throughput to the MBS through the associated SBS. The MBS will calculate the current total achievable link throughput, and then feed it back to each UE, which makes each UE can calculate its complete reward $r_{i,t+1}$. When the access link state of UE i changes, each SBS j estimates its current access link state $c_{ji,t}$, and sends the state to UE i. It should be noted that taking into account the blockage characteristics of the mmWave link, we use the sub-6 GHz band to reliably transmit access requests and signaling information. Therefore, we can see that through information exchange, the state observations of each UE at step $t+1$ can be completely obtained, which can be used as the basis for action selection at step $t+1$. Then, the tuple $\{s_{i,t}, a_{i,t}, r_{i,t+1}, s_{i,t+1}\}$ of each UE i is stored in its replay memory for the update of the Q network.

Furthermore, at the end of each episode e, a mini-batch of memory D_i is randomly sampled from the replay memory as training data for the Q-evaluate network at each UE agent i. With the target Q-value calculated by the Q-target network, the loss function of the Q-evaluate network can be written as

$$L_i = \sum_{D_i}[r_{i,t+1} + \gamma \cdot Q(s_{i,t+1}, \underset{a}{\operatorname{argmax}}\, Q(s_{i,t+1}, a; \theta_t); \theta_t^-) - Q(s_{i,t}, a_{i,t}; \theta_t)]^2. \tag{45}$$

Stochastic gradient descent is used to update the Q-evaluate network. The Q-target network of each UE agent i is updated in every C episodes according to the fixed Q-target.

After completing training of the multi-agent system, the trained Q-evaluate network of each UE i can be used to make user association and backhaul bandwidth allocation decisions in the current system scenario. At this time, each agent i can independently choose action without the assistance of the network controller. Specifically, when the access link state changes, each UE agent i uses the state observation with e and ε from the last training episode as input to the Q-evaluate network. According to the output of the Q-evaluate network, the action with the maximal Q-value is chosen by each UE i. The information exchange in the process of taking action and observing state is similar to the Q-value based action selection in the training stage. Thereafter, at each time t, all the UEs will be associated with the corresponding SBS and the corresponding backhaul bandwidth will be allocated to each SBS.

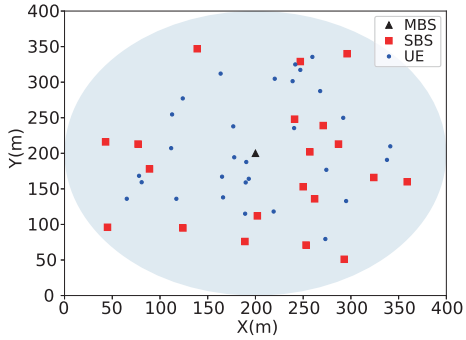


Fig. 3. The deployment scenario of the HetNet used in our simulation study.

TABLE IV
SIMULATION PARAMETERS

Parameter	Value
MBS transmit power P_0	40 dBm
SBS transmit power P_j	30 dBm
Backhaul total bandwidth W_{bk}	6 GHz
Number of bandwidth resource blocks L	300
Access total bandwidth W_{ac}	1 GHz
Noise power density N_0	-114 dBm/MHz
Maximum number of UEs served by each BS N_s	20
BS antenna main lobe gain G_{MBS}^{max}, G_{SBS}^{max}	20dB, 10dB
BS antenna side lobe gain G_{MBS}^{min}, G_{SBS}^{min}	-10dB, 0dB
UE antenna main and side lobe gains G_{UE}^{max}, G_{UE}^{min}	5dB, 0dB
Main lobe Beamwidths ω_{MBS}, ω_{SBS} and ω_{UE}	30°, 40°, 50°
28 GHz LOS channel parameters $\bar{n}_{ac}$ and σ_{ac}	2.1, 3.6 dB
28 GHz NLOS channel parameters $\bar{n}_{ac}$ and σ_{ac}	3.4, 9.7 dB
73 GHz LOS channel parameters $\bar{n}_{bk}$ and σ_{bk}	2.0, 4.2 dB
28 GHz link state parameters $\frac{1}{a_{out}}$, b_{out}, and $\frac{1}{a_{los}}$	50 m, 1.8, 50 m

V. SIMULATION RESULTS AND DISCUSSIONS

In this section, we evaluate the performance of the proposed MADDQN method through simulations. Specifically, we first investigate the impact of the selfish factor δ on the performance of our scheme. Then, we compare our scheme with the other four schemes in terms of the throughput performance.

A. Simulation Setup

Consider a two-tier HetNet as shown in Fig. 3, where the SBSs and UEs are randomly distributed within a radius of 200 m centered at the MBS. In the HetNet, small cells are densely deployed, and we set the density of SBSs to be more than 100 BSs/km^2 accordingly [49], [50]. MmWave communications are used in the HetNet, where the access links and the backhaul links use the 28 GHz band and the 73 GHz band, respectively. The link parameters are determined by fitting the equations to the measured data in downtown Manhattan via maximum likelihood estimation [40]. The settings of simulation parameters are summarized in Table IV.

In the MADDQN method, the Q-evaluate network of each UE has the same structure as the Q-target network, i.e., having three fully connected hidden layers with 400, 350, and 300

neurons, respectively. The replay memory size is 150,000 and the minibatch size is 1,000. Rectified linear unit (ReLU), i.e., $\mathrm{ReLU}(x) = \max(x, 0)$, is adopted as the activation function. The RMSProp optimizer [51] is used to update the Q-evaluate network, where the learning rate is set to 0.0001. The discount rate γ is set to 0.9. The total training steps T of each episode is 1,000, and the total number of episodes E is set on the basis of ensuring the convergence of the algorithm. The exploration rate ε is set to attenuate linearly from 1 to 0.002 with the increase of episode e over the first $80\%E$ training episodes and be stable at 0.002 afterwards.

It is worth noting that the dynamic changes of the system are reflected in the dynamic and random access link blockage, which results in dynamic changes in the access link state $c_{ji,t}$ over time. The probability of each access link state is given by (8-10). For the proposed MADDQN method, we first perform algorithm training using Algorithm 1 and then test the trained MADDQN. In the training stage, the state of the access link between each UE and each SBS changes once at each training time step. Since each episode contains 1,000 training steps, the access link states change 1,000 times in each episode, which facilitates the MADDQN algorithm to learn the underlying correlation between the dynamics of the link blockage, the joint optimization strategies, and the system throughput performance. In the testing stage, we test the performance of the algorithm over 1,000 time steps, and the access link state between each UE and each SBS changes with time steps. At each time step, all UEs make quick decisions with their trained Q-networks based on the current access link state.

In order to show the advantage of our proposed algorithm in the improvement of link throughput, the proposed user association and backhaul bandwidth allocation scheme based on trained MADDQN is compared with the three baseline schemes for HetNets and one baseline schemes for the macrocell:

1) *HL* (heuristic based user association and load based backhaul bandwidth allocation): the user association scheme is proposed in [52]. Since we do not consider optimization of power and beam width, we adapt the algorithm as shown in Algorithm 2. When the access link between SBS j and UE i is not in the outage state, we treat it as a feasible user association. The algorithm first orders all the feasible user associations according to their respective SNRs, and then validates in order whether the current association can increase the total access link throughput of the HetNet. The MBS allocates backhaul bandwidth proportional to the load on each small cell.
2) *SL* (SNR based user association and load based backhaul bandwidth allocation): each UE is associated with the SBS which can provide the maximum SNR, and the MBS allocates backhaul bandwidth proportional to the load on each small cell. If the number of UEs requested to associate with a SBS exceeds N_s, the SBS selects N_s UEs with the maximum SNR for access, and other UEs select the SBS providing the maximum SNR among the remaining SBSs to associate.
3) *DA* (distance based user association and equal backhaul bandwidth allocation): Each UE is associated with the nearest SBS, and the backhaul bandwidth is evenly allocated to each SBS. If the number of UEs requested to associate with a SBS exceeds N_s, the SBS selects the nearest N_s UEs for access, and other UEs select the nearest SBS among the remaining SBSs to associate.
4) *MBS-only* (MBS serves UEs directly): There are no small cells deployed in the macro cell. Each UE is associated with the MBS. The MBS communicates with UEs at 28 GHz band, and allocates equal bandwidth to each UE.

Algorithm 2: Heuristic scheme for User Association.

1: Set $y_{ij,t} = 0, \forall i \in U, \forall j \in B$ at time t;
2: Get the $SNR_{ij,t}$ if the access link state $c_{ji,t} \neq 1$ at time t;
3: Sort the $SNR_{ij,t}$ values in descending order into $Z_t = \{z_{1,t}, z_{2,t}, \ldots, z_{k,t}, \ldots, z_{K,t}\}$. $k = \phi(i, j)$ denotes the mapping between k and the SBS-UE pair (i, j);
4: Initialize $R_t^{ac} = 0$;
5: **while** $k \leq K$ **do**
6: Set $y_{k,t} = 1$;
7: Compute $R_t^{ac}(k)$;
8: **if** $R_t^{ac}(k) > R_t^{ac}(k-1)$ and constraint (28) is satisfied **then**
9: Set $y_{k,t} = 1$;
10: **else**
11: Set $y_{k,t} = 0$;
12: **end if**
13: **end while**

B. Choice of the Selfish Factor

Since the choice of the selfish factor δ defined in (37) has a significant impact on the throughput performance of our scheme, we now investigate it for better choice.

First, in Fig. 4, we test the effect of different δ on the convergence performance of the algorithm. There are 30 UEs, 20 SBSs and a MBS in the HetNet as Fig. 3 shows, and the number of training episodes E is set to 3,000. The average link throughput of each episode represents the average of the link throughput of 1,000 training steps in this episode. If the action selection at a training step does not satisfy the constraint (27) and (28), we set the throughput of this training step to 0. As we can see, when δ is 0 and 0.2, the fluctuation of convergence curve is small, and it can converge stably in about 2400 training episodes. However, as δ increases, the convergence curve fluctuates more and more sharply. In addition, the convergence value of the average link throughput is maximum at $\delta = 0.2$ and decreases with increasing δ when $\delta > 0.2$. To better explain these observations, we test the trained MADDQN of different δ in the same scenario. The test results are shown in Figs. 5 and 6.

Fig. 5 shows the average total link throughput of effective decisions achieved by the trained MADDQN over 1,000 time steps under different δ. The effective decision means that the user association and the backhaul bandwidth allocation strategy made by MADDQN can meet the constraints (27) and (28). It can be seen that with the increase of δ, the average total link throughput

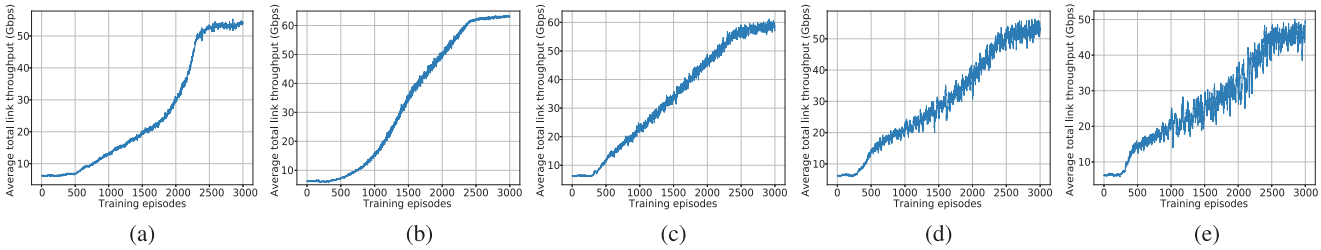


Fig. 4. Convergence performance on the choice of δ for MADDQN. (a) $\delta = 0$. (b) $\delta = 0.2$. (c) $\delta = 0.5$. (d) $\delta = 0.8$. (e) $\delta = 1$.

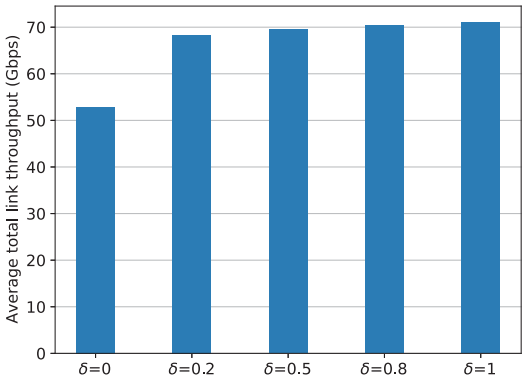


Fig. 5. Average total link throughput of effective decisions achieved by MADDQN under different δ.

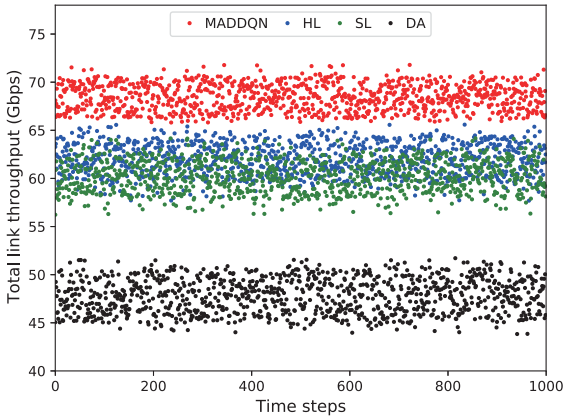


Fig. 7. Total link throughput with different time steps.

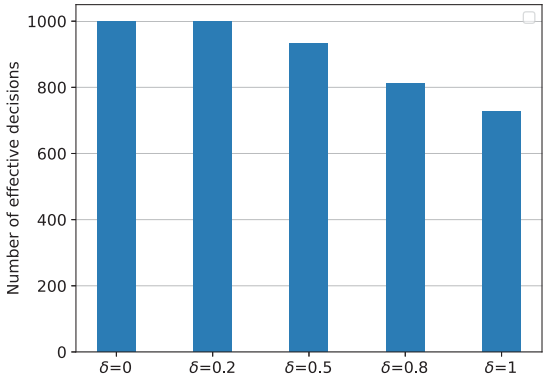


Fig. 6. Number of effective decisions of MADDQN under different δ.

for effective decisions is improved. This is because each UE can make action selections independently by its Q-network based on its state observations, and does not have access to information about other UEs' action selections. So the backhaul resource blocks in the network are difficult to be fully utilized. In other words, Equation (44) is hard to hold. As δ increases, UEs tend to occupy more resource blocks to improve their throughput $R_{i,t}$. This is reflected in the larger $l_{ij,t}$ in action selection. The utilization of backhaul resources in the network is more sufficient at this time, thus improving the average throughput of the network.

However, the increase in δ also brings a problem. Fig. 6 shows the number of effective decisions made by MADDQN in 1,000 time steps under different δ. As we can see, when δ is 0 and 0.2, all the 1,000 decisions are effective decisions. But when δ increases to 0.5, 0.8, and 1, the proportion of effective decisions decreases to 93.2%, 81.1%, and 72.8%, respectively. This is because as δ increases, the UE's aggressive resource block occupancy strategy causes the allocated backhaul resources to exceed the total available backhaul resources in the network, i.e., $\sum_{j\in B}\sum_{i\in U} l_{ij,t} > L$. For the actual communication system, this situation will cause network congestion, which is unacceptable. Based on the observations in Figs. 5 and 6, we can see that setting δ to 0.2 achieves a balance between improving throughput and satisfying the backhaul resource constraint. Therefore, $\delta = 0.2$ is adopted in the subsequent simulations.

C. Comparison With Other Schemes

Fig. 7 shows the total link throughput performances of the four schemes at 1,000 time steps in the scenario shown in Fig. 3. As we can see, since the access link state changes over time steps, the total link throughput of all the scheme varies at different time steps. This intuitively reflects the impact of the dynamic changes in the access link state on the system performance. Besides, in 1,000 time steps, MADDQN can always achieve the highest total link throughput compared with the other three baseline schemes. This means that our algorithm can better adapt to the dynamic mmWave scenario. Compared with the other three schemes, the average total link throughput of MADDQN of 1,000 time steps achieves improvements of 10.1%, 13.2%, and 42.7%, respectively.

In addition, as shown in Fig. 7, the performance of the HL scheme is similar to that of the SL scheme. This is because both HL and SL schemes make user association decisions based on

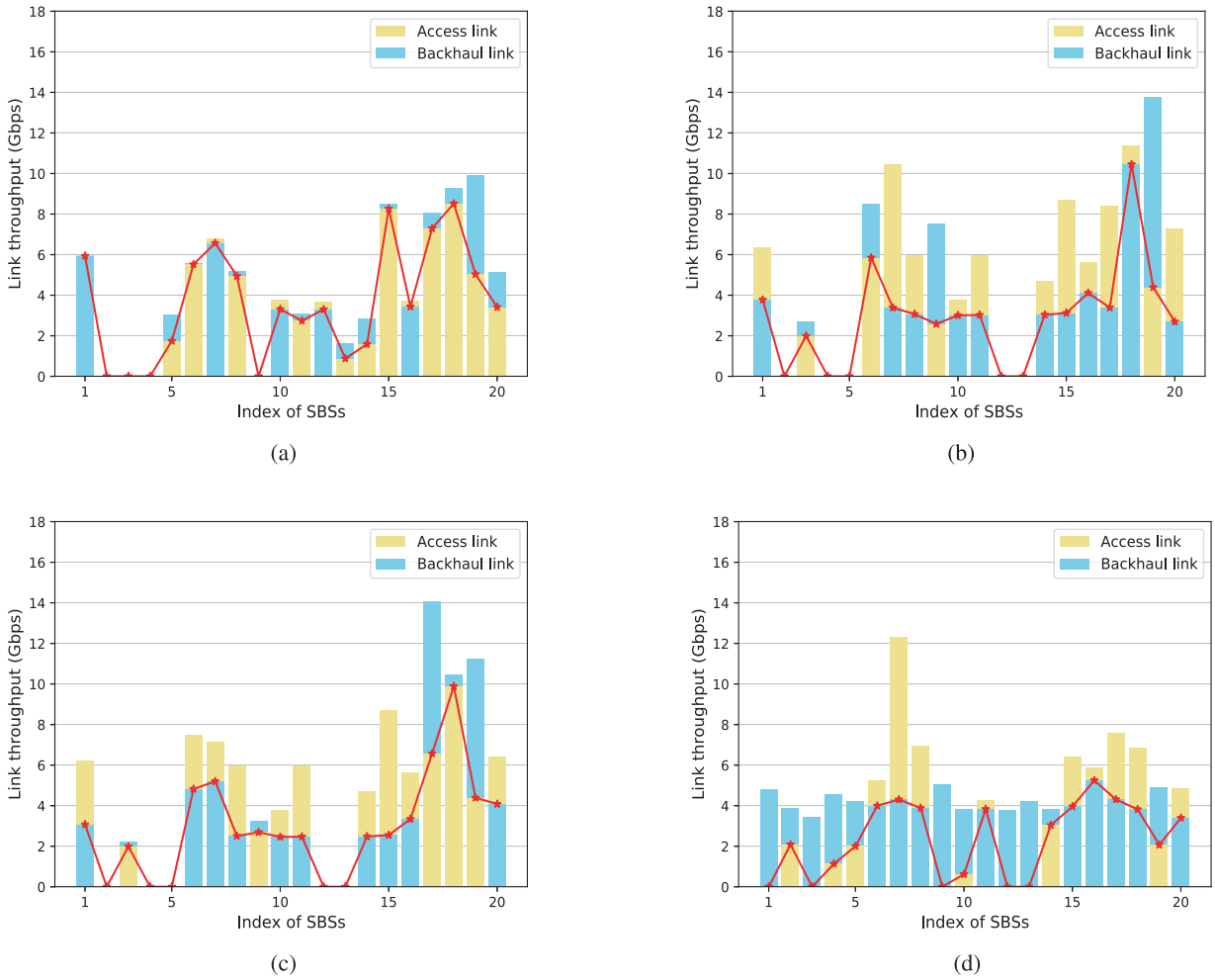


Fig. 8. Link throughput of each SBS at one time step under the four schemes: (a) MADDQN, (b) HL, (c) SL and (d) DA.

the SNR information of the network, and their backhaul resource allocation schemes are the same. The HL scheme, however, takes into account the impact of the user association strategy on the throughput of the access links, so the throughput performance is slightly better than that of the SL scheme. As for the DA scheme, it cannot adjust the strategies according to the network dynamics, so the performance is the worst.

Then, under different schemes, we test the access and backhaul link throughput of each SBS at one time step as Fig. 8 shows. The red mark in Fig. 8 represents the actual link throughput of each SBS, which is determined by the small one between the backhaul link throughput and the access link throughput. In Fig. 8, we can observe whether the access link throughput and the backhaul link throughput are matched for each SBS. From an overall view, MADDQN is most advantageous in balancing the access and backhaul link throughput, which is attributed to the joint design of access and backhaul. As Fig. 8(a) shows, SBSs with higher access throughput can always be allocated more backhaul resources to obtain higher backhaul link throughput. Accordingly, higher actual link throughput of these SBSs can be achieved. This explains why MADDQN has the best throughput performance in Fig. 7.

In comparison, the SL and HL scheme adopts the load based backhaul resource allocation strategy. However, due to the difference between the access link states of different UEs, the SBS that serves more UEs does not necessarily achieve higher access link throughput, such as the SBS 9 and 19 in Fig. 8(b) and the SBS 17 and 19 in Fig. 8(c). Allocating more backhaul resources to these SBSs can not increase the actual link throughput. On the contrary, it is harmful to the increase of the actual link throughput because of reducing the available backhaul resources for the remaining SBSs. As we can see in Fig. 8(b) and Fig. 8(c), the actual link throughput of some SBSs is only half of the access link or backhaul link throughput, which limits the actual link throughput of the HetNet.

Furthermore, the user association and backhaul resource allocation of the DA scheme are independent of each other. In Fig. 8(d), since SBS 1, 3, 9, 12, and 13 have no UE associated, the actual link throughput of these SBSs is equal to 0. As the backhaul resources are evenly allocated among the SBSs, the backhaul resources on these SBSs are not utilized. At the same time, the evenly allocated backhaul resources can not meet the requirements of the remaining SBSs with high access link throughput, such as SBS 7, 8, 15, 17, and 18. Therefore, the

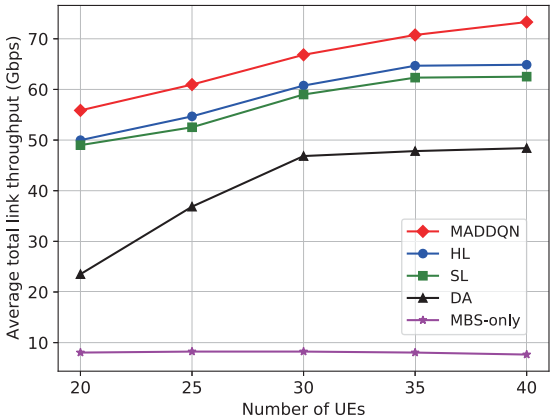


Fig. 9. Average total link throughput under different numbers of UEs.

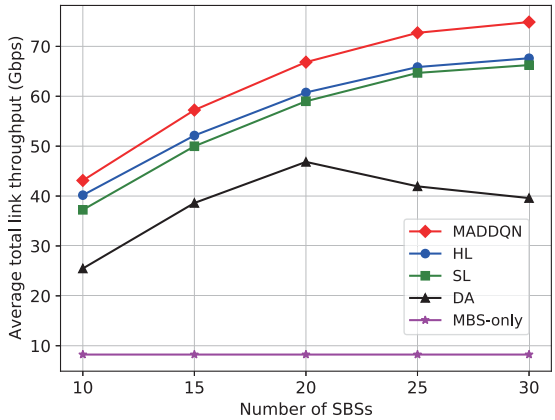


Fig. 10. Average total link throughput under different number of SBSs.

actual link throughput of these SBSs is limited by the backhaul resources.

In order to test the scalability of our MADDQN scheme, we evaluate the total link throughput performance in different scenarios. First, under different numbers of UEs, the throughput performance of the five schemes, including four schemes for HetNets and one scheme for macrocell, is shown in Fig. 9. There are 20 SBSs and a MBS in the HetNet as shown in Fig. 3. When the number of UEs is 20, 25, 30, 35, and 40, $l_{\max}$ is set to 18, 15, 12, 10, and 9, respectively. The number of training episodes E is set to 3,000. In each scenario, after the training of MADDQN, we test the performance of the trained MADDQN and the other four chemes over 1,000 time steps. The total link throughput of these 1,000 time steps is averaged to obtain the average total link throughput, which is used to evaluate the throughput performance of five schemes. As shown in Fig. 9, the average total link throughput of MADDQN under different numbers of UEs is always higher than that of the other four baseline schemes. When the number of UEs is 20, 25, 30, 35 and 40, the improvements in the total link throughput of MADDQN over HL are 11.8%, 11.5%, 10.1%, 9.39%, and 12.99%, respectively.

Besides, we can see that with the increasing number of UEs, the average total link throughput of the four schemes for HetNets increases. However, due to the limited resources of the HetNet, the improvements of the four schemes get smaller when more UEs are served by the HetNet. In contrast with the other three schemes, the network throughput growth of the MADDQN scheme is less constrained. The reason is that the MADDQN scheme optimizes the user association and backhaul bandwidth allocation jointly, which allows for more flexible and efficient utilization of the resource in the HetNet. In addition, in Fig. 9, we can see that the dense deployment of mmWave small cells can provide a significant boost in system throughput compared to the MBS-only scheme. For example, when the number of UEs is 30, the average total link throughput of MADDQN scheme for HetNet is 7.88 times higher than that of the MBS-only scheme. This is because with the dense deployment of small cells, the distance from UEs to SBSs can be reduced, which achieves higher SNR and enhances network coverage. Besides, UEs in different small cells can be served in the same spectrum through frequency reuse, so the network spectrum efficiency can be improved.

Finally, we examine the performance of the four schemes for HetNets under different numbers of SBSs, and the results are presented in Fig. 10. The MBS-only scheme is also used as a comparison. There are 30 UEs and a MBS in the HetNet as Fig. 3 shows. The number of SBSs is set to 10, 15, 20, 25, and 30, respectively and the number of training episodes E is 3,000. Similarly, performance of five schemes is evaluated with the average total link throughput of 1,000 time steps. As shown in Fig. 10, the average total link throughput of MADDQN under different numbers of SBSs is higher than that of the other four baseline schemes. When the number of SBSs is 10, 15, 20, 25, and 30, the improvements in average total link throughput of MADDQN over HL are 7.3%, 9.7%, 10.1%, 10.4% and 10.7%, respectively.

In addition, under different number of SBSs in the HetNet, the throughput of the four schemes for HetNets is always significantly greater than that of the MBS-only scheme. In the HetNet, with the increasing number of SBSs, there are more available access bandwidth resources for UEs due to the frequency reuse in different small cells. Besides, UEs will find SBSs closer to them with better service quality. The access load of each SBS is also reduced. So as we can see in Fig. 10, the throughput of the MADDQN, HL and SL schemes can be improved. Compared with the MBS-only scheme, higher throughput gains can be achieved with these three schemes when there are more SBSs in the HetNet. However, since the increase in the number of small cells causes more inter-cell interference and the backhaul resources in the network are limited, the improvement of the total link throughput gradually slows down with the increase in the number of SBS.

Unlike the above three schemes, the average total link throughput of the DA scheme decreases with the increasing number of SBSs when there are more than 20 SBSs in the HetNet. The reason is that when there are more SBSs in the HetNet, the available backhaul resources for each SBS become less due to the equal backhaul bandwidth allocation. Besides, there may be more SBSs with no UE association in the HetNet, which causes more serious waste of backhaul resources. Although the access

link throughput has been improved by the dense deployment of SBSs, the actual total link throughput is limited by the backhaul link throughput.

In contrast with the other four schemes, the MADDQN scheme we proposed guarantees the balance of the access and backhaul throughput with the joint design of access and backhaul, so it can provide greater performance gains than the other four schemes when more SBSs are deployed in the HetNet. This also reflects that MADDQN is more suitable for the mmWave small cells dense deployment scenarios with a large number of SBSs.

VI. Conclusion

In this article, we investigated the problem of user association and backhaul bandwidth resource allocation in two-tier mmWave HetNets, where small cells are densely deployed and two different mmWave bands are allocated to access and backhaul. We formulated the joint user association and backhaul resource allocation problem and then transformed the problem into a Markov game. A joint design scheme based on MADRL was proposed for the maximization of the long-term total link throughput. The proposed scheme treated each UE as an agent and allowed each UE to learn the optimal policy autonomously by DDQN based on its state observations. Through extensive training, each UE can dynamically adjust the policy to the time-varying link state. Simulation results showed that the proposed MADRL scheme could adapt to the dynamic mmWave link states, and achieve high total link throughput under various system configurations, and outperformed three baseline schemes.

Due to the appropriate deployment of SBSs, we assume the LOS transmission between MBS and SBSs in this article. However, the mmWave backhaul link may also be affected by link blockage in real communication systems. Therefore, we will consider a more realistic channel model for backhaul in future work. Besides, multi-hop wireless backhaul helps to further improve network coverage and combat mmWave link blockage, enabling more flexible backhaul connectivity. In the IAB networks with multi-hop backhaul, the backhaul path selection is introduced and the end-to-end latency, throughput, and fairness are critical performance metrics. The joint design of access and backhaul in such networks is an open research problem. Moreover, we will evaluate the impact of introducing direct communications between MBS and UEs on the system performance and design an effective joint optimization algorithm for the system. In addition, we will include UE mobility and changes in the number of UEs as additional cases of environment dynamics for future work.

References

[1] M. N. Islam, A. Sampath, A. Maharshi, O. Koymen, and N. B. Mandayam, "Wireless backhaul node placement for small cell networks," in *Proc. 48th Annu. Conf. Inf. Sci. Syst.*, 2014, pp. 1–6.

[2] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," in *Springer Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, Apr. 2015.

[3] R. Baldemair et al., "Ultra-dense networks in millimeter-wave frequencies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 202–208, Jan. 2015.

[4] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1Gbps/UE in cellular systems: Understanding ultra- dense small cell deployments," *IEEE Commun. Surv. Tut.*, vol. 17, no. 4, pp. 2078–2101, Fourthquarter 2015.

[5] M. Polese et al., "Integrated access and backhaul in 5G mmWave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, Mar. 2020 .

[6] 3rd Generation Partnership Project (3GPP), "Study on Integrated Access and Backhaul," 3GPP, Sophia Antipolis technology park, France, Tech. Rep. 38.874, 2018.

[7] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth partitioning and downlink analysis in millimeter wave integrated access and backhaul for 5G," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8195–8210, Dec. 2018.

[8] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3251–3268, May 2016.

[9] R. Liu, Q. Chen, G. Yu, and G. Y. Li, "Joint user association and resource allocation for multi-band millimeter-wave heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8502–8516, Dec. 2019.

[10] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2052–2069, Oct. 2015.

[11] Y. Liu, L. Lu, G. Y. Li, Q. Cui, and W. Han, "Joint user association and spectrum allocation for small cell networks with wireless backhauls," *IEEE Wireless Commun. Lett.*, vol. 5, no. 5, pp. 496–499, Oct. 2016.

[12] A. Khodmi, S. B. Rejeb, N. Agoulmine, and Z. Choukair, "A joint power allocation and user association based on non-cooperative game theory in an heterogeneous ultra-dense network," *IEEE Access*, vol. 7, pp. 111790–111800, Aug. 2019.

[13] Z. Su et al., "User association and wireless backhaul bandwidth allocation for 5G heterogeneous networks in the millimeter-wave band," *China Commun.*, vol. 15, no. 4, pp. 1–13, Apr. 2018.

[14] Y. Liu, X. Fang, P. Zhou, and K. Cheng, "Coalition game for user association and bandwidth allocation in ultra-dense mmWave networks," in *Proc. IEEE/CIC Int. Conf. Commun. China*, 2017, pp. 1–5.

[15] N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surv. Tut.*, vol. 21, no. 4, pp. 3133–3174, Fourthquarter 2019.

[16] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surv. Tut.*, vol. 24, no. 1, pp. 1–30, Firstquarter 2022.

[17] X. Shen et al., "AI-Assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 45–66, 2020.

[18] K. Xiao, S. Mao, and J. K. Tugnait, "TCP-Drinc: Smart congestion control based on deep reinforcement learning," *IEEE Access J.*, vol. 7 pp. 11892–119042019.

[19] S. Shen, T. Zhang, S. Mao, and G.-K. Chang, "DRL-based channel and latency aware radio resource allocation for 5G service-oriented RoF-mmWave RAN," *IEEE/OSA J. Lightw. Technol.*, vol. 39, no. 18, pp. 5706–5714, Sep. 2021.

[20] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, U.K.: MIT Press, 2018.

[21] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 5, no. 3/4, pp. 279–292, May 1992.

[22] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[23] G. Tesauro, "Extending Q-Learning to general adaptive multi-agent systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 871–878.

[24] A. Alwarafy, M. Abdallah, B. S. Çiftler, A. Al-Fuqaha, and M. Hamdi, "The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 322–365, 2022.

[25] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.

[26] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: A. utility perspective," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1025–1039, Jun. 2015.

[27] Y. Chen, J. Li, W. Chen, Z. Lin, and B. Vucetic, "Joint user association and resource allocation in the downlink of heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5701–5706, Jul. 2016.

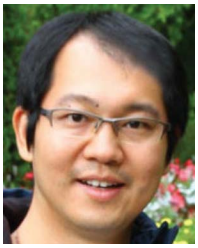
[28] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 823–831, Apr. 2016.
[29] C. Chaieb, Z. Mlika, F. Abdelkefi, and W. Ajib, "On the user association and resource allocation in hetnets with mmWave base stations," in *Proc. IEEE 28th Annu. Int. Symp. Personal, Indoor, Mobile Radio Commun.*, 2017, pp. 1–5.
[30] K. Khawam, S. Lahoud, M. E. Helou, S. Martin, and F. Gang, "Coordinated framework for spectrum allocation and user association in 5G HetNets with mmWave," *IEEE Trans. Mobile Comput.*, vol. 21, no. 4, pp. 1226–1243, Apr. 2022.
[31] W. Hao, M. Zeng, Z. Chu, S. Yang, and G. Sun, "Energy-efficient resource allocation for mmWave massive MIMO HetNets with wireless backhaul," *IEEE Access*, vol. 6, pp. 2457–2471, 2018.
[32] S. Ni, J. Zhao, H. H. Yang, and Y. Gong, "Enhancing downlink transmission in MIMO HetNet with wireless backhaul," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6817–6832, Jul. 2019.
[33] S. Aboagye, A. Ibrahim, and T. M. N. Ngatched, "Frameworks for energy efficiency maximization in HetNets with millimeter wave backhaul links," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 83–94, Mar. 2020.
[34] M. Feng and S. Mao, "Dealing with limited backhaul capacity in millimeter-wave systems: A deep reinforcement learning approach," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 50–55, Mar. 2019.
[35] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.
[36] N. Zhao, Y. -C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.
[37] H. Yang, J. Zhao, K. -Y. Lam, Z. Xiong, Q. Wu, and L. Xiao, "Distributed deep reinforcement learning-based spectrum and power allocation for heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6935–6948, Sep. 2022.
[38] M. Sana, A. De Domenico, W. Yu, Y. Lostanlen, and E. Calvanese Strinati, "Multi-agent reinforcement learning for adaptive user association in dynamic mmWave networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6520–6534, Oct. 2020.
[39] Y. Niu et al., "Energy-efficient scheduling for mmWave backhauling of small cells in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2674–2687, Mar. 2017.
[40] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.
[41] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.
[42] M. R. Akdeniz et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.
[43] T. S. Ajani, A. L. Imoize, and A. A. Atayero, "An overview of machine learning within embedded and mobile devices optimizations and applications," *Sensors*, vol. 21, no. 13, Jun. 2021, Art. no. 4412.
[44] Y. Deng, "Deep learning on mobile devices: A review," in *Proc. SPIE Mobile Multimedia/Image Process., Secur., Appl.*, 2019, pp. 52–66.
[45] J. Lee et al., "On-device neural net inference with mobile GPUs," 2019, *arXiv:1907.01989*.
[46] J. Foerster et al., "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1146–1155.
[47] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," in *Innovation in Multi-Agent Systems and Applications-1*, vol. 310, D. Srinivasan and L. C. Jain Eds. Berlin, Germany: Springer, 2010, pp. 183–221.
[48] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.
[49] T. Zhang, J. Zhao, L. An, and D. Liu, "Energy efficiency of base station deployment in ultra dense HetNets: A stochastic geometry analysis," *IEEE Wireless Commun. Lett.*, vol. 5, no. 2, pp. 184–187, Apr. 2016.
[50] T. Ding, M. Ding, G. Mao, Z. Lin, A. Y. Zomaya, and D. López-Pérez, "Performance analysis of dense small cell networks with dynamic TDD," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9816–9830, Oct. 2018.
[51] S. Ruder, "An overview of gradient descent optimization algorithms," Sep. 2016, *arXiv:1609.04747*.
[52] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, "Deep learning-based beam management and interference coordination in dense mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 592–603, Jan. 2019.

Ziqi Guo was born in Shandong, China, in 2000. He received the B.E. degree in communication engineering from Beijing Jiaotong University, Beijing, China, in 2021. He is currently working toward the M.S. degree with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University. His research interests include mmWave wireless communications and reconfigurable intelligent surface.

Yong Niu (Senior Member, IEEE) received the B.E. degree in electrical engineering from Beijing Jiaotong University, Beijing, China, in 2011, and the Ph.D. degree in electronic engineering from Tsinghua University, in 2016. From 2014 to 2015, he was a Visiting Scholar with the University of Florida, Gainesville, FL, USA. He is currently an Associate Professor with State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University. His research interests include networking and communications, including millimeter wave communications, device-to-device communication, mediumaccess control, and software-defined networks. He was a Technical Program Committee Member for IWCMC 2017, VTC 2018-Spring, IWCMC 2018, INFOCOM 2018, and ICC 2018. He was the Session Chair of IWCMC 2017. He was the recipient of the Ph.D. National Scholarship of China in 2015, Outstanding Ph.D. Graduates and Outstanding Doctoral Thesis of Tsinghua University in 2016, Outstanding Ph.D. Graduates of Beijing in 2016, Outstanding Doctorate Dissertation Award from the Chinese Institute of Electronics in 2017, and the 2018 International Union of Radio Science Young Scientist Award.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University, Brooklyn, NY, USA, in 2004. He is currently a Professor and Earle C. Williams Eminent Scholar Chair in electrical and computer engineering with Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grid. He is a Distinguished Lecturer of IEEE Communications Society (2021-2022), IEEE Council of RFID (2021-2022), Distinguished Lecturer (2014-2018), and a Distinguished Speaker of IEEE Vehicular Technology Society (2018-2021). He is with the Editorial Board of IEEE/CIC China Communications, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, ACM GetMobile,IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE MULTIMEDIA, IEEE NETWORK, and IEEE NETWORKING LETTERS. He was the co-recipient of the 2021 IEEE Internet of Things Journal Best Paper Award, 2021 IEEE Communications Society Outstanding Paper Award, IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, IEEE ComSoc MMTC 2018 Best Journal Paper Award and 2017 Best Conference Paper Award, Best Demo Award of IEEE SECON 2017, Best Paper Awards of IEEE GLOBECOM 2019, 2016, and 2015, IEEE WCNC 2015, IEEE ICC 2013, and 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a Member of the ACM.

Ruisi He (Senior Member, IEEE) received the B.E. and Ph.D. degrees from Beijing Jiaotong University (BJTU), Beijing, China, in 2009 and 2015, respectively. Since 2015, he has been with State Key Laboratory of Advanced Rail Autonomous Operation, BJTU, where he has been a Full Professor since 2018. He has been a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA, University of Southern California, Los Angeles, CA, USA, and Universit Catholique de Louvain, Leuven, Belgium. He has authored and coauthored five books, three book chapters, more than 200 journal articles and conference papers, as well as several patents. His research interests include wireless propagation channels, railway and vehicular communications, and 5G and 6G communications. Dr. He is a Member of the European Cooperation in Science and Technology. He is the Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE Antennas and Propagation Magazine, IEEE COMMUNICATIONS LETTERS, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY. He is also the Lead Guest Editor of the IEEE JOURNAL ON SELECTED AREA IN COMMUNICATIONS and IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION. He is the Early Career Representative (ECR) for Commission C, International Union of Radio Science (URSI). He was the recipient of URSI Issac Koga Gold Medal in 2020, IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2019, URSI Young Scientist Award in 2015, and five best paper awards in conferences.



Ning Wang (Member, IEEE) received the B.E. degree in communication engineering from Tianjin University, Tianjin, China, in 2004, the M.A.Sc. degree in electrical engineering from The University of British Columbia, Vancouver, BC, Canada, in 2010, and the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2013. From 2004 to 2008, he was with the China Information Technology Design and Consulting Institute, as a Mobile Communication System Engineer, specializing in planning and design of commercial mobile communication networks, network traffic analysis, and radio network optimization. From 2013 to 2015, he was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia. Since 2015, he has been with the School of Information Engineering, Zhengzhou University, Zhengzhou, China, where he is currently an Associate Professor. He also holds adjunct appointments with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, and the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada. His research interests include resource allocation and security designs of future cellular networks, channel modeling for wireless communications, statistical signal processing, and cooperative wireless communications. He was with the technical program committees of international conferences, including the IEEE GLOBECOM, IEEE ICC, IEEE WCNC, and CyberC. He was the Finalist of the Governor Generals Gold Medal for Outstanding Graduating Doctoral Student from the University of Victoria in 2013.

Zhangdui Zhong (Fellow, IEEE) received the B.E. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 1983 and 1988, respectively. He is currently a Professor and an Advisor of Ph.D. students with Beijing Jiaotong University, where he is also the Chief Scientist of State Key Laboratory of Advanced Rail Autonomous Operation. He is the Director of the Innovative Research Team, Ministry of Education, Beijing, and the Chief Scientist of the Ministry of Railways, Beijing. He is an Executive Council Member of the Radio Association of China, Beijing, and the Deputy Director of the Radio Association, Beijing. His research interests include wireless communications for railways, control theory, and techniques for railways, and GSM-R systems. His research has been widely used in railway engineering, such as the Qinghai-Xizang railway, Datong CQinhuangdao Heavy Haul railway, and many high-speed railway lines in China. He has authored and coauthored seven books, five invention patents, and more than 200 scientific research papers in his research area. He was the recipient of the Mao YiSheng Scientific Award of China, Zhan TianYou Railway Honorary Award of China, and Top 10 Science/Technology Achievements Award of Chinese Universities.

Bo Ai (Fellow, IEEE) received the M.S. and Ph.D. degrees from Xidian University, Xian, China, in 2002 and 2004, respectively. He was an Excellent Postdoctoral Research Fellow with Tsinghua University, Beijing, China, in 2007. He was a Visiting Professor with the EE Department, Stanford University, Stanford, CA, USA, in 2015. He is currently with Beijing Jiaotong University, Beijing, China, as a Full Professor and a Ph.D. Candidate Advisor. He is also the Deputy Director of State Key Laboratory of Advanced Rail Autonomous Operation and International Joint Research Center. He is one of the main responsible people for Beijing Urban rail operation control system International Science and Technology Cooperation Base, the Member of the Innovative Engineering based jointly granted by Chinese Ministry of Education and the State Administration of Foreign Experts Affairs. He has authored and coauthored eight books and more than 300 academic research papers in his research area. He has held 26 invention patents. He has been the Research Team Leader of 26 national projects and was the recipient of some important scientific research prizes. He has been notified by Council of Canadian Academies (CCA), that based on Scopus database. He has been listed as one of the Top 1 authors in his field all over the world. He has also been Feature Interviewed by *ELECTRONICS LETTERS (IET)*. His research interests include the research and applications of channel measurement and channel modeling and dedicated mobile communications for rail traffic systems.