Network Access Selection for URLLC and eMBB Applications in Sub-6GHz-mmWave-THz Networks: Game Theory Versus Multi-Agent Reinforcement Learning

Nguyen Thi Thanh Van, Nguyen Le Tuan, Nguyen Cong Luong, Tien Hoa Nguyen, *Member, IEEE*, Shaohan Feng, Shimin Gong, Dusit Niyato, *Fellow, IEEE* and Dong In Kim, *Fellow, IEEE*

***Abstract*—We investigate a heterogeneous network (HetNet) including sub-6GHz base stations (BSs), mmWave BSs, and THz BSs to support enhanced mobile broadband (eMBB) users and ultra-reliable low-latency communication (URLLC) users. We particularly investigate a user-centric network in which the users locally and dynamically select and switch among BSs over time to achieve their highest utility. Two types of users have different Quality of Service (QoS) requirements. Thus, we design two types of utility functions specifically for the eMBB users and URLLC users. Then, to model the dynamic selection behavior of the users, we propose to use a fractional game with the power-law memory. The fractional game allows the eMBB users and the URLLC users to incorporate their past strategies into their current selection, thus improving their utility. Furthermore, we consider the case that the BSs cooperate with each other, and we model the network selection of the users as a multi-agent problem. Then, we propose to use a multi-agent deep reinforcement learning (MADRL) algorithm that enables the URLLC users and eMBB users to make their network selection decision online to achieve their long-term utility. Various simulation results are provided to demonstrate the scalability and effectiveness of the proposed approaches. Particularly, compared with the classical game, the fractional game is able to achieve a higher utility but incurs a higher network adaptation cost. Moreover, the different types of URLLC users (in terms of latency and reliability requirements) and the number of URLLC users in the network significantly affect the total utility and the network selection strategies of the eMBB users. Importantly, given the cooperation among the BSs, the the MADRL outperforms both the classical and fractional games in terms of total network utility.**

***Index Terms*—URLLC and eMBB, sub-6GHz-mmWave-THz communications, dynamic network selection, fractional game, multi-agent deep reinforcement learning.**

I. Introduction

Ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) are promising services for critical requirements on wireless transmission of emerging applications in the next generation wireless networks [2]. In particular, the eMBB service accommodates applications that requires the data transmission during a long time interval with high data rate. The eMBB service is thus used for emerging AR/VR media, ultra HD, and 360^0 streaming video. Meanwhile, URLLC services are to transmit data with short length that requires ultra-high reliability with a low packet error rate, i.e., 10^{-7}, and low latency [3]. The URLLC services are thus suitable for mission critical applications such as autonomous vehicles and remote surgery. There are a number of previous works, e.g., [3]–[7], investigating the network system with the coexistence of the URLLC users and eMBB users. The common system model of these works consists of a single sub-6 GHz base station (BS) that serves multiple URLLC and eMBB users. Then, the authors investigated an optimization problem that aims to optimize resources including spectrum and power to maximize the data rate of the eMBB users while guaranteeing the QoS requirements of the URLLC users. Here, the QoS of the URLLC users are in terms of latency, data rate, and decoding error probability. Then, different algorithms are used to solve the resource decision problem. For example, the work in [5] proposed a Lyapunov-based optimization framework, and the work in [3] proposed to leverage a deep reinforcement learning (DRL) to find the optimal policy on the resource decision accounting for the variations of the arrived URLLC packets and the channels.

The aforementioned works focus on resource allocation in which there is a single sub-6 GHz BS serving both eMBB users and URLLC users. Given the rapid growth of emerging eMBB and URLLC applications such as virtual realization and autonomous control systems as well as their high density, supporting the eMBB services and the URLLC services imposes challenges to the existing sub-6GHz bands due to the limited bandwidth. Meanwhile, mmWave and THz bands are able to provide the vast available bandwidth of tens of GHz. Thus, the mmWave and THz bands well satisfy the requirements of the eMBB services and URLLC services. The work in [8] proposed to deploy multiple mmWave BSs to serve the URLLC and eMBB users. However, the study is limited to some tens

N. T. T. Van is with the Faculty of Electrical and Electronic Engineering, PHENIKAA University, Hanoi 12116, Vietnam, Vietnam. E-mail: van.nguyenthithanh@phenikaa-uni.edu.vn.

N. C. Luong is with the Faculty of Computer Science, PHENIKAA University, Hanoi 12116, Vietnam. E-mail: luong.nguyencong@phenikaa-uni.edu.vn.

T. H. Nguyen and L. T. Nguyen are with the School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam. E-mails: hoa.nguyentien@hust.edu.vn, tuan.nl214130@sis.hust.edu.vn

S. Feng is with the School of Information and Electronic Engineering (Sussex Artificial Intelligence Institute), Zhejiang Gongshang University, Hangzhou 310018, China. E-mail: feng_shaohan@mail.zjgsu.edu.cn.

S. Gong is with the School of Intelligent Systems Engineering, Sun Yat-sen University, China. Email: gongshm5@mail.sysu.edu.cn.

D. Niyato is with the College of Computing and Data Science, Nanyang Technological University, Singapore. Email: dniyato@ntu.edu.sg.

D. I. Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea. E-mail: dongin@skku.edu.

An earlier version of this paper [1] was presented at the IEEE Wireless Communications and Networking Conference (WCNC), Dubai, United Arab Emirates, in April 2024

of the users. A recent work in [9] investigated a heterogeneous network (HetNet) consisting of Sub-6GHz, mmWave, and THz BSs. However, heterogeneous users consisting of eMBB and URLLC types are not considered.

In this paper, we investigate a HetNet including sub-6GHz base stations (BSs), multiple mmWave BSs, and multiple THz BSs to support the users using the URLLC service, namely URLLC users, and the users using the eMBB service, i.e., eMBB users. The sub-6GHz BSs that is able to provide long communication range can guarantee the requirements on network coverage of the users, while the mmWave and THz BSs that can provide large bandwidth can support the requirement on high data rate of the users. However, the coexistence of multiple network tiers including sub-6GHz, mmWave, and THz communications as well as different types of users including URLLC users and eMBB users raises resource management issues including network access selection as well as user association due to the following reasons. First, different network tiers may belong to different network providers who offer different prices for using their network services. Second, there exist different types of users, i.e., URLLC users and eMBB users, that have different critical requirements on transmission. Specifically, the URLLC users require low latency and ultra-reliability, while the eMBB users aim to maximize the data rate. Third, there is a high density of BSs that are different in network capacity and network coverage, which increases the frequent handover of the users. To coordinate the user-BS association, centralized optimization algorithms can be used. However, they typically require a large number of overhead communications and face high computation complexity, especially in the large-scale system.

To address the above challenge, decentralized approach based on evolutionary game has been recently used. Particularly, evolutionary game models the evolution process of network selection strategies of the users. By using the replicator dynamics equations, evolutionary game enables the user to locally and dynamically select the network to obtain the higher utility. An equilibrium is obtained as the users have no incentive to switching their network strategies. Since each user only performs the network selection, the computation complexity of the game at the user side has a low computation complexity with $\mathcal{O}(1)$, which is suitable to be used in large-scale systems. Evolutionary game has been recently been used in several works, e.g., [10]–[15] for the network user access and resource allocation. Particularly, the work in [10] proposed the evolutionary game for the service provider selection of the eMBB users in a HetNet. The same model and game approach are also found in [11]. However, in [11], the switching cost is introduced as a penalty in the utility function of each user. The reason is that as the user switches among the different BSs, the resource cost for the channel estimation and communication occurs. The game approaches used in [10] and [11] are considered to be the classical game. A fractional game as a generalized game has been proposed in recent works [15]–[17]. By taking the power-law memory (PLM) into account the network selection strategies of the users, the fractional game is considered to be an generalized game approach that improves the utility of the users compared with the classical game.

Despite of its merit advantages, the evolutionary game may require the users to repeat the network selection many times for the convergence, causing the high latency to the users for selecting the best network. We further consider the case that the BSs cooperate with each other to share the network state. Then, we develop a multi-agent deep reinforcement learning (MADRL) for the decision-making of the users [18]. With MADRL, each user equipped with a DRL algorithm can make its best action decision, e.g., network selection, depending its network state observation to receive a long-term reward. The MADRL is pre-trained offline, and then the users can select its best network online given any network state. This is important in dynamic systems such as the sub-6GHz-mmWave-THz network where the channels may vary fast over time. Given the cooperation condition, the MADRL is expected to achieve higher utility than the game approaches. MADRL has been used to solve emerging issues in future networks such as dynamic spectrum access [19], dynamic user-BS association [20], and power control [21]. The contributions of our work are summarized as follows:

- We investigate a time-variant network access selection problem of URLLC users and eMBB users in a sub-6GHz-mmWave-THz network. The URLLC users aim to satisfy their latency requirement, while the eMBB users aim to maximize their data rate. We propose two different utility functions with network service cost for the URLLC users and eMBB users to satisfy their QoS.
- To model the dynamic network adaptation of the URLLC users and eMBB users over time, we leverage the fractional game with replicator dynamic processes and effects of power-law memory. The replicator dynamic process helps the users to gradually achieve the highest utility after switching among the BSs, given the strategies of other users. Meanwhile, the accounting of the effects of power-law memory to the users' decision-making helps them to better select the network, thus improving the utility.
- We consider the case that there is a cooperation among the BSs in sharing network states, e.g., the number of users selecting each BS in the past and the total network utility. We model the network selection of the URLLC and eMBB users as a multi-agent problem. We develop an MADRL algorithm for the network selection problem which allows each URLLC user and eMBB user to find its own optimal policy to maximize its long-term utility.
- Simulations with hundreds users are provided to demonstrate the effectiveness and scalability of the proposed game and MADRL approaches. The results further show that the proposed game is able to reach its equilibrium even with outdated information. Also, the fractional game has higher utility than the classical game, but it occurs higher switching cost. Meanwhile, given the cooperation among the BSs, the MADRL with an optimal policy outperforms both the game approaches.

The rest of the paper is organized as follows. In Section II, we present the HetNet model including the utility function design specific for the two types of users, i.e., URLLC users and eMBB Users. We present the fractional game in III and the MADRL algorithm in Section IV. Simulation results are provided and discussed in Section V, and the conclusions of the paper are given in Section VI.

TABLE I
LIST OF NOTATIONS FREQUENTLY USED IN THIS PAPER.

Notation	**Description**
a_i, s_i, r_i	Action, state, and reward of user i
$d_{b,i}$	Distance between BS b and user i
g_i	Total discounted reward over next time slots of user i
$h_{b,i}$	Small-fading channel between BS b and user i
L_b	Number of antennas of BS b
N_b^{u}	Numbers of URLLC users selecting BS b
N_b^{e}	Numbers of eMBB users selecting BS b
p_b^{tx}	Transmit power of BS b
$p_{m,i}^{\mathrm{LoS}}$, $p_{z,i}^{\mathrm{LoS}}$	LoS probability of BS m, BS z corresponding to user i
$Q_{\pi_i}(\dot{})$	Q-value function of user i
$r_{b,u}$	Transmission rate achieved by URLLC user u
$r_{b,e}$	Transmission rate achieved by eMBB user e
$\mathcal{S}, \mathcal{M}, \mathcal{T}$	Sets of sub-6GHz BSs, mmWave BSs, and THz BSs
u, e, b	Indexes of URLLC user, eMBB user, and BS
$v_{b,e}$	Valuations of eMBB user e as associated with BS b
$v_{b,u}$	Valuations of URLLC user u as associated with BS b
$W_{b,u}$	Expected bandwidth achieved by URLLC user u
$W_{b,e}$	Expected bandwidth achieved by eMBB user e
$x_{b,i}$	Probability that user i selecting BS b
γ	Discount factor
ε_u	Decoding error probability requirement of URRLC user u
ξ	Order of the fractional derivative
$\Pi_{b,i}$	Utility of user i as associated with BS b BS b
π_i	Policy of user i
$\Pi_{b,e}$	Utility of eMBB user b as associated with BS b
τ_u	Latency requirement requirement of URRLC user u

II. SYSTEM MODEL

We consider an HetNet consisting of a set $\mathcal{U}$ of U URLLC users and a set $\mathcal{E}$ of E eMBB users. The HetNet consists of three tiers. The first tier consists of a set $\mathcal{S}$ of S sub-6GHz BSs deployed to provide sub-6GHz communication services to the users. The second tier is composed of a set $\mathcal{M}$ of M mmWave BSs providing mmWave communication services. The third tier consists of a set $\mathcal{T}$ of T THz BSs which is responsible for providing THz communication services to the users. The sub-6GHz BSs are deployed to provide wide network coverage to the users, and thus we assume that each sub-6GHz BS is equipped with an omnidirectional antenna [11], [9]. In contrast, the mmWave BSs and THz BSs aim to provide high throughput to individual users. Due to the high attenuation of mmWave and THz signals, we consider that highly directional beamforming antenna arrays are used for the mmWave BSs and THz BSs [9]. We use indexes of s, m, and z to specify sub-6GHz BS $s \in \mathcal{S}$, mmWave BS $m \in \mathcal{M}$, THz BS $z \in \mathcal{T}$, respectively. We also use an index b to specify BS $b \in \{s, m, z\}$. We use indexes u and e to specify URLLC user $u \in \mathcal{U}$ and eMBB user $e \in \mathcal{E}$, respectively. In addition, we use an index i to specify user $i \in \{u, e\}$ and t to specify the index of time slot or time instant t.

A. Sectorized Antenna Model

To model radiation patterns of antenna arrays in all the BSs, we use the sectorized antenna model [9], [11], [22], [23] in which the antenna gain for user i at a certain location associated with BS b is expressed by [9]

$$G_{b,i} = \begin{cases} G_b^{\max}, & |\theta| \leq \frac{\theta_b}{2}, \\ G_b^{\min}, & \text{otherwise}, \end{cases} \tag{1}$$

where $G_b^{\max}$ and $G_b^{\min}$ are the maximum and minimum antenna gains corresponding to the direction of the main-lobe and that of the side-lobes of BS b, respectively, $\theta \in [-\pi, \pi]$ represents the boresight direction angle, and θ_b refers to the main-lobe beamwidth of BS b. The values of θ_b, $G_b^{\max}$, and $G_b^{\min}$ are determined based on the number of array antenna elements and the array geometry [9]. We consider that the mmWave BSs and THz BSs are equipped with uniform linear antenna arrays in which the space between antenna elements equals half of the operating wavelength. Then, θ_b, $G_b^{\max}$, and $G_b^{\min}$ are

$$\begin{cases} \theta_b = 2 \arcsin\left\{\frac{2.782}{\pi L_b}\right\} \\ G_b^{\max} = \frac{2\pi L_b^2 \sin\{1.5\pi/L_b\}}{\theta_b L_b^2 \sin\{1.5\pi/L_b\} + (2\pi - \theta_b)} \\ G_b^{\min} = \frac{2\pi}{\theta_b L_b^2 \sin\{1.5\pi/L_b\} + (2\pi - \theta_b)}, \end{cases} \tag{2}$$

where L_b is the number of antennas of BS b.

B. Blockage Model

The mmWave communications is very sensitive to obstacles existed between the BSs and the users due to their high propagation loss. In this paper, for the mmWave communications, we consider only the Light-of-Sight (LoS) signal, and that of the none-LoS (NLoS) signals is ignored as the NLoS signal is proved to be much weaker than the LoS signal, i.e., lower than 20 dB [11], [24]. The probability that a LoS link exists between a BS and a user depends on the distance between the BS and the user and the size of obstacles. Then, the LoS probability, denoted by $p_{m,i}^{\mathrm{LoS}}$, between mmWave BS and user i is [25]

$$\min\left(\frac{H_1}{d_{m,i}}, 1\right)\left(1 - \exp(-\frac{d_{m,i}}{H_2})\right) + \exp\left(-\frac{d_{m,i}}{H_2}\right), \tag{3}$$

where $d_{m,i}$ is the distance between BS m and user i, $H_1 = 18$ m, and $H_2 = 63$ m.

C. Received Power at Users

The users selecting different BSs will receive different powers as presented in the following.

1) Sub-6GHz propagation model: We denote p_s^{tx} as the transmit power of sub-6GHz BS $s \in \mathcal{S}$. When user i selects sub-6GHz BS $s \in \mathcal{S}$, the receive power at user i is

$$p_{s,i} = p_s^{\mathrm{tx}} G_s G_i h_{s,i} \left(\frac{c}{4\pi f_s}\right)^2 d_{s,i}^{-\alpha_s}, \tag{4}$$

where $h_{s,i}$ is the small-scale fading, G_s is the antenna gain of the BS, G_i is the antenna gain of user i, f_s is the operating frequency of the BS, α_s is the path loss exponent corresponding to the sub-6GHz band, and c is the speed of light.

2) MmWave propagation model: We denote p_m^{tx} as the transmission power of mmWave BS $m \in \mathcal{M}$. When user i selects the BS, the receive power at user i is

$$p_{m,i} = p_{m,i}^{\text{LoS}} p_m^{\text{tx}} G_{m,i} G_i h_{m,i} \left(\frac{c}{4\pi f_m} \right)^2 d_{m,i}^{-\alpha_m}, \qquad (5)$$

where $p_{m,i}^{\text{LoS}}$ is the probability that a LoS link exists between BS m and user i determined according to (3), $h_{m,i}$ is the small-scale fading, $G_{m,i}$ is the antenna gain of the BS, α_m is the path loss exponent corresponding to the mmWave band, and f_m is the operation frequency of the BS.

3) THz propagation model: We denote p_t^{tx} and $G_{z,i}$ as the transmission power and the antenna gain of THz BS $z \in \mathcal{T}$, respectively. Similar to the mmWave communications, we consider the effect of LoS and ignore that of the NLoS on the data rate achieved by the users. The power received by user i by selecting BS z is [26]

$$p_{z,i} = p_z^{\text{tx}} G_{z,i} G_i \left(\frac{c}{4\pi f_z} \right)^2 d_{z,i}^{-\alpha_z} e^{-k_\alpha(f_z) d_{z,i}}, \qquad (6)$$

where $d_{z,i}$ is the distance between the BS and the user, f_z denotes the operating frequency of BS z, c is the speed of light, $k_a(f_z)$ refers to the medium absorption coefficient that depends on carrier frequency and the composition of the transmission medium at a molecular level. For example, given $f_z = 2$ THz, $k_a(f_z) = 2.3 \times 10^{-5}$ m^{-1} for oxygen [27].

D. Transmission Rate Achieved by URLLC and eMBB users

We denote $r_{b,u}$ as the transmission rate achieved by URLLC user $u \in \mathcal{U}$ and $r_{b,e}$ as the transmission rate achieved by eMBB user $e \in \mathcal{E}$ by selecting BS $b \in \{s, m, z\}$. We consider that at the beginning of a time slot, a set $\mathcal{N}_b^{\text{u}}$ of N_b^{u} URLLC users and a set $\mathcal{N}_b^{\text{e}}$ of N_b^{e} eMBB users select b. We denote W_b as the total bandwidth of BS b. Since multiple users can exist in the coverage area of a BS, a round-robin (RR) scheduling mechanism or channel-aware proportional fair (PF) algorithm is deployed to schedule the spectrum resource. It is also assumed that the users are capable of operating at sub-6GHz, mmWave, and THz bands, but each user is solely served by a single BS at each time slot. By selecting BS b, each user $i \in \{u, e\}$ is expected to achieve the amount of bandwidth of $W_{b,i} = \frac{W_b}{N_b^{\text{u}} + N_b^{\text{e}}}$. Note that the URLLC users have stringent QoS requirements on transmission latency and decoding error probability. For this, packets from the URLLC users have very small sizes. Thus, we leverage the capacity analysis in finite blocklength regime [28] rather than the Shannon capacity. Particularly, the transmission rate achieved by URLLC user u as selecting BS b is approximated by

$$r_{b,u} \approx \frac{W_{b,u}}{\ln 2} \left[\ln \left(1 + \frac{p_{b,u}}{N_0 W_b} \right) - \sqrt{\frac{V_{b,u}}{\tau_u W_{b,u}}} Q_{\text{URLLC}}^{-1}(\varepsilon_u) \right], \qquad (7)$$

where $W_{b,u} = \frac{W_b}{N_b^{\text{u}} + N_b^{\text{e}}}$ is the bandwidth achieved by URLLC user u by selecting BS b, N_0 is the noise power spectral density, $p_{b,u}$ is the receive power at URLLC user u determined according to (6), τ_u and ε_u are the latency requirement and decoding error probability requirement of URRLC user u, respectively, $Q_{\text{URLLC}}^{-1}(\cdot)$ is the inverse of Q-function, and $V_{b,u} = 1 - \left(1 + \frac{p_{b,u}}{N_0 W_b} \right)^{-2}$ is the channel dispersion. As SNR at the URLLC user is higher than 5 dB, and we can have $V_{b,u} \approx 1$. Additionally, the transmission rate achieved by eMBB user e by selecting BS b is determined by

$$r_{b,e} = W_{b,e} \log_2 \left(1 + \frac{p_{b,e}}{N_0 W_b} \right), \qquad (8)$$

where $W_{b,e} = \frac{W_b}{N_b^{\text{u}} + N_b^{\text{e}}}$ is the bandwidth achieved by eMBB user $e \in \mathcal{N}_b^{\text{e}}$ by selecting BS b, $p_{b,e}$ is the receive power at the user that is determined according to (6).

E. Utility Achieved by URLLC and eMBB Users

1) Expected data rate achieved by URLLC users and eMBB users: In the context of evolutionary game, a set of users selecting the same BS constitutes a population or a group. We denote $x_{b,u}$ as the probability that user u in the URLLC user set selects BS $b \in \{s, m, z\}$. Also, we denote $x_{b,e}$ as the probability that user e in the eMBB user set selects BS b. There are N_b^{u} URLLC users and N_b^{e} eMBB users selecting BS b. Therefore, user u selects BS b with a probability of $x_{b,u} = N_b^{\text{u}}/U$, and user e selects BS b with a probability of $x_{b,e} = N_b^{\text{e}}/E$. As such, the expected number of URLLC users selecting BS b is $x_{b,u}U$, and the expected number of eMBB users selecting BS b is $x_{b,e}E$. Note that the bandwidth of the BS is equally shared by both the URLLC and eMBB users. Therefore, the expected bandwidth achieved by user u and user e by selecting BS b is $\overline{W}_{b,u} = \overline{W}_{b,e} = \frac{W_b}{x_{b,u}U + x_{b,e}E}$. Consequently, the expected data rate achieved by URLLC user u is

$$\overline{r}_{b,u} \approx \frac{\overline{W}_{b,u}}{\ln 2} \left[\ln \left(1 + \frac{p_{b,u}}{N_0 W_b} \right) - \sqrt{\frac{\overline{V}_{b,u}}{\tau_u \overline{W}_{b,u}}} Q^{-1}(\varepsilon_u) \right],$$

where $\overline{V}_{b,u} = 1 - \left(1 + \frac{p_{b,u}}{N_0 W_b} \right)^{-2}$. The expected data rate achieved by eMBB user e by selecting BS b is

$$\overline{r}_{b,e} = \overline{W}_{b,e} \log \left(1 + \frac{p_{b,e}}{N_0 W_b} \right). \qquad (9)$$

2) Utility of URLLC users: We denote $v_{b,u}$ as the valuation of one data bit downloaded to user $u \in \mathcal{U}$ by selecting BS b. Here, the valuation indicates how much the user is willing to buy one bit by selecting the network service provided by the service provider. Since the user uses the network service from the BS, it will pay a service fee to the service provider who owns the BS. We denote $\lambda_{b,u}$ as the price per data unit by using BS b. For the URLLC users, the reliability and latency are the critical metrics. A low spectral efficiency modulation and coding scheme will be adopted so as to ensure the ultra-reliability of the URLLC traffic according to the 3GPP standard [29]. For the latency, the whole URLLC packet needs to be transmitted within a latency requirement. In particular, we denote L_u as the length of the packet of URLLC user u. Also, we denote τ_u as the latency requirement of URLLC user u. To satisfy the low latency requirement of the URLLC packet, i.e., transmitting the URLLC packet L_u within τ_u, the minimum data rate achieved by user u should be no less than $r_u^{\rm req} = \frac{L_u}{\tau_u}$, i.e., $\overline{r}_{b,u} \geq r_u^{\rm req}$. Then, we propose a utility function achieved by URLLC user u that is proportional to its data rate as follows:

$$\begin{aligned}\Pi_{b,u} =& v_{b,u}\left(\overline{r}_{b,u} - r_u^{\rm req}\right) - C_b\overline{r}_{b,u} \\ =& (v_{b,u} - C_b)\overline{r}_{b,u} - v_{b,u}r_u^{\rm req},\end{aligned} \tag{10}$$

where $v_{b,u}$ is the valuation of one data bit downloaded to URRLC user u as selecting BS b. In general, $v_{b,u}$ of the user is the same over the BSs. C_b is the cost per data bit, e.g., \$$10^{-5}$/bit, set by BS b. As seen, the utility dramatically decreases once the latency requirement of the URLLC user is not satisfied.

3) Utility of eMBB users: We denote $\Pi_{b,e}$ as the utility achieved by user $e \in \mathcal{E}$ by selecting BS b. Different from URLLC users, the eMBB users select a BS so as to maximize its throughput, and thus we define the utility obtained by eMBB user e by selecting BS b as follows:

$$\Pi_{b,e} = v_{b,e}\overline{r}_{b,e} - C_b\overline{r}_{b,e} = \left(v_{b,e} - C_b\right)\overline{r}_{b,e}, \tag{11}$$

where $v_{b,e}$ is the valuation of one data bit downloaded to eMBB user e by selecting BS b.

III. Fractional Game Approach

We first model the BS selection strategies of the URLLC users and eMBB users and then analyze the evolutionary equilibrium.

A. Game Formulation

Each user makes its BS selection and receives the corresponding utility. The users selecting the same BS constitute a group. As such, an URLLC user and an eMBB user belong to a group if they select the same BS. The expected utility of user $i \in \mathcal{E} \cup \mathcal{U}$ over all the BSs at time t is

$$\overline{\Pi}_i = \sum_{b\in\mathcal{S}\cup\mathcal{M}\cup\mathcal{T}} x_{b,i}\Pi_{b,i}. \tag{12}$$

To model the BS adaptation of the users, the classical game [30] leverages the replicator dynamics as follows

$$\frac{\rm dx}{\rm dt} = \mu x_{b,i}[t]\left(\Pi_{b,i}[t] - \overline{\Pi}_i[t]\right), \tag{13}$$

where $[t]$ refers to the current time instant, $\frac{\rm dx}{\rm dt}$ is the first-order derivative of $x_{b,i}$ with respect to time, $x_{b,i}[t_0] = x_{b,i}^0$ is the initial strategy of the user in group, μ is the learning rate of the users. The evolutionary game defined in (13) represents the population strategy evolution of the users over time, and the game converges to the evolutionary equilibrium, which is defined as the set of stable fixed points of the replicator dynamics. It can be seen from (13) that the strategy selection changes between two consecutive time instances of the user depends only on the instantaneous utility, and does not depend on its past strategy. This is namely an economic process without memory effect. In the real scenarios including our system model, the users are able to remember their previous strategies that have an impact on the current selection decision, thus improving the utility achieved by the users. This is the power-law memory (PLM) concept in economic processes [16] and also the key idea of the fractional game. Thus, to understand the fractional game, we present the economic processes with the PLM. Accordingly, a typical economic process is composed of an exogenous variable as the input and endogenous variable as the output. We denote $\Psi[t]$ as the endogenous variable of the process at the current time instant. In particular for our work, $\Psi[t]$ is the BS selection strategy of the user. Also, we denote $\Upsilon[\tau], \tau \in (-\infty, t)$ as the exogenous variable at the past instant τ. Then, $\Psi[t]$ depends on the past changes of the exogenous variable as $\Psi[t] = \Xi_0^t\left(\Upsilon\left[\tau\right]\right) + \Psi\left[0\right]$ [16], [31], where $\Psi\left[0\right]$ is the value of $\Psi[t]|_{t=0}$ and $\Xi_0^t\left(\Upsilon\left[\tau\right]\right)$ is

$$\Xi_0^t\left(\Upsilon\left[\tau\right]\right) := \int_0^t \Omega_\xi\left(t-\tau\right)\Upsilon\left[\tau\right]{\rm d}\tau. \tag{14}$$

In (14), $\Omega_\xi\left(t-\tau\right)$ captures the memory dynamics and $\xi \in (0, 2)$ represents the order of the fractional derivative. Function $\Omega_\xi\left(t-\tau\right)$ is defined by [31]:

$$\Omega_\xi(t-\tau) = \frac{1}{\Gamma(\xi)(t-\tau)^{1-\xi}} \tag{15}$$

where $\Gamma\left(\xi\right)$ is the is the Gamma function, and its form is $\Gamma\left(\xi\right) = \int_0^{+\infty} z^{\xi-1}e^{-z}{\rm d}z$. The physical meaning of $\Omega_\xi\left(t-\tau\right)$ is that it represent the impact of $\Upsilon[\tau]$ on $\Psi[t]$ in the economic process. To further understand this impact, we consider a special case in which $\Psi[0] = 0$, and we have $\Psi[t] = \Xi_0^t\left(\Upsilon\left[\tau\right]\right)$. Then, we have the Leibniz integral rule as follows:

$$\frac{\rm d\Psi}{\rm dt} = \Omega_\xi[0]\Upsilon[t] + \int_0^t \left[\frac{\rm d}{\rm dt}\Omega_\xi\left(t-\tau\right)\right]\Upsilon[\tau]{\rm d}\tau. \tag{16}$$

As observed from (16), $\Psi[t]$ depends on both the current input $\Upsilon[t]$ and past input $\Upsilon[\tau]$ for all $\tau \in [0, t)$. We can rewrite (16) as

$${}_0^{\rm C}D_t^\xi\Psi[t] = \Upsilon[t], \tag{17}$$

where ${}_0^{\rm C}D_t^\xi\Psi[t]$ is the right-sided Caputo-type fractional derivative of $\Psi[t]$ with ξth-order, which is given by [31]

$${}_0^{\rm C}D_t^\xi\Psi[t] = \frac{1}{\Gamma(\lceil\xi\rceil - \xi)}\int_0^t \frac{\Psi^{(\lceil\xi\rceil)}[\tau]}{(t-\tau)^{\xi+1-\lceil\xi\rceil}}{\rm d}\tau. \tag{18}$$

To incorporate past experience of the users, we propose to leverage a fractional game formulated by [31]:

$${}_0^{\rm C}D_t^\xi x_{b,i}[t] = \mu x_{b,i}[t]\left(\Pi_{b,i} - \overline{\Pi}_i\right), \forall i \in \{u, e\}, b \in \{s, m, z\}. \tag{19}$$

The fractional game defined in (19) allows the user to incorporate its network selection strategies at the past time instants (along with the returned average utility) into its current decision-making on BS selection. Note that as $\xi = 1$, then the game defined in (19) becomes the classical game as expressed in (13). In the next section, we will show that the fractional game expressed in (19) has a unique equilibrium, which guarantees the system stability of the system.

B. Game Equilibrium

With the proposed game, the users adapt their network selection strategies over time. The game equilibrium is achieved as the users do not change their strategies. In this section, we prove the existence of an equilibrium to the game defined in (19). The proof is analyzed similar to some recent works [15], [16], [17] as follows.

First, we denote $\mathcal{N}_b = \mathcal{N}_b^{\mathrm{u}} \cap \mathcal{N}_b^{\mathrm{e}}$ as the set of all the users, i.e., URLLC users and eMBB users, selecting BS $b, b \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{T}$. We further define $\boldsymbol{\Upsilon}[t] \triangleq [x_{b,i}[t]]_{i\in\mathcal{N}_b, b\in\mathcal{S}\cup\mathcal{M}\cup\mathcal{T}}$ and $\mathbf{E}\left(\boldsymbol{\Upsilon}[t]\right) \triangleq \left[\mu x_{b,i}[t]\left[u_{b,i}[t] - \overline{u}_i[t]\right]\right]_{i\in\mathcal{N}_b, b\in\mathcal{S}\cup\mathcal{M}\cup\mathcal{T}}$, then (19) is reformulated by

$$ {}_{0}^{\mathrm{C}}D_t^{\xi}\boldsymbol{\Upsilon}[t] = \mathbf{E}\left(\boldsymbol{\Upsilon}[t]\right), \tag{20}$$

where $\boldsymbol{\Upsilon}[0] = [x_{b,i}[0]]_{i\in\mathcal{N}_b, b\in\mathcal{S}\cup\mathcal{M}\cup\mathcal{T}}$ is the initial strategy. Let $e_{b,i}$ be an element of vector $\mathbf{E}$. We assume that $e_{b,i}$ is a twice-differentiable function, and its derivation $\frac{\partial}{\partial x_{b,i}} e_{b,i}$ is bounded, i.e., $\exists B \in \mathbb{R}^+ : |e_{b,i}(\hat{\boldsymbol{\Upsilon}}[t]) - e_{b,i}(\tilde{\boldsymbol{\Upsilon}}[t])| < B||(\hat{\boldsymbol{\Upsilon}}[t]) - \tilde{\boldsymbol{\Upsilon}}[t]||_{\mathcal{L}_1}$, satisfying the Lipschitz condition. Then, the game in (20) can be converted the equivalent problem as follows:

$$\boldsymbol{\Upsilon}[t] = \boldsymbol{\Upsilon}[0] + {}_0\mathrm{I}_{\mathrm{t}}^{\xi}\mathbf{E}(\boldsymbol{\Upsilon}[\mathrm{t}]), \forall \mathrm{t} \in \mathcal{T}, \tag{21}$$

where

$${}_0\mathrm{I}_{\mathrm{t}}^{\xi}\mathbf{E}(\boldsymbol{\Upsilon}[\mathrm{t}]) = \int_0^{\mathrm{t}} \frac{(\mathrm{t}-\tau)^{\xi-1}}{\Gamma(\xi)}\mathbf{E}(\boldsymbol{\Upsilon}[\tau])\mathrm{d}\tau \tag{22}$$

is the fractional integral [32].

The equivalence between (20) and (21) is proved as follows. First, we take the $\lceil\xi\rceil$-th order derivative of $\boldsymbol{\Upsilon}[t]$ in (21) t as follows:

$$\begin{aligned}\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[t] =& \frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\left[{}_0\mathrm{I}_{\mathrm{t}}^{\xi}\mathbf{E}\left(\boldsymbol{\Upsilon}[\mathrm{t}]\right)\right] = {}_{0}^{\mathrm{RL}}D_t^{\lceil\xi\rceil-\xi}\mathbf{E}\left(\boldsymbol{\Upsilon}[t]\right) \\ =& F^{\xi}[t]\mathbf{E}\left(\boldsymbol{\Upsilon}[0]\right) + {}_{0}^{\mathrm{C}}D_t^{\xi}\mathbf{E}\left(\boldsymbol{\Upsilon}[t]\right),\end{aligned} \tag{23}$$

where $F^{\xi}[t] = \frac{t^{\xi-\lceil\xi\rceil}}{\Gamma(1-\lceil\xi\rceil+\xi)}$, and ${}_{0}^{\mathrm{RL}}D_t^{\xi}\mathbf{E}\left(\boldsymbol{\Upsilon}[t]\right)$ is the ξth-order left-sided Riemann-Liouville fractional derivative given by

$$\begin{aligned}&{}_{0}^{\mathrm{RL}}D_t^{\lceil\xi\rceil-\xi}\mathbf{E}\left(\boldsymbol{\Upsilon}[t]\right) \\ =&\frac{1}{\Gamma(1-\lceil\xi\rceil+\xi)}\frac{\mathrm{d}}{\mathrm{dt}}\int_0^t \frac{\mathbf{E}(\boldsymbol{\Upsilon}[\tau])}{(t-\tau)^{\lceil\xi\rceil-\xi}}\mathrm{d}\tau \\ =&\frac{1}{\Gamma(1-\lceil\xi\rceil+\xi)}\frac{\mathrm{d}}{\mathrm{dt}}\int_0^t \theta^{\xi-\lceil\xi\rceil}\mathbf{E}(\boldsymbol{\Upsilon}[t-\theta])\mathrm{d}\theta \\ =&\frac{1}{\Gamma(1-\lceil\xi\rceil+\xi)}\Big[t^{\xi-\lceil\xi\rceil}\mathbf{E}(\boldsymbol{\Upsilon}^0) + \int_0^t \theta^{\xi-\lceil\xi\rceil} \\ &\times \frac{\mathrm{d}}{\mathrm{dt}}\mathbf{E}(\boldsymbol{\Upsilon}[t-\theta])\mathrm{d}\theta\Big] \\ =&\frac{1}{\Gamma(1-\lceil\xi\rceil+\xi)}\Big[t^{\xi-\lceil\xi\rceil}\mathbf{E}(\boldsymbol{\Upsilon}^0) + \int_0^t (t-\tau)^{\xi-\lceil\xi\rceil} \\ &\times \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{E}(\boldsymbol{\Upsilon}[\tau])\mathrm{d}\tau\Big] \\ =&F^{\xi}[t]\mathbf{E}(\boldsymbol{\Upsilon}^0) + {}_0\mathrm{I}_{\mathrm{t}}^{\xi}\Big[\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\mathbf{E}(\boldsymbol{\Upsilon}[\mathrm{t}])\Big].\end{aligned} \tag{24}$$

Given any $\sigma \in (0, t)$, the upper bound of the $\mathcal{L}^1$ norm of $F^{\xi}[t]\mathbf{E}\left(\boldsymbol{\Upsilon}[0]\right)$ is [16]

$$\left\|F^{\xi}[t]\mathbf{E}(\boldsymbol{\Upsilon}[0])\right\|_{\mathcal{L}^1} \leq \left\|F^{\xi}(\sigma)\mathbf{E}(\boldsymbol{\Upsilon}[0])\right\|_{\mathcal{L}^1}. \tag{25}$$

Based on the bounding condition of $\frac{\partial}{\partial x_{b,i}}e_{b,i}$, Equations (23) and (24), and Inequality (25), we have

$$\begin{aligned}\left\|\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[t]\right\|_{\mathcal{T}} \leq& \left\|F^{\xi}(\sigma)\mathbf{E}\left(\boldsymbol{\Upsilon}[0]\right)\right\|_{\mathcal{L}^1} + \left\|{}_0\mathrm{I}_{\mathrm{t}}^{\xi}\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\mathbf{E}\left(\boldsymbol{\Upsilon}[\mathrm{t}]\right)\right\|_{\mathcal{T}} \\ \leq& \left\|F^{\xi}(\sigma)\mathbf{E}\left(\boldsymbol{\Upsilon}[0]\right)\right\|_{\mathcal{L}^1} + \left\|{}_0\mathrm{I}_{\mathrm{t}}^{\xi}\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[\mathrm{t}]\right\|_{\mathcal{T}}\mathrm{AB},\end{aligned} \tag{26}$$

where $\|x\|_{\mathcal{T}} = \int_{\mathcal{T}} \exp\left(-\mu t\right)\|x\|_{\mathcal{L}^1}\,\mathrm{dt}$ and A is the cardinality of $\{\,(i,k)|\, i \in \mathcal{N}_b, b \in \mathcal{S} \cap \mathcal{M} \cap \mathcal{T}\}$. By substituting (B.2) in Appendix in [17] into (26), we have

$$\begin{aligned}\left\|\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[t]\right\|_{\mathcal{T}} \leq& \left\|F^{\xi}(\sigma)\mathbf{E}\left(\boldsymbol{\Upsilon}[0]\right)\right\|_{\mathcal{L}^1} + \frac{AB}{\mu^{\xi}}\left\|\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{ds}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[t]\right\|_{\mathcal{T}} \\ \Leftrightarrow \left\|\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[t]\right\|_{\mathcal{T}} \leq& \frac{1}{1-\frac{AB}{\mu^{\xi}}}\left\|F^{\xi}(\sigma)\mathbf{E}(\boldsymbol{\Upsilon}[0])\right\|_{\mathcal{L}^1}.\end{aligned} \tag{27}$$

In the case that $1 > \frac{AB}{\mu^{\xi}}$, i.e., μ is large, $\left\|\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[t]\right\|_{\mathcal{T}}$ has an upper bound. Then, consider $\xi \in (0, 1)$ and take the $\lceil\xi\rceil$th-order derivative of $\boldsymbol{\Upsilon}[t]$, we have

$$\begin{aligned}{}_{0}^{\mathrm{C}}D_t^{\xi}\boldsymbol{\Upsilon}[t] =& {}_0\mathrm{I}_{\mathrm{t}}^{\lceil\xi\rceil-\xi}\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\boldsymbol{\Upsilon}[\mathrm{t}] \\ =& {}_0\mathrm{I}_{\mathrm{t}}^{\lceil\xi\rceil-\xi}\Big\{\frac{\mathrm{t}^{\xi-\lceil\xi\rceil}}{\Gamma\left(1-\lceil\xi\rceil+\xi\right)}\mathbf{E}\left(\boldsymbol{\Upsilon}[0]\right) \\ &+ {}_0\mathrm{I}_{\mathrm{t}}^{\xi}\Big[\frac{\mathrm{d}^{\lceil\xi\rceil}}{\mathrm{dt}^{\lceil\xi\rceil}}\mathbf{E}\left(\boldsymbol{\Upsilon}[\mathrm{t}]\right)\Big]\Big\} = \mathbf{E}(\boldsymbol{\Upsilon}[\mathrm{t}]).\end{aligned} \tag{28}$$

(28) demonstrates that the fractional game defined in (20) is equivalent to the problem expressed in (21). Thus, by proving the uniqueness of the solution to the problem in (21), we can guarantee that the solution is unique to the fractional game in (20). We denote $\mathcal{D}_{\boldsymbol{\Upsilon}}$ as the feasible domain of $\boldsymbol{\Upsilon}$. Also, we denote F as an operator defined by $F : \mathcal{D}_{\boldsymbol{\Upsilon}} \mapsto \mathcal{D}_{\boldsymbol{\Upsilon}}$. Based on Appendix B.3 in [17], we have

$$\left\|F\hat{\boldsymbol{\Upsilon}}[t] - F\tilde{\boldsymbol{\Upsilon}}[t]\right\|_{\mathcal{T}} < \frac{AB}{\mu^{\xi}}\left\|\hat{\boldsymbol{\Upsilon}}[t] - \tilde{\boldsymbol{\Upsilon}}[t]\right\|_{\mathcal{T}}. \tag{29}$$

Note that in the case of $\mu^{\xi} \geq AB$, we have $\|F\hat{\boldsymbol{\Upsilon}}[t] - F\tilde{\boldsymbol{\Upsilon}}[t]\|_{\mathcal{T}} < \|\hat{\boldsymbol{\Upsilon}}[t] - \tilde{\boldsymbol{\Upsilon}}[t]\|_{\mathcal{T}}$. This means that F satisfies the fixed point theorem and that there exists a unique solution to the problem given in (21). Based on (29), we can prove the stability of the solution to the problem given in (21). This is well presented in the existing studies, e.g., [16], [17]. Thus, we omit the proof of the stability. Instead, the stability of the game is illustrated and discussed in Section V.

IV. Multi-Agent Deep Reinforcement Learning

The fractional game approach can be considered to be the fully decentralized solution in which there is no cooperation among the BSs. In this section, we consider that the BSs cooperate with each other, and we present an MADRL algorithm for the decision making of the URLLC and eMBB users. In general, the MADRL algorithm is related to the fractional game since both of them enables each user to autonomously make local decision based on received utility. Both approaches are thus suitable to large-scale systems. Otherwise, the reasons for using the MADRL, rather than a centralized optimization

Algorithm 1 Fractional game for the network selection of the URLLC and eMBB users

```
Initialization: $T_{\max}$, $\mu$, $S$, $M$, $T$, $W_s, W_m, W_z$, $v_{s,e}, v_{m,e}, v_{t,e}, v_{s,u}, v_{m,u}, v_{z,u}, C_s, C_m, C_z, \alpha_s, \alpha_m, \alpha_z$, $x_{b,i} \forall b \in \{s, m, z\}, i \in \{u, e\}$. Each user randomly chooses an BS and its service;
repeat
    for $i \in \{e, u\}$ do
        BS $b$ calculates the utility (i.e., $\Pi_{b,i}$) based on (11) for eMBB users and (10) for URLLC users;
        BS $b$ sends the utility information and probabilities of $x_{b,i}$ to the user;
        The user calculates the expected utility, i.e. , $\bar{\Pi}_i$, according to (12);
        if $\bar{\Pi}_i[t] < \Pi_{b,i}[t]$ then
            Randomly choose a BS that has $\Pi_{b,i}[t]$ higher than $\bar{\Pi}_i[t]$;
        end if
    end for
    $t \leftarrow t + 1$;
until $t > T_{\max}$;
```

algorithm or a single-agent DRL, are as follows. First, conventionally each user makes its BS selection decision to achieve its own maximum utility. Second, the channel state at different time slot may be time-varying. Third, before making the BS decision, the user has no prior knowledge of the selection of other users and of the utility that it receives. Fourth, the BS selection decision of the user has an impact on the data rate and utility of other users. Finally, we investigate a large-scale system with some hundreds of URLLC users and eMBB users, and traditional learning algorithms like single-agent DRL faces large action and state space issues. However, to effectively use the MADRL, we make some assumptions as follows. First, there are high data rate links among the BSs. This assumption is reasonable as the optical wired backhaul links can be used. Second, each BS cooperates with each others to inform the total number of URLLC and eMBB users associated with it in the previous time slot as well as a system reward. Given the optical wired backhaul links, these information can shared shared among all BSs within in real time. Third, each user is equipped with a memory to store its dataset. This assumption is reasonable since the existing devices are mostly equipped with memories with large capacities. The key ideas of the proposed MADRL algorithm are as follows. First, each user, i.e., URLLC user or eMBB user, as an agent takes an action by selecting BS $b, b \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{T}$. Similar to the game approach as presented in Section III, BS b applies the RR scheduling mechanism to equally divide its bandwidth to its associated users. The BS then calculates the utility according to (10) if the user is the URLLC user and according to (11) if the user is the eMBB user. Different from the game approach that considers the average utility for each user over the users selecting one BS, the proposed MADRL algorithm aims to maximize the average utility for each user over all the users in the network. This is to balance the traffic load among the BSs, which benefits from the HetNet. This requires the BSs to communicate the utility of its users with each others, which may incur overhead communication costs. However, the links among the BSs are assumed to be have very high capacity, and the overhead communication latency is ignored. Then, the BSs calculate the average utility of each user and broadcasts this information to its associated users. This is considered to be the immediate reward achieved by the user. In the followings, we present the details of the proposed MADRL algorithm. For ease of presentation, in this section, we use the index of i to commonly refer to both the URLLC user or eMBB user, where $i \in \{1, \ldots, U + E\}$. Particularly, i represents the URLLC user if $i \in \{1, \ldots, U\}$ and the eMBB user if $i \in \{U + 1, \ldots, U + E\}$.

A. Action space

The mathematical formulation of the action space for each user agent i is as follows:

$$a_i[t] = \{b | b \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{T}\} \tag{30}$$

with $\mathcal{S}$, $\mathcal{M}$ and $\mathcal{T}$ are respectively sets of S sub-6GHz BSs, M mmWave BSs, and T THz BSs.

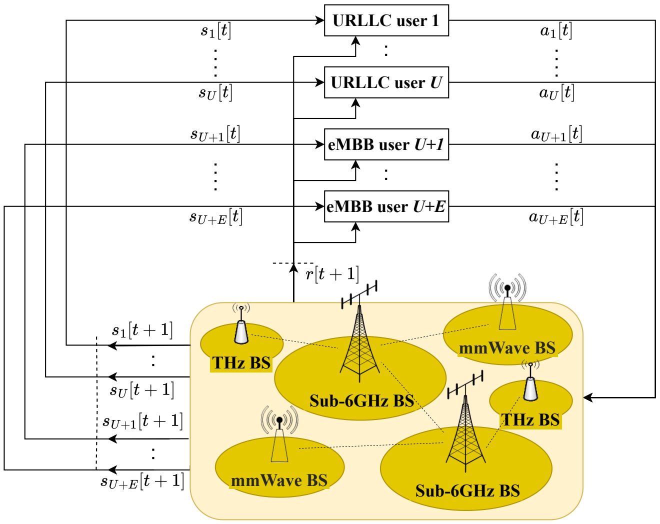


Fig. 1. Flowchart of the proposed MADRL for network selection of URLLC and eMBB users in the sub-6GHz-mmWave-THz network.

B. State space

Different from the game approach in which all the involved channels are fixed, we consider the dynamic and practical wireless environment in which the small-scale fading channel coefficients, i.e., $h_{s,i}, h_{m,i}, i \in \{1, \ldots, U + E\}, s \in \mathcal{S}, m \in \mathcal{M}$, vary over time slots. These channels directly impact the data rate and utility achieved by the users, and thus we include them as state elements observed by the users. To observe these values, the channels need to be estimated at the beginning of each time slot. For this, the users send pilot sequence signals to the BSs [19]. The BSs estimate the channels and send back to the users. Similar to the game approach, to help the users to have more information to make their decisions properly, we introduce the history experiences of the users as the state elements. Here, the history experiences consist of the action, reward, and the numbers of users selecting the different BSs in the previous time slot. The BSs can broadcast these information at the end of the previous time slot. Mathematically, the state space of user $i, i \in \{1, \dots, U+E\}$ is defined as

$$\begin{aligned} s_i[t] =& \Big\{ h_{s,i}[t], h_{m,i}[t], a_i[t-1], r_i[t], \\ & \Big\{ N_s^u[t-1] + N_s^e[t-1], N_m^u[t-1] + N_m^e[t-1], \\ & N_z^u[t-1] + N_z^e[t-1] \Big\}_{\forall s \in \mathcal{S}, \forall m \in \mathcal{M}, \forall z \in \mathcal{T}} \Big\} \end{aligned} \tag{31}$$

where $a_i[t-1]$ is the action of user i in the previous time step $t-1$ and $r_i[t]$ is the individual reward it receive for that action afterward, $N_b^u[t-1]$ and $N_b^e[t-1]$ are the numbers of URLLC users and eMBB users, respectively, selecting BS $b, b \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{T}$, in the previous time step $t-1$. It is highlighted that the elements of state s_i used as the input of the neural networks of the user may be very different in the value. To eliminate the errors caused by biased results and to improve the speed and stability of the convergence, the state elements of s_i are normalized to the range of $[0, 1]$ by their maximum values. Particularly, the individual reward is normalized by $r_i[t]/r_i^{\max}$, where $r_i^{\max}$ is the maximum value of r_i which is determined through observing the numerical simulations, meanwhile the numbers of URLLC users and eMBB users which selecting BS $b, b \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{T}$ in the previous time step $t-1$ are normalized by $N_b^u[t-1]/(U+E)$ and $N_b^e[t-1]/(U+E)$, respectively, where U and E are the maximum values of N_b^u and N_b^e, respectively.

C. Reward Design

We denote $r_i, i \in \{1, \dots, U+E\}$ as the individual reward obtained by user i. The user selects a BS to maximize its own utility $\Pi_{b,i}$. Thus, r_i is proportional to the utility, i.e., $\Pi_{b,i}$. Moreover, to motivate the user to move to the BS that leads to a utility higher than the current utility, we define a parameter of β close to the maximum value of $\Pi_{b,i}$. The maximum value of $\Pi_{b,i}$ can be obtained by observing some first iterations. Then, the user achieves a reward of $r_0 \geq \max \Pi_{b,i}$ if $\Pi_{b,i} \geq \beta$. Mathematically, the reward of user i is defined by

$$r_i[t+1] = \begin{cases} r_0, & \text{if } \Pi_{b,i} \geq \beta, \\ \Pi_{b,i}, & \text{otherwise.} \end{cases} \tag{32}$$

In (32), we need to use a certain value, i.e., r_0, as a target value so as to provide the users an optimal value for making its decision. This is not only enhances the probability of convergence, but also improves the training speed considerably. Since the optical wired backhaul links are deployed among the BSs. The BSs can cooperatively calculate a sum reward and share among them in real time as follows

$$r[t+1] = \sum_{i=1}^{U+E} r_i[t+1]. \tag{33}$$

Then, the BSs broadcast this reward to the users as their states.

D. Learning Algorithm

The MADRL approach consists of two stages: the centralized training stage and the decentralized implementation stage.

1) Centralized training stage: The user, i.e., URLLC user or eMBB user, learns to select its BS with the aim of maximizing the long-term reward. To helps the user to find its optimal policy, we leverage the reinforcement learning algorithm based on the double deep Q-learning algorithm [33]. The whole MADRL algorithm is shown in Algorithm 2. We denote π_i as the policy of user i. The learning algorithm uses two neural networks, i.e., an online network with weights of w_i and a target network with weights of w_i'. The online network estimates Q-value functions of all feasible actions at the output, and the target network is used to estimate the target value. We denote $Q_{\pi_i}(s_i, a_i)$ as the Q-value function when user i selects action a_i given state s_i. Then, we have

$$Q_{\pi_i}(s_i, a_i) = \mathrm{E}_{\pi_i}\big[g_i[t] | s_i[t], a_i[t]\big], \tag{34}$$

where $g_i[t]$ represents the total discounted rewards accumulated over all the next time slots from the current time slot defined by

$$g_i[t] = \sum_{k=0}^{\infty} \gamma^k r[t+k+1], 0 \leq \gamma \leq 1. \tag{35}$$

where γ is the discount factor. We denote $Q(s_i, a_i; w_i)$ as the parameterized Q-value function. Then, the target value at each time step t, denoted as $z_i[t]$, is determined by

$$\begin{aligned} z_i[t] =& r[t+1] + \gamma Q\big(s_i[t+1], \\ & \operatorname{argmax}_{a_i} Q(s_i[t+1], a_i; w_i[t]); w_i'[t]\big). \end{aligned} \tag{36}$$

At each time step, the user takes its BS selection action according to the ϵ-greedy policy. The value of ϵ gradually decreases over training episodes to guarantee that the algorithm is able to explore the network environment during the early states of the training and to exploit the learned experiences in the later episodes. The user saves the tuple of $(s_i[t], a_i[t], r[t+1], s_i[t+1])$ as a transition (or a sample) in its own replay memory. The user randomly selects a mini-batch $\mathcal{B}_{\text{batch}}$ of B_{batch} samples to update the weight of w_i of the online network to minimize the loss function defined by

$$\arg\max_{w_i} f_i(w_i[t]) = \sum_{\mathcal{B}_{\text{batch}}} \left[z_i[t] - Q(s_i[t], a_i[t]; w_i[t])\right]^2 \tag{37}$$

2) Decentralized implementation stage: At each time step t, user i observes the local environment $s_i[t]$ as expressed in (31). Then, it chooses the BS following its trained DDQN. These is no need to share the reward between the users in this stage, and thus this state is considered to be the decentralization stage.

To solve the problem given in (37), popular algorithms such as Adam optimizer [34] and RMSProp optimizer [35] can be used. In this work, we leverage the Adam optimizer that is considered to be the combination of the momentum technique [36], the RMSProp optimizer, and a bias correction technique. The Adam optimizer is thus expected to improve the reward, compared with the RMSProp optimizer. The Adam algorithm is implemented from step 24 to step 27 in Algorithm 2. Particularly, m_i and v_i are the first momen estimates, respectively. $\hat{m}_i$ and $\hat{v}_i$ are the bias corrections of m_i and v_i, respectively, which are introduced to adapt the learning rate. Moreover, η_1, η_2, ς, and ι are the constants.

Algorithm 2 MADRL algorithm for the network selection of the URLLC and eMBB users

```
Initialization: $S$, $M$, $T$, $W_s, W_m, W_z$, $v_{s,e}, v_{m,e}, v_{z,e}, v_{s,u}, v_{m,u}, v_{z,u}, C_s, C_m, C_z, \alpha_s, \alpha_m, \alpha_z$, $\eta_1, \eta_2, \varsigma, \iota$;
The users set $w_i$ and $w'_i$ to random values, $v_i[0] = 0, s_0[i] = 0$;
for each episode $k$ do
    for $i \in \{e, u\}$ do
        if $k \equiv 0 \pmod 4$ then
            $w'_i = w_i$
        end if
    end for
    for each time step $t$ do
        for $i \in \{e, u\}$ do
            User $i$ observes state $s_i[t]$ and selects action $a_i[t]$ following $\epsilon$-greedy policy;.
            BS $b$ equally divides its bandwidth to its associated users and calculates the utility according to (10) if the user is the URLLC user and according to (11) if the user is the eMBB user;
            BS $b$ calculates reward $r_i[t]$ according to (32);
        end for
        The BSs cooperatively calculate the sum reward $r[t+1]$ according to (33);
        BS $b$ broadcasts the information of individual reward $r_i[t]$, sum reward $r[t+1]$, and channel coefficients $h_{s,i}$ and $h_{m,i}$ to its associated users;
        for $i \in \{e, u\}$ do
            User $i$ observes state $s_i[t+1]$;
            User $i$ stores $(s_i[t], a_i[t], r[t+1], s_i[t+1]$ in its replay memory;
            User $i$ randomly samples mini-batches $\mathcal{B}$ from memory replay;
            User $i$ updates $w_i$ to minimize loss function $f_i(w[t])$:
            repeat
                $g_i[t] := \nabla_{w_i} f_i(w_i[t-1])$;
                $m_i[t] \leftarrow \eta_1.v_i[t-1] + (1-\eta_1).g_i[t]$;
                $v_i[t] \leftarrow \eta_2.v_i[t-1] + (1-\eta_2).g_i^2[t]$;
                $\hat{m}_i[t] \leftarrow m_i[t]/(1-\eta_1^t)$;
                $\hat{v}_i[t] \leftarrow v_i[t]/(1-\eta_2^t)$;
                $w_i[t] \leftarrow w_i[t-1] - \varsigma.\hat{m}_i[t]/(\sqrt{\hat{v}_i[t]} + \iota)$;
            until $f_i(w_i[t])$ converges;
        end for
    end for
    The BSs broadcast the information of $\{N_b^u[t] + N_s^b[t]\}_{u \in \mathcal{U}, e \in \mathcal{E}, b \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{T}}$ to the users;
end for
```

V. Performance Evaluation

In this section, we provide simulation results to demonstrate the effectiveness of the fractional games and the MADRL schemes. We consider an HetNet that consists of 1 sub-6GHz BS, 2 mmWave BSs and 3 THz BSs. The coordinates (in meter) of the sub-6GHz BS, mmWave BSs, and THz BSs are listed in Table II. In particular, the sub-6GHz BS is located at the coordinate of $[-5, 0]$ m. The number of URLLC users is 100, and the number of eMBB users is 100. The users are randomly distributed in an area of 200 m × 150 m where the center is the sub-6GHz BS. Simulation parameters related to the BSs, users, and wireless channels are shown in Table III.

TABLE II
Coordinates of the BSs and users.

BS type	Service No.	Coordinate (in meter)
mmWave BSs	1	$[25, 0]$
	2	$[-20, 0]$
Thz BSs	1	$[0, 10]$
	2	$[-10, -10]$
	3	$[-10, 10]$
Sub-6GHz BS	1	$[-5, 0]$
User group	1	$[0, 5]$

A. Performance Evaluation of the Fractional Games

To evaluate the fractional games, we introduce the classical game that is the fractional game with the fractional factor set to $\xi = 1.0$. The fractional games are obtained by setting $\xi > 1.0$ or $\xi < 1.0$. Particularly, we consider the fractional games with $\xi = 1.1$ and $\xi = 0.7$. Note that different values of ξ are also considered during the performance evaluation.

1) Network selection strategy convergence: First, we show the selection strategy convergence of the users. Figs. 2(a), (b), and (c) illustrate the portions of the users selecting different BSs with the classical game, i.e., $\xi = 1.0$, and fractional games, i.e., $\xi = 0.7$ and $\xi = 1.1$, respectively. As seen, the strategies of the users with all the games are able to converge to the stable values. However, the convergence speeds of the fractional games are slower than that of the classical game. Particularly, the classical game converges at time instant of $t = 20$, while the fractional games, i.e., $\xi = 0.7$ and 1.1, converge at $t = 100$. The reason can be that with the fractional games, the users account for their past strategy information that makes the strategy selection longer.

2) Network selection strategy adaptation: Strategy adaptation refers to how often the users change their selection strategy in an time instant. The strategy adaptation shows the dynamics that leads to the high utility for the users. Fig. 3(a) shows the average strategy adaptation values of the games over time instants. Note that from the time instant $t > 20$, the games start their convergences. As seen, the fractional game with $\xi = 1.1$ has the highest strategy adaptation value and that with $\xi = 0.7$ has the lowest value. Meanwhile, strategy adaptation value of the classical game is between those of the two fractional games. In other words, the fractional game with $\xi = 1.1$ is the highest dynamic, followed by the classical game and the fractional game with $\xi = 0.7$. Again, the high dynamics enables the high frequency in changing its BS selection to achieve a high utility, but this also causes a large network resource cost to the users. The benefits and shortcoming of the games are discussed in the next sections.

3) Utility of users with different games: Next, it is important to demonstrate the improvement of the fractional game compared with the classical game. Fig. 3(b) shows the total

TABLE III
SIMULATION PARAMETERS.

Parameter	Value	Parameter	Value
$\{s, m, z\}$	$\{1, 2, 3\}$	$\{L_s, L_m, L_z\}$	$\{1, 8, 8\}$
N_0	-174 dBm	$\{G_i, G_s\}$	$\{0 \text{ dBi}, 0 \text{ dBi}\}$
$\{W_s, W_m, W_z\}$	$\{20 \text{ MHz}, 100 \text{ MHz}, 500 \text{ MHz}\}$ [9]	h_0	$\exp(1)$
$\{L_u, \tau_u\}$	$\{0.5 \text{ ms, } 256 \text{ bits }\}$	$\{h_{s,i}, h_{m,i}\}$	$\{h_0, h_0\}$
$\{v_{s,e}, v_{m,e}, v_{z,e}\}$	$\$\{2 \times 10^{-6}, 4 \times 10^{-7}, 2 \times 10^{-7}\}$/bit	μ	$\exp(-2)$
$\{\overline{W}^{\text{blk}}, \overline{L}^{\text{blk}}\}$	$\{10 \text{ m}, 10 \text{ m}\}$	$\{p_s^{\text{tx}}, p_m^{\text{tx}}, p_z^{\text{tx}}\}$	$\{40 \text{ dBm}, 30 \text{ dBm}, 30 \text{ dBm}\}$ [9]
$\{\alpha_s, \alpha_m, \alpha_z\}$	$\{2.1.2.5, 3.0\}$ [9]	$\{f_s, f_m, f_z\}$	$\{2.4 \text{ GHz}, 73 \text{ GHz}, 0.8 \text{ THz}\}$ [9]
ε	10^{-7}	λ^{blk}	100 blockages/km^2
$\{v_{s,u}, v_{m,u}, v_{z,u}\}$	$\$\{2.2 \times 10^{-6}, 6 \times 10^{-7}, 3 \times 10^{-8}\}$/bit	$\{C_s, C_m, C_z\}$	$\{10^{-6}, 5 \times 10^{-8}, 10^{-8}\}$ \$/bit
$k_a(f_z)$	0.03 m^{-1}	c	3×10^8 m/s

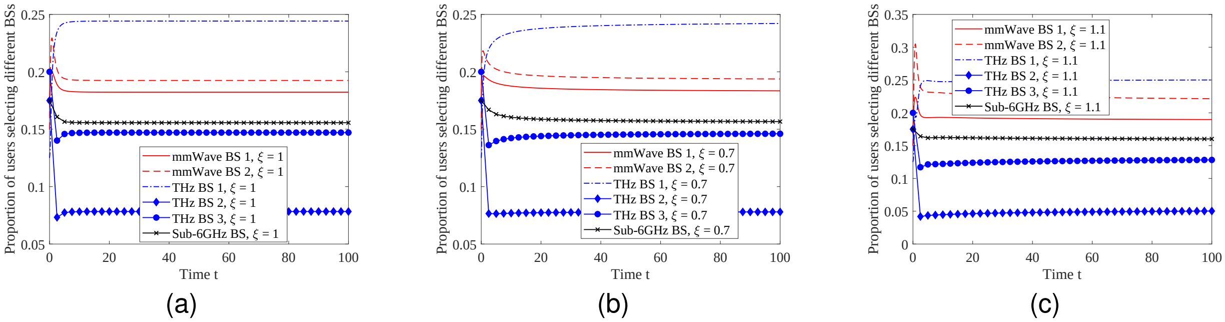


Fig. 2. Proportion of users selecting different BSs.

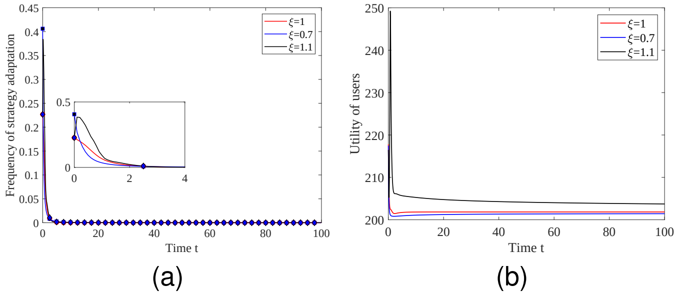


Fig. 3. (a) The strategy adaptation frequency of the users and (b) Total utility of the URLLC and eMBB users.

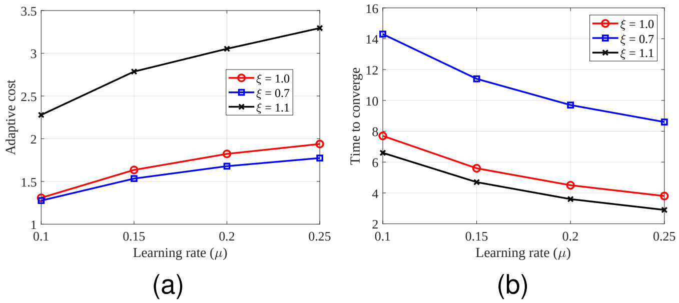


Fig. 4. (a) Adaptive cost and (b) convergence time of different games.

utility of all the users, i.e., URRLC users and eMBB users, in the network as the different games are used. As seen, the total utility obtained by the URLLC and eMBB users in the fractional game with $\xi = 1.1$ is higher than that in the classical game, i.e., $\xi = 1.0$. This result confirms that by accounting the past experiences, the fractional game allows the users to make better strategic decisions.

4) Adaptive cost and convergence time: The network selection adaptation of the users causes the energy and network resource consumption as well as communication overhead. In general, the higher network selection adaptation causes higher resource consumption and communication overhead. For this, we introduce the concept of adaptive cost that represents the network resource consumption. As shown in Fig. 4(a), the adaptive cost obtained by the fractional game with $\xi = 1.1$ is the highest and the adaptive cost obtained by the fractional game with $\xi = 0.7$ is the lowest. The reason is explained based on Section V-A1. Particularly, the game with $\xi = 1.1$ has the highest selection strategy frequency, meaning that the users have the highest network switching frequencies. Thus, the resource consumption is the highest. The results in Fig. 4(a) also show that as the learning rate increases, the adaptive cost of the games increase due to the high strategy adaptation frequency. However, this helps the users to spend less time for the convergence as shown in Fig. 4(b).

5) Direction field of the replicator dynamics and utility of the eMBB users and URLLC users: Figs. 5(a) and (b) illustrate the direction field of the replicator dynamics of eMBB user 1 and URLLC user 1, respectively. As seen, the users adapt their strategies over time that are presented by the directions of the arrows to reach the equilibrium. These results demonstrate how the game stability is obtained. At the equilibrium, although the users may select the different network, their utility should be the same as shown in Figs. 6(a), (b), and (c). The reason is that at the equilibrium, the utility of the users selecting any network is equal to their expected

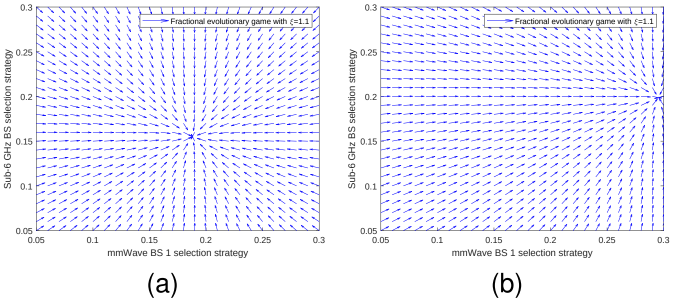


Fig. 5. Direction field of replicator dynamics of (a) eMBB users and (b) URLLC users as the game with $\xi = 1.1$ is used.

utilization.

6) Impact of the number of URLLC users on the total utility of the eMBB users: For this, we consider the proportion of the eMBB users selecting a specific BS, e.g., sub-6GHz BS, with the fractional game, as shown in Fig. 7(a). As seen, the proportion of the eMBB users selecting the sub-6GHz BS decreases as the number of URLLC users increases. The reason is that as the number of URLLC users increases, a larger amount of bandwidth of the sub-6GHz are allocated to the URLLC users. Therefore, the eMBB users attempt to move to other BSs to achieve higher utility. Therefore, the proportion of the eMBB users selecting the sub-6GHz BS decreases. Fig. 7(a) further shows that different types of URLLC users in terms of data rate requirement (defined by L_u/τ_u) impacts the utility of the eMBB users. As seen, as there are URLLC users with higher data rate requirements in the network, the utility of the eMBB users selecting the sub-6GHz BS decreases. The reason is that such a URLLC user requires more bandwidth to satisfy its data rate requirement, which decreases the bandwidth allocated to the eMBB users and further their utility.

7) Impact of the number of URLLC users on the network selection strategies of the eMBB users: As shown in Fig. 7(b), as the number of URLLC users increases, the eMBB users switch from the sub-6GHz and mmWave BSs to the THz BSs, i.e., THz BS 2. The reason is that the THz BSs have a vast amount of bandwidth and hence can provide the eMBB users higher data rates and utility compared with the sub-6GHz and mmWave BSs.

8) Impact of information delay on the game convergence: To adapt their network selection strategies, the URLLC users and eMBB users need the knowledge of the average utility information transmitted from the BSs at each time slot. This can cause high overhead communication. Thus, similar to the works [16] and [17], we now discuss how the convergence is guaranteed as the URLLC users and eMBB users use the past information. We assume that the users use the past information delayed by δ time units. Then, the game formulation is

$$ {}_{0}^{\mathrm{C}}D_t^{\xi} x_{b,i}[t] = \exp{(\mu)} x_{b,i}[t-\delta] \left(\Pi_{b,i}[t-\delta] - \bar{\Pi}_i[t-\delta]\right) $$

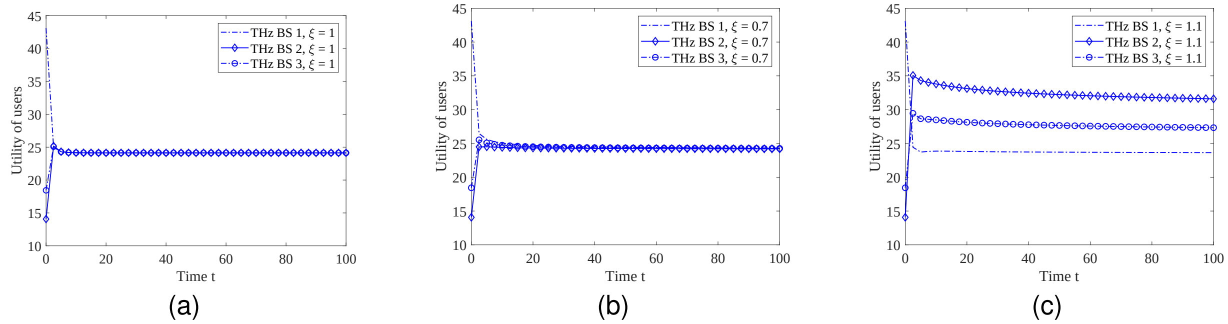
for all $i \in \{e, u\}$, $b \in \{s, m, z\}$, where $\xi \in (0, 2), \xi \neq 1$.

To illustrate the impact of information delay on the game convergence, we consider the proportion of the users selecting THz BS1. Figs. 8(a), (b), and (c) show the proportions of the users selecting THz BS1 for the games with $\xi = 1, 0.7$ and 1.1, respectively. In these figures, $\delta = 0$ means that the games use the current information, and no past information is used. It can be seen that all the games are able to converge to the stable values as the past information is used, i.e., $\delta > 0$. This implies that the games can use the past information for their network selection strategies. However, the convergence speed in this case is slower, especially as the fractional game with $\xi = 1.1$ is used. However, when δ is large, the convergence of the games is not guaranteed. For example, the convergence of the fractional game with $\delta = 1.1$ and the game with $\delta = 1$ are not guaranteed with $\delta > 28$ and $\delta > 43$, respectively. In this case, we can say that the outdated information cannot be used for the users to make their network selection. These results further show that the fractional game with $\xi = 1.1$ is less robust to the outdated information since it requires fresher information for the convergence.

B. Performance Evaluation of MADRL Algorithm

In this section, we discuss simulation results obtained by the proposed MADRL algorithm, namely ADAM-MARL. To evaluate the proposed MADRL, we implement the fractional games and classical game as discussed in Section V-A as baseline schemes. We also introduce the popular multi-agent reinforcement learning based on RMSProp optimizer [35], namely RMSProp-MARL, as a different baseline scheme. It is known as non-momentum version of Adam, which adaptively adjusts the learning rate based on the variation in the gradients only. For the network system, we use the same simulation parameters similar to the game approach that are listed in Tables II and III. This is to provide the fair evaluation among the algorithms. Furthermore, we consider that the channels of $\{h_{s,i}\}$ and $\{h_{m,i}\}$ follow a Gaussian distribution of $\mathcal{CN}(h_0, \sigma_h)$ with $\sigma_h = 0.005h_0$. For the reinforcement learning implemented at each user, the online network and target network have the same structure. Particularly, the two networks use a neural network consisting of 3 fully connected hidden layers with $256, 128$, and 64 neurons. The rectified linear unit (ReLU) function is used as the activation function at the neural nodes. We set the learning rate to 0.001 and the mini-batch size to 2000. It is noted that if the mini-batch size is too large, the computation time and memory usage of the learning algorithm increase, which results in reducing its ability to generalize. Conversely, a too small mini-batch size can result in instability during the learning process. The algorithm may converge slowly or fail to converge. Moreover, a too small mini-batch size can hinder the learning of rare samples or diverse representations in the data. Our tuning experience shows that a value of mini-batch size of 2000 is able to address the above trade-off. The training process of the reinforcement learning algorithm at each user consists of 300 episodes, and each episode is implemented in 150 steps. The exploration rate, i.e., ϵ, linearly decreases from 1 to 0.02 over the first 200 episodes and remains constant thereafter.

First, we show the convergence of the multi-agent reinforcement learning algorithms. As shown in Fig. 9(a), both the learning algorithms are able to converge to their reward values.

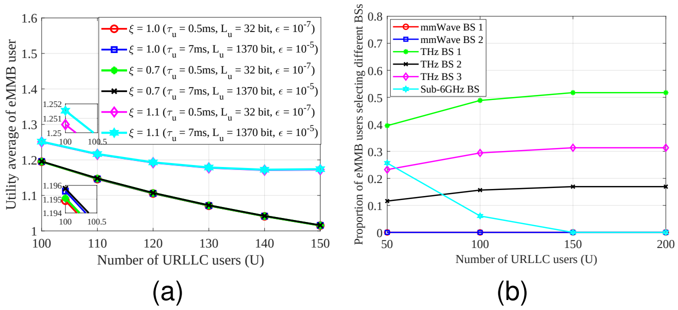


(a)

(b)

(c)

Fig. 6. Utility achieved by the users with the game approaches.



(a)

(b)

Fig. 7. Impact of the number of URLLC users on (a) the utility of eMBB users and (b) the network selection strategies of the eMBB users.

This demonstrates the feasibility of the learning algorithms in effectively solving the network selection problem of the users. Compared with the RMSProp-MARL, the proposed ADAM-MARL algorithm seems to converge faster. However, the reward achieved by the ADAM-MARL is much higher than that achieved by the RMSProp-MARL. This demonstrates the effectiveness of Adam compared with the RMSProp due to the addition of the momentum term.

We compare the utility obtained by the proposed game and ADAM-MARL approaches. As shown in Fig. 9(b), for the given number of users, the total utility obtained by the proposed ADAM-MARL algorithm is always higher than that obtained by the game approaches including the fractional game ($\xi = 1.1$). This demonstrates the benefit of the cooperation among the BSs as well as the advantage of the learning algorithm. The improvement of the ADAM-MARL compared with the game approaches maintains even with the large number of users, i.e., 160. This implies the scalability of the proposed ADAM-MARL algorithm.

VI. Conclusions

In this paper, we have investigated a HetNet network including sub-6GHz BSs, mmWave BSs, and THz BSs to support eMBB users and URLLC users. In the network, both types of the users are able to dynamically select the BSs over time to achieve their highest utility. We have designed two types of utility functions specific for the eMBB users and URLLC users. To model the selection behavior of the users, we have formulated the fractional game with PLM, which allows the users to incorporate the past experiences into their decisions to achieve higher utility. Furthermore, we have considered the case that the BSs cooperate with each other in sharing the network state. Then, we have developed a multi-agent deep reinforcement learning algorithm that allows the URLLC users and eMBB users to make their network selection decision online to achieve their long-term utility. We have provided various simulation results to demonstrate the scalability and effectiveness of the proposed fractional game and the multi-agent learning approaches. Particular for the fractional game, its advantages (utility improvement) and shortcomings (high network adaptation cost) are presented. Moreover, how the different types and the number of URLLC users in the network impact the total utility and the network selection strategies of the eMBB users is well discussed. Finally, we have shown that the multi-agent deep reinforcement learning outperforms both the classical and fractional games in terms of utility.

References

[1] N. C. Luong, F. Shaohan, S. Gong, and D. Niyato, "Dynamic network selection for urllc and embb applications in sub-6ghz-mmwave-thz networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2024, pp. 1–6.

[2] H. Zarini, N. Gholipoor, M. R. Mili, M. Rasti, H. Tabassum, and E. Hossain, "Resource management for multiplexing embb and urllc services over ris-aided thz communication," *IEEE Transactions on Communications*, vol. 71, no. 2, pp. 1207–1225, 2023.

[3] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach," *EEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, 2021.

[4] X. Yuan, Y. Zhu, Y. Hu, B. Ai, and A. Schmeink, "Optimal user grouping and analytical joint resource allocation design in hybrid bc-tdma assisted urllc networks," *IEEE Transactions on Wireless Communications*, 2023.

[5] P. Yang, X. Xi, K. Guo, T. Q. Quek, J. Chen, and X. Cao, "Proactive uav network slicing for urllc and mobile broadband service multiplexing," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3225–3244, 2021.

[6] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2018.

[7] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between embb and urllc in 5g wireless networks," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1736–1749, 2020.

[8] R. Liu, G. Yu, J. Yuan, and G. Y. Li, "Resource management for millimeter-wave ultra-reliable and low-latency communications," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1094–1108, 2020.

[9] K. Humadi, I. Trigui, W.-P. Zhu, and W. Ajib, "User-centric cluster design and analysis for hybrid sub-6ghz-mmwave-thz dense networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7585–7598, 2022.

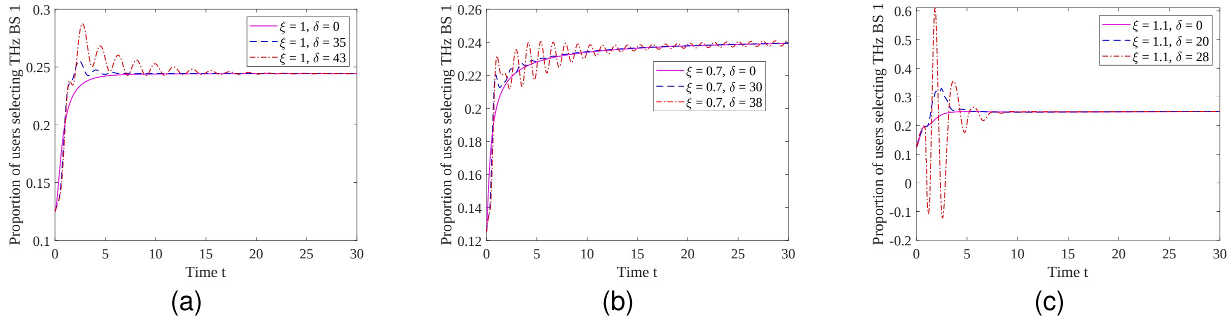


Fig. 8. Proportions of the users selecting different BSs with different information delay values.

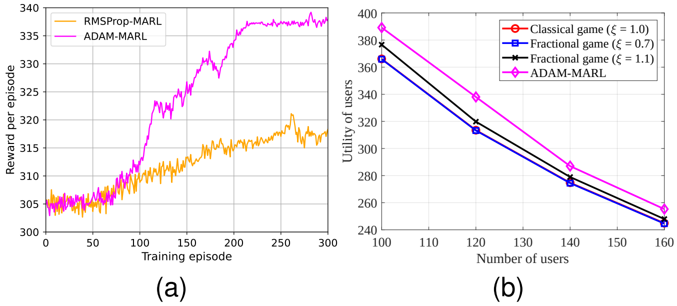


Fig. 9. (a) Convergence of the MARL algorithms and (b) total utility obtained by the game and ADAM-MARL approaches.

[10] C. Xu, M. Sheng, V. S. Varma, T. Q. Quek, and J. Li, "Wireless service provider selection and bandwidth resource allocation in multi-tier hcns," *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 5108–5124, 2016.

[11] C. Skouroumounis and I. Krikidis, "An evolutionary game for mobile user access mode selection in sub-6 ghz/mmwave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5644–5657, 2022.

[12] N. Saha and R. Vesilo, "An evolutionary game theory approach for joint offloading and interference management in a two-tier hetnet," *IEEE Access*, vol. 6, pp. 1807–1821, 2017.

[13] C. Dai, K. Zhu, R. Wang, and Y. Xu, "Decoupled multiple association in full-duplex ultra-dense networks: An evolutionary game approach," in *IEEE International Conference on Communications*, 2019, pp. 1–6.

[14] B. Huang and A. Guo, "A dynamic hierarchical game approach for user association and resource allocation in hetnets with wireless backhaul," *IEEE Wireless Communications Letters*, vol. 13, no. 1, pp. 59–63, 2023.

[15] S. Feng, D. Niyato, X. Lu, P. Wang, and D. I. Kim, "Dynamic model for network selection in next generation hetnets with memory-affecting rational users," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1365–1379, Apr. 2020.

[16] N. T. T. Van, N. C. Luong, S. Feng, H. T. Nguyen, K. Zhu, T. Van Luong, and D. Niyato, "Dynamic network service selection in intelligent reflecting surface-enabled wireless systems: Game theory approaches," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 5947–5961, 2022.

[17] N. T. T. Van, N. C. Luong, S. Feng, V.-D. Nguyen, and D. I. Kim, "Evolutionary games for dynamic network resource selection in rsma-enabled 6g networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1320–1335, 2023.

[18] T. Li, K. Zhu, N. C. Luong, D. Niyato, Q. Wu, Y. Zhang, and B. Chen, "Applications of multi-agent reinforcement learning in future internet: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1240–1279, 2022.

[19] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.

[20] 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 Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.

[21] L. Zhang and Y.-C. Liang, "Deep reinforcement learning for multi-agent power control in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2551–2564, 2020.

[22] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2014.

[23] M. Di Renzo and W. Lu, "System-level analysis and optimization of cellular networks with simultaneous wireless information and power transfer: Stochastic geometry modeling," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2251–2275, 2016.

[24] C. Han, A. O. Bicen, and I. F. Akyildiz, "Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2402–2412, May 2015.

[25] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, 2017.

[26] C. Han, A. O. Bicen, and I. F. Akyildiz, "Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2402–2412, 2015.

[27] J. M. Jornet and I. F. Akyildiz, "Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the terahertz band," *IEEE Trans Wireless Commun.*, vol. 10, no. 10, pp. 3211–3221, Oct. 2011.

[28] C. She, C. Liu, T. Q. Quek, C. Yang, and Y. Li, "Ultra-reliable and low-latency communications in unmanned aerial vehicle communication systems," *IEEE Transactions on communications*, vol. 67, no. 5, pp. 3768–3781, 2019.

[29] Techplayon, "5g nr modulation and coding scheme – modulation and code rate," 2020. [Online]. Available: https://www.techplayon.com/5g-nr-modulation-and-coding-scheme-modulation-and-code-rate/

[30] N. C. Luong, N. T. T. Van, S. Feng, H. T. Nguyen, D. Niyato, and D. I. Kim, "Dynamic network service selection in irs-assisted wireless networks: A game theory approach," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 5, pp. 5160–5165, May 2021.

[31] S. Feng, D. Niyato, X. Lu, P. Wang, and D. I. Kim, "Dynamic game and pricing for data sponsored 5g systems with memory effect," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 4, pp. 750–765, 2020.

[32] I. Podlubny, "Fractional differential equations," *Mathematics in science and engineering*, vol. 198, pp. 41–119, 1999.

[33] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *AAAI*, vol. 30, no. 1, 2016, pp. 2094–2100.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, 2015, pp. 1–15.

[35] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop, coursera: Neural networks for machine learning," *University of Toronto, Technical Report*, vol. 6, 2012.

[36] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, Atlanta, USA, 2013, pp. 1139–1147.