DQN Based User Association Control in Hierarchical Mobile Edge Computing Systems for Mobile IoT Services

Yunseong Lee[a], Arooj Masood[a], Wonjong Noh[b,*], Sungrae Cho[a,*]

[a] *The School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, Republic of Korea*
[b] *The School of Software, Hallym University, Chuncheon 24252, Republic of Korea*

Abstract

This paper proposes a new user association control method to maximize system utility in terms of system throughput, edge throughput, and handover cost in mmWave-based HMEC (mm-HMEC) systems. We first formulated a non-convex mathematical programming model and then proposed centralized and distributed deep Q-network (DQN)-based range expansion bias control algorithms. We discovered that the proposed schemes provided polynomial communication overhead and computation complexity. Through simulations, we evaluated the proposed centralized and distributed schemes under various user equipment (UE) mobility models: random waypoint mobility model (RWM), random direction mobility model (RDM), and Manhattan mobility model (MAN). We first confirmed that the proposed distributed control achieves almost the same performance as the proposed centralized control. Second, with 10 slave-MEC servers, the proposed distributed control provides 212.17%, 62.07%, 46.36%, and 48.63% enhanced utility; 33.68%, 18.19%, 9.28%, and 12.66% enhanced average cell throughput; 14.32%, 38.32%, 63.87%, and 24.57% enhanced edge throughput; and 39.09%, 25.42%, 29.32%, and 25.95% reduced handover cost compared to random cell range expansion (CRE), standard CRE, dynamic CRE, and Q-learning-based CRE, respectively.

Keywords: Machine learning, Heterogeneous computing system, Resource allocation, Mobile networks, Edge computing, Mobility based services

1. Introduction

In future computing systems, the availability of computing resources is expected to become an increasingly challenging issue owing to the extensive growth in mobile traffic demands. To address this issue, a new computing architecture, a hierarchical mobile edge computing (HMEC) networks that consists of a master-MEC server (M-MECS) and a slave-MEC server (S-MECS), is emerging, and millimeter wave (mmWave) between 30 and 300 GHz have been attracting growing attention for deployment [1]. The mmWave based HMEC (mm-HMEC) systems have many benefits compared with existing MEC systems. First, the application of mmWave to the fronthaul link increases the degrees of freedom in the network configuration without installing a wired infrastructure. Second, the mmWave frequency band has a wide transmission bandwidth that supports high-speed transmission. Third, mmWave supports directional transmission (via massive antenna installations) and has a small transmission range which results in low interference and enables efficient spatial reuse of spectrum in a hierarchical architecture [2]. Finally, HMEC allows user content to be cached closer to the user by storing it in S-MECS. Some example services can be efficiently deployed in mm-HMEC networks.

example 1. The mobile virtual reality (VR) services, such as surgical training, military training game, education service, and streaming video require very high speed transmission of Gbps and ultra-low latency of milliseconds. Also, the self-driving vehicle services in vehicle-to-everything (V2X) environment, which continuously collect information for autonomous driving applications from road side unit (RSU), demand continuous communications with guaranteed minimum rate for the stable self-driving. That is, the wider bandwidth of fronthaul and access links, directional antenna of M/S-MECS, and closer contents caching at S-MECS of mm-HMEC enable to support those service requirements. Especially, S-MECS can be mounted on the ceiling in indoor VR service [3] and on the high-altitude RSU in self-driving service [4]. Therefore, the line of sight (LOS) transmission between S-MECS and service users can be easily obtained with little attenuation in communication.

Although mm-HMEC offers many benefits, there are some challenges to be addressed [1]. First, low complexity beamforming control techniques with mmWave-based massive antenna have to be studied [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Second, orthogonal or non-orthogonal resource scheduling in wireless fronthaul between M-MECS and S-MECS and wireless access between S-MECS and user equipment (UE), should

[⋆] This work was supported in part by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea Government (MSIT) under Grant 2018-0-00889 (A study on core technology of 5G mobile communication using millimeter wave band) and the National Research Foundation of Korea (NRF) grant funded by the Korean Government (Ministry of Science and ICT) under Grant 2020R1F1A1069119.

[*] Corresponding author

Email addresses: yslee@uclab.re.kr (Yunseong Lee), arooj@uclab.re.kr (Arooj Masood), wonjong.noh@hallym.ac.kr (Wonjong Noh), srcho@cau.ac.kr (Sungrae Cho)

be studied [15, 16, 17, 18, 19, 20]. Third, because of the short propagation distance and directional antenna angle of mmWave, frequent handovers should be addressed [21, 22, 23, 24]. In particular, densely deployed S-MECSs and high-mobility UEs can cause a much greater number of handovers in such networks. Finally, to make the most of the new infrastructure, mobile users should be actively pushed onto the S-MECSs that will often be lightly loaded. It can provide higher performance over time by offering UEs many more resource blocks than M-MECS. That is, a more balanced user association [25, 26, 27, 28, 29, 30] reduces the load on the M-MECS, allowing it to serve its remaining users better.

In this work, we focus on a deep Q-learning network (DQN)-based user association control that enhances average throughput, edge throughput, and handover cost in mm-HMEC systems. The main contributions of this study are summarized as follows:

- Unlike the proposed algorithm [25, 26, 27, 28, 29, 30], we considered improving the system throughput, edge throughput, and handover efficiency simultaneously in the mm-HMEC systems. We summarized the features and differences of and from some representative related works in Table 1.

- We first formulated an optimization problem to determine the best range expansion policy as a non-convex mixed-integer programming model, and then proposed DQN-based centralized and distributed algorithms to solve the problem.

- In the proposed DQN algorithms, we designed a reward function that consists of the average system throughput, edge users throughput, and handover cost. By controlling the weights of the reward factors, we can apply the proposed DQN to various mm-HMEC systems with different objectives.

- We analyzed the proposed centralized and distributed algorithm to provide polynomial communication overhead and computation complexity.

- We performed extensive simulations over various UE mobility models: the random waypoint model (RWM), random direction mobility (RDM), and Manhattan mobility model (MAN). We compared the proposed controls with other benchmark schemes: random cell range expansion (CRE), standard CRE, dynamic CRE, and Q-learning-based CRE. We first confirmed that the proposed distributed scheme achieved almost the same performance as the proposed centralized scheme. Comparing other schemes, under 4 S-MECSs, the proposed schemes have on average (in terms of all mobility models, and centralized/distributed operations) 20.85% higher system utility than the best benchmark scheme. Moreover, when we increase the number of S-MECSs to 10, the proposed schemes provide on average 42.38% higher system utility than the best benchmark scheme. That is, when we increased the number of S-MECSs, we can see that the performance gain (with respect to the best benchmark scheme) increases with non-increased communication overhead, and reduced computation complexity and end-to-end latency.

The remainder of this paper is organized as follows. In Section 2, we describe related works. In Section 3, we describe the mm-HMEC system model considered in this work. Section 4 presents the proposed centralized and distributed DQN-based user association control schemes. Section 5 presents operation procedures, and then analyzes computing complexity, communication overhead, and end-to-end latency. In Section 6, we describe an evaluation of the proposed scheme. Finally, we present our conclusions in Section 7. The key notations used in this paper are summarized in Table 2.

2. Related Work

2.1. Optimization based approach

Some works [5, 6, 7, 8, 9, 10, 11, 12, 13, 14] studied directional antenna-based beamforming controls in mm-HMEC systems. Hadi *et al.* [5] proposed a 3-dimensional beamformimg design to mitigate interference and improve resource utilization in a two-tier mm-HMEC. Zhai *et al.* [6] proposed hybrid beamforming to achieve high throughput and capability for backhaul links. Castanheira *et al.* [7] proposed a low complexity hybrid beamforming techniques for massive mm-HMEC to mitigate interference. Sung *et al.* [8] proposed a UE controlled beam switching mechanism that allows UEs to switch from the serving beam to a target beam without random access delay. Hao *et al.* [9] proposed a cooperative beamforming scheme to minimize fronthaul transmission delay from edge server to user. Nor *et al.* [10] proposed a light fidelity based low-complex beamforming training scheme that reduces outage probability. Nor *et al.* [11] proposed a Li-Fi location based beamforming technique in indoor environment that improves beamforming complexity. Mubarak *et al.* [12] proposed a Wi-Fi localization based low-complex beamforming training and positioning technique. Zhou *et al.* [13] proposed an imperfect channel state information based low-complex beamforming training that reduces the power consumption. Wang *et al.* [14] proposed an intelligent reflecting surface based flexible beamforming training and iterative positioning algorithm that enhances the estimation of angle of arrival and angle of departure.

Some works [15, 16, 17, 18, 19, 20] studied fronthaul and backhaul resource allocation, scheduling, and traffic load control in mm-HMEC systems. Na *et al.* [15] proposed a directional routing and backhaul link scheduling algorithm to reduce end-to-end delay and avoid the deafness problem in mmWave-based smart manufacturing system network. Ding *et al.* [16] proposed a QoS aware full-duplex concurrent scheduling algorithm for mm-HMEC backhaul to improve the number of flows and satisfy QoS requirements. Semiari *et al.* [17] proposed a game-theoretic optimal backhaul link matching scheme to improve the average sum rate in mm-HMEC. Li *et al.* [18] proposed a heuristic resource and user scheduling algorithm for mm-HMEC to maximize data rate with lower end-to-end delay.

Table 1: Summary of Key Features and Differences

Ref.	Utilized Technique	Performance Metrics	Antenna model	Features
[25]	Per-tier bias control using convex optimization	Coverage probability, handover	Omnidirectioanl	LTE-HMEC, HPPP, UE mobility
[26]	Per-BS heuristic bias control through system utility comparison	Throughput	Omnidirectioanl	LTE-HMEC, ABS, PFTF scheduling
[27]	Per-BS bias control using particle swarm optimization	Coverage probability, throughput	Omnidirectioanl	5G-HMEC, JT-CoMP
[28]	Per-tier static bias control	Coverage probability	Omnidirectioanl	5G-HMEC, CoMP, user-centric BS clustering
[29]	Per-BS bias control using stochastic geometric analysis and system utility function	Coverage probability	Omnidirectioanl	5G-HMEC, three-tier network (macro, pico, femto), HPPP
[30]	Per-BS bias control using Q-learning	Coverage probability, throughput	Omnidirectioanl	LTE-HMEC, edge UE
Proposed	Per-beam bias control using deep Q-Network	System throughput, edge throughout, handover	Directional	mm-HMEC, edge UE

ABS = Almost blank subframes, BS = Base station, HPPP = Homogeneous Poisson Point Process, JT-CoMP = Joint Transmission-Coordinated Multipoint Transmission, PFTF = Proportional Fair in Time and Frequency, Ref. = Reference

Yang *et al.* [19] proposed an adaptive traffic allocation strategy to reduce end-to-end delay in mm-HMEC. Jia *et al.* [20] proposed an effective contents delivery scheme in mm-HMEC backhaul to reduce average delay of delivering content.

Some works [21, 22, 23, 24] studied handover controls in mm-HMEC systems. Ren *et al.* [21] proposed multiple beams cooperation (MBC)-based inter-beam handover schemes: jointly optimized dynamic inter-beam handover (JOD-IBH) scheme and seamless inter-beam handover mechanism. They improved the inter-beam handover performance by achieving a balance between the handover failure rate and resource occupation rate. Mazzavilla *et al.* [22] proposed a handover approach that considers dynamic channel load and handover overhead. Guidolin *et al.* [23] proposed a context-aware handover scheme that jointly considers traffic load, UE mobility, and cell size. Zang *et al.* [24] proposed a user mobility-aware handover scheme between S-MECS and M-MECS. It considered the Gauss-Markov mobility model to determine the time spent in S-MECS.

Some works [25, 26, 27, 28, 29] studied user association controls in mm-HMEC systems. Sadr *et al.* [25] proposed a range expansion control to maximize system performance, such as coverage probability with and without accounting for user mobility in multi-tier networks. Al-Rawi *et al.* [26] studied the impact of dynamically changing the range of low power nodes (LPNs). They proposed a simple heuristic method that adapts the size of the LPNs to load and interference situation. Shami *et al.* [27] utilized joint transmission coordinated multipoint (JT-CoMP) and particle swarm optimisation (PSO) to assign each S-MECS a specific biasing value to balnce and control the load among BSs while the overall throughput of the system is still maximized. Subsequently, Shaddad *et al.* [28] utilized coordinated multi-point transmission (CoMP) to reduce interference and proposed user centric clustering where a user can be served by a number of MECSs. Jiang *et al.* [29] proposed a novel system parameter (biasing factor) to change the user association to achieve load balancing of the entire system and better QoS of users by solving a network-wide utility maximization problem.

2.2. Machine learning based approach

Conventional optimization approaches struggle with a heavy computation load in dynamic environments that require real-time solutions. However, recent developments in machine learning algorithms have significantly reduced the computational complexity of solutions generated at different times through the optimization of a learning model over all possible system state realizations.

Some studies [31, 32] applied machine learning-based beamforming controls in mm-HMEC systems. Zhang *et al.* [31] proposed a convolutional neural networks based coordinated beamforming scheme to maximize energy efficiency in mm-HMEC. Kim *et al.* [32] proposed an online learning-based beamforming scheme to quantify beam interference and maximize successful transmissions in ultra-dense mm-HMEC.

Some studies [33, 34] proposed machine learning based resource controls in mm-HMEC systems. Ryu *et al.* [33] proposed a deep reinforcement learning based power control scheme to maximize energy efficiency and user throughput for backhaul link in mm-HMEC. Vu *et al.* [34] proposed a reinforcement learning based backhaul path selection algorithm to reduce latency in mm-HMEC.

Some studies [35, 36, 37, 38, 39] proposed machine learning based handover controls in mm-HMEC systems. Yan *et al.* [35] proposed a k-nearest neighbor (KNN)-based handover control mechanism. It predicts vehicle positions to subsequently use them to pre-activate target mmW-RRUs in mmWave vehicle-to-infrastructure (V2I) networks. Sun *et al.* [36] proposed a multi-armed bandit (MAB)-based handover control method in mmWave cellular networks. It exploits user post-handover mobility trajectory and LoS blockage to maximize user-BS connection time after each handover. Mollel *et al.* [37] proposed a DQN-based handover (optimal MECS selection) control method in mmWave-based ultra-dense networks to maximize the user throughput. It considers the received SNR from the serving BS and prolonged user connectivity. Sun *et al.* [38] proposed a handover control with reinforcement learning (RL) by considering

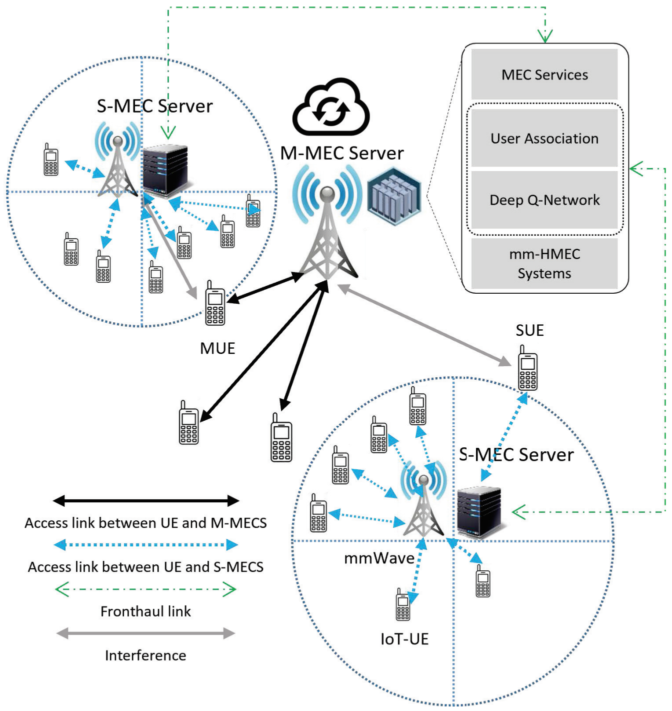


Figure 1: mm-HMEC system architecture

mmWave channel characteristics, UE service requirements, and user densities. Junhong *et al.* [39] proposed a deep RL-based MECS activation scheme in the presence of continuous traffic demands in HMEC systems. In this approach, traffic arrivals were predicted using a deep neural network (DNN) by exploiting the space-time correlation of data arrival patterns.

Some studies [30, 40] proposed machine learning based user association controls in mm-HMEC systems. Kudo *et al.* [30] proposed a Q-learning-based range expansion control to minimize network service outages while improving average throughput. Khan *et al.* [40] proposed a distributed (RL)-based user association algorithm, in which each MECS is equipped with a local RL agent to learn association policy and select local actions. It is demonstrated that each independently trained RL network performs user association actions with low control overhead and low computational complexity by maximizing the mobile user experience in terms of network-wide access while guaranteeing a minimum level of service for all UE.

3. System Model

3.1. Network Model

We considered an mm-HMEC system architecture consisting of an M-MECS and several S-MECSs. The M-MECS is denoted by m, and the set of S-MECSs is denoted by $\mathbb{S} = \{s_1, ..., s_S\}$. The set of UEs is denoted by $\mathbb{U} = \{u_1, ..., u_U\}$. The M-MECS and S-MECSs exchange their information using mmWave-based fronthaul links, and the M/S-MECSs and UEs share their information using mmWave based access links. In the proposed system, the S-MECS operates a directional array antenna composed of a fixed number of beams, and each fixed beam has a specific transmission power and antenna angle.

Table 2: Key Notation Descriptions

Notation	Description
$\mathbb{B}$	Set of all HMECs in mm-HMEC systems
$\mathbb{S}$	Set of S-MECSs in mm-HMEC systems
$\mathbb{U}$	Set of UEs in mm-HMEC systems
$\mathbb{D}$	Set of beams of MECS
m	M-MECS
RSRP_b^d	Reference signal received power of the dth directional beam in BS b
bias_b^d	Bias value for dth directional beam in MECS b
p_b	Transmit power of MECS b
g_b	Channel gain between MECS b and UE u
δ_u^b	SINR of UE allocated with MECS b
W_b	MECS bandwidth
N_b	Number of UEs associated with MECS b
T_t^u	Throughput of UE u at time t
e_u	Indicator value for edge UE
c_u^b	Indicator value for the link between UE u and MECS b
v_{s_i}	Number of edge UEs in S-MECS s with i directional beam
v_t	Vector of the number of edge UEs in all S-MECSs at time t
ψ_{min}	Minimum throughput required for edge UE
ψ_t	Vector of the average throughput of edge UEs in all S-MECSs at time t
μ_t^η	Average throughput of all UEs at time t
ρ_t	Vector of the number of UEs assigned to all S-MECS at time t
h_t^u	Handover cost of UE u at time t
m_t	Number of UEs assigned to the M-MECS at time t
w	Weight value
$\bar{y}$	Target Q-value
θ	Parameter of training Q-network
γ	Discounting factor
β	Minibatch
τ	Delay time
f	Computing cycle frequency
I	Packet size

3.2. User Association Model

We assume that each UE is associated with either the M-MECS or S-MECS. Each user selects its serving MECS, $b_{serving}$, based on the downlink signal strength in the following manner:

$$b_{serving} = \arg\max_{b\in\mathbb{B}, d\in\mathbb{D}_b} (\mathrm{RSRP}_b^d + \mathrm{bias}_b^d) \text{ in [dB]}, \qquad (1)$$

where $\mathbb{B}$ is the set of all MECSs; $\mathbb{D}_b$ is the set of directional beams equipped in MECS b; RSRP_b^d is the reference signal received power from beam d of MECS b; bias_b^d is the range expansion bias, which is zero for the M-MECS and x ($x \geq 0$)

for the low-power S-MECSs. The UE associated with the M-MECS and S-MECS are referred to as the user in the M-MECS (MUE) and the user in the S-MECS (SUE), respectively. This user association scheme is called cell range expansion (CRE) [41, 42], which is a standard user association approach included in the 3GPP Rel. 11 [43]. Each MECS's beam has its own range expansion bias value bias$_b^d$. The user association can be different according to the change in the range expansion bias. First, if the virtually biased beam of a serving S-MECS becomes stronger, and the beam of an M-MECS becomes weaker, then some MUEs can move toward the S-MECS. In this situation, UE can conduct a handover from the M-MECS to the S-MECS. Second, if the virtually biased beam of a serving S-MECS becomes weaker and the beam of an M-MECS becomes stronger, then some SUEs can move toward the M-MECS. In this situation, UE can conduct a handover from the S-MECS to the M-MECS. Finally, an SUE is located at the boundary region of the S-MECS's directional beams and receives multiple side-lobe beams. In this case, SUE can conduct a handover to another beam of the S-MECS.

3.3. Interference Model

In mm-HMEC systems, various types of interference are encountered, such as between S-MECSs, M-MECS and S-MECS. This interference has a significant impact on the performance of mm-HMEC systems. The signal-to-interference-plus-noise ratio (SINR) measurement process differs according to the types of serving MECS. The SINR of an MUE is defined as follows:

$$\delta_u^m = \frac{p_m g_u^m}{\sum\limits_{x\in\mathbb{S}} \sum\limits_{y\in\mathbb{D}} p_{x_y} g_u^{x_y} + \sigma^2}, \tag{2}$$

where p_m is the transmit power of the serving M-MECS m; g_u^m is the channel gain between the serving M-MECS m and UE u; p_{x_y} denotes the transmission power of the yth beam of the S-MECS x; $g_u^{x_y}$ denotes the channel gain between the yth beam of the S-MECS x and UE u; σ^2 denotes noise power. Similarly, the SINR of an SUE is defined as follows:

$$\delta_u^{s_i} = \frac{p_{s_i} g_u^{s_i}}{\sum\limits_{x\in\mathbb{S},} \sum\limits_{y\in\mathbb{D}, y\neq i} p_{x_y} g_u^{x_y} + \sigma^2}, \tag{3}$$

where p_{s_i} is the transmission power of the ith beam of the serving S-MECS s; $g_u^{s_i}$ is the channel gain between the ith beam of the serving S-MECS s and UE u. According to each SINR, the throughput for MUE and SUE are defined as follows:

$$T_t^{u,m} = \frac{W_m}{N_m} \log_2 (1 + \delta_u^m), \tag{4}$$

$$T_t^{u,s_i} = \frac{W_{s_i}}{N_{s_i}} \log_2 (1 + \delta_u^{s_i}), \tag{5}$$

where W_m denotes the bandwidth of the M-MECS ; W_{s_i} denotes the bandwidth of the ith beam of the S-MECS s; N_m denotes the number of UE associated with the M-MECS m; N_s denotes the number of UE associated with the i-th beam in the S-MECS s.

3.4. Problem Formulation

In this work, we target some multimedia based mobile unicast or multicast services such as VR/AR/Metaverse streaming service, V2X application services, or factory automation service in mm-HMEC, which will be very dominant services in future. For the efficient supports for those kinds of services, besides that the minimum throughput requirement should be guaranteed, in terms of user-perspective, high throughput should be provided as much as possible, and in terms of system-perspective, low handover cost (in fact, 5G system targets zero handover latency, but it demands many signal processing or computation cost for the seamless soft-handover) should be provided.

The existing range expansion technique presents several challenges. For example, when the bias value is too high, throughput degradation occurs by accommodating more UEs than an S-MECS can accommodate. Otherwise, when the bias is low, the utilization of the S-MECS decreases because UEs at the S-MECS's boundary are accommodated by the M-MECS. Moreover, if the bias values are not well-coordinated among beams, they result in frequent or unnecessary handovers or excessively suppress handovers so that UE loses opportunities to connect to MECS that provides higher throughput. Therefore, we need to design a system that allows each beam of the S-MECS to dynamically adjust its own range expansion bias according to the state of the network. More specifically, we aim to maximize throughput and minimize handover cost under edge throughput constraint by dynamically controlling the range expansion bias, as follows:

$$\max_{\mathbf{B}} \quad \sum_{t=0}^{\mathbf{T}} \mathbf{U_B}(t) \tag{6}$$

$$\text{s.t} \quad \sum_{b\in\mathbb{B}} \sum_{j\in\mathbb{D}} c_u^{b_j} \leq 1, \forall u \in \mathbb{U}, \tag{7}$$

$$c_u^{b_j} \in \{0, 1\}, \qquad \forall b \in \mathbb{B}, \forall j \in \mathbb{D}, \forall u \in \mathbb{U}, \tag{8}$$

$$\Psi_{\mathbf{B}}(t) \geq \psi_{min}, \quad \forall t \in \mathbf{T}. \tag{9}$$

In (6), the objective utility function $\mathbf{U_B}(t)$ is defined as a weighted sum of the average throughput $\Lambda_{\mathbf{B}}(t)$ and average handover cost $\Omega_{\mathbf{B}}(t)$ at time t,

$$\mathbf{U_B}(t) = w_1 \Lambda_{\mathbf{B}}(t) + w_2 \Omega_{\mathbf{B}}(t), \tag{10}$$

where $0 < w_1 < 1$ and $0 < w_2 < 1$ denote the weight values of the throughput and handover cost, respectively. Here, the average throughput $\Lambda_{\mathbf{B}}(t)$ for all UEs at time t is defined as follows:

$$\Lambda_{\mathbf{B}}(t) = \frac{1}{|\mathbb{U}|} \sum_{u\in\mathbb{U}} T_t^u, \tag{11}$$

where $\mathbf{B}$ is the bias vector for MECS beams; $|\mathbb{U}|$ is the number of UEs; T_t^u is the throughput of UE u at time t. In addition, the average handover cost for all UEs at time t is defined as follows:

$$\Omega_{\mathbf{B}}(t) = -\frac{1}{|\mathbb{U}|} \sum_{u\in\mathbb{U}} h_t^u, \tag{12}$$

where h_t^u is 1 if the handover for UE u occurs at time t, otherwise, it is 0. That is, the average handover cost means the average number of handover occurred. Each handover can incur handover costs, such as signal processing overhead, energy consumption, and etc. This cost can be implicitly reflected in the coefficient w_2. In (7) and (8), each UE is associated with only one beam, and $c_u^{b_j}$ denotes the indicator variable (i.e., $c_u^{b_j} = 1$ when UE u is associated with the j-th beam in MECS b; otherwise, $c_u^{b_j} = 0$). In (9), the average throughput of the edge UEs should be greater than the minimum rate, ψ_{min}. Here, if a UE belongs to the lower 5% in terms of throughput, the UE is classified as an edge UE. The average throughput of the edge UEs is defined as follows:

$$\Psi_{\mathbf{B}}(t) = \frac{1}{|\mathbb{E}|} \sum_{j \in \mathbb{E}} T_t^j, \tag{13}$$

where $\mathbb{E}$ and $|\mathbb{E}|$ are the set of edge users and number of edge users, respectively.

Remark 1. The problem in (6)–(7) is a mixed-integer programming (MIP) problem because of the existence of multiple discrete and continuous variables. Moreover, MIP problems are known to be NP-hard by nature, and finding an optimal solution usually requires exponential time complexity [44]. Because the channels between UE and MECSs dynamically change over time, a large number of possible channel realizations can be generated, which poses a challenge when applying conventional optimization solutions in real time. Therefore, to employ real-time user association control, we propose a novel solution based on deep RL.

4. DQN-based User Association Control for mm-HMEC Systems

In this section, we propose a DQN-based bias control model for mm-HMEC systems, whose overall architecture is shown in Fig. 2.

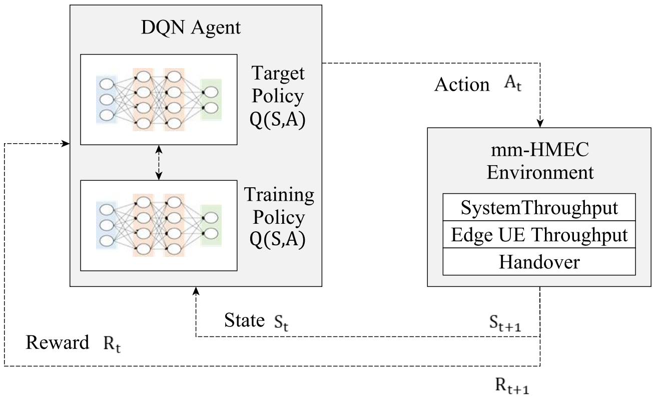


Figure 2: DQN architecture for the proposed user association control

4.1. Overview of RL and Deep Q-Learning

RL is a robust machine learning technique that aims to maximize long-term rewards by interacting with the environment [45]. The outstanding features of RL are trial-and-error search, which presents a trade-off between exploration and exploitation, and the reward, which implies that an agent can consider either immediate or cumulative reward in the long run using discounting factors [45]. In RL, the environment can be described as a Markov decision process (MDP) wherein the state space, state-transition probability, and reward function are not necessarily required [46]. RL can be classified into model-based and model-free types, depending on the existence of an environmental state-transition probability model. Although model-based RL must perform supervised learning with an inherent model, model-free RL can train the parameters starting from zero. Furthermore, some recent studies have shown that model-free RL can handle DNN effectively [47, 48]. Gu *et al.* [49] proposed combining the strengths of model-based and model-free RL to accelerate learning.

Q-learning is a typical model-free RL algorithm wherein state-action Q-values are stored in a reference table or evaluated by a nonlinear approximator, such as DNN. Deep Q-learning was initially introduced to teach machines how to play games without human intervention [46]. This algorithm uses a neural network called DQN to process the raw-state representation input directly. Given the neural network parameter θ, the Q-value function can be represented by $Q(\mathcal{S}, \mathcal{A}; \theta)$, where s and a denote the state vector and action, respectively. The neural network is trained by updating θ to approximate the Q-value based on the interaction experiences of the agent. Mnih *et al.* [47] proved that deep Q-learning is more advantageous than conventional Q-learning, owing to its higher performance and faster convergence. However, deep Q-learning occasionally learns unrealistically high Q-values because it includes a maximization step, which tends to overestimate Q-values. For example, a DQN estimates a current-state Q-value from the achieved reward and a discounted next-state Q-value, which is overestimated as the maximum value among the estimations over the possible actions. To solve this problem, a double Q-learning algorithm, fully named deep double Q-learning [50], was introduced to decompose the max operation into action selection and action evaluation. This algorithm uses two separated DQNs: one as the primary network for action selection and the other as the target network for action evaluation. Compared with the original deep Q-learning algorithm, this algorithm has the following three improvements.

1. *Feature set:* We determine the state features to feed into the neural networks, utilizing the hierarchical layers of tiled convolution filters to exploit local spatial correlations and allow extraction of high-level features from raw input data [46, 51].
2. *Experience replay mechanism:* The algorithm stores experience tuples $\mathrm{ex}(t) = \langle \mathcal{S}_t, \mathcal{A}_{\mathcal{S}_t}, \mathcal{R}_t, \mathcal{S}_{t+1} \rangle$ in a replay memory pool $M(t) = \{\mathrm{ex}(1), \ldots, \mathrm{ex}(t)\}$. The learning process is performed using random samples from the memory pool rather than directly using consecutive samples,

as in Q-learning. This allows the network to learn efficiently by randomly considering any experience instead of focusing on the immediate experience. The algorithm also breaks down the correlations between observations to achieve better stability.

3. *Target Q-network:* We adopt a second neural network to update the target Q-value. The target network is set to reduce the correlations between the target and estimated Q-values, which can also help improve the stability of the algorithm.

Multiple episodes are implemented during the training phase of the deep double Q-learning algorithm. In each episode, a state is observed, and then the agent selects an action for exploration or exploitation based on the ϵ-greedy strategy. The algorithm prefers exploration at the beginning with a reasonably randomized policy and slowly moves toward the exploitation of a deterministic policy. Next, the system performs the selected action, receives a reward, and observes the next state. The experience tuple is then saved to the replay memory for training in later steps. Random batches of experience are sampled from the replay memory and fed into the neural networks for training. A loss function is formulated between the estimated and target Q values. The algorithm then updates the network parameters by minimizing the loss function at each iteration. The loss function is minimized by a minibatch stochastic gradient descent (SGD) algorithm, which has the advantages of a relatively low computation cost and rapid training speed. The loss function $L(\theta)$ can be expressed as

$$L(\theta) = \mathop{\mathbb{E}}_{\langle \mathcal{S},\mathcal{A}_s,\mathcal{R},\mathcal{A}' \rangle}\left[\left(\bar{y} - \; Q(\mathcal{S},\mathcal{A}_s;\theta)\right)^2\right], \tag{14}$$

where $\bar{y}$ is the target Q-value of the target Q-network and θ denotes the parameter of the training Q-network. The target Q-value is calculated as follows:

$$\bar{y} = \mathcal{R} + \gamma Q\left(\mathcal{S}', \underset{\mathcal{A}_{\mathcal{S}'}}{\operatorname{argmax}}\, Q(\mathcal{S}',\mathcal{A}_{\mathcal{S}'};\theta);\bar{\theta}\right), \tag{15}$$

where γ is the discounting factor, and $\bar{\theta}$ denotes the parameters of the target Q-network. Here, $\bar{\theta}$ is updated after every G steps.

4.2. Centralized User Association Control

In this model, to determine the optimal bias for the S-MECS, the M-MECS receives all information and trains the DQN model based on it. To train the proposed DQN model, the MEC exploits the following states, actions, and reward function.

1. States: The state of agent at time t is defined as

$$\mathcal{S}_t = \{\rho_t, m_t, \psi_t, v_t\}, \tag{16}$$

where $\rho_t = \left\{\rho_t^1, \cdots, \rho_t^{|\mathbb{S}|}\right\}$ denotes a vector of the number of UEs assigned to all S-MECSs at time t, and ρ_t^n consists of the number of UEs assigned to S-MECS n; m_t denotes the number of UEs assigned to the M-MECS at time t; $\psi_t = \left\{\psi_t^1, \cdots, \psi_t^{|\mathbb{S}|}\right\}$ denotes a vector of the average throughput for edge users in all S-MECSs at time t; v_t denotes the number of edge users at time t. In other words, when the number of directional beams is $|\mathbb{D}|$,

$$v_t = \begin{bmatrix} v_t^{1_1} & \cdots & v_t^{1_{|\mathbb{D}|}} \\ \vdots & \ddots & \vdots \\ v_t^{S_1} & \cdots & v_t^{S_{|\mathbb{D}|}} \end{bmatrix}, \tag{17}$$

where $v_t^{s_i}$ denotes the number of edge users in the directional beam i of the S-MECS s at time t.

Algorithm 1 Centralized User Association Control with DQN

```
Initialize replay memory $D$
Initialize $Q$ with random parameter $\theta$
Initialize $Q^{'}$ with $\bar{\theta} = \theta$
Initialize $\eta = 0$ and $\epsilon = 0.08$
while $\eta < 5000$ do
    Initialize sequence $\mathcal{S}_1 = \{\rho_t, m_t, \psi_t, v_t\}$
    Initialize preprocessed sequence $\phi_1 = \phi(\mathcal{S}_1)$
    Initialize $t = 0$
    while $t < T$ do
        Receive status reports from all S-MECSs and MUEs
        Sample $z$ uniformly between 0 and 1
        if $z < \epsilon$ then
            Select a random action $\mathcal{A}_t$
        else
            Select $\mathcal{A}_t = \max_{\mathcal{A}}(\phi(\mathcal{S}_t), \mathcal{A}; \theta)$
        end if
        Execute $\mathcal{A}_t = \{\mathcal{B}_t^1, \cdots, \mathcal{B}_t^{|\mathbb{S}|}\}$
        Set a reward $\mathcal{R}_t = \mathbf{U_B}(t) + w_3\mathbf{S}(\Psi_\mathbf{B}(t) - \psi_{min})$
        Set $\mathcal{S}_{t+1}$ by $\mathcal{S}_t, \mathcal{A}_t, \mathcal{R}_t$
        Preprocess $\phi_{t+1} = \phi(\mathcal{S}_{t+1})$
        Store transition $(\phi_t, \mathcal{A}_t, \mathcal{R}_t, \phi_{t+1})$ in $D$
        if $\eta > 100$ then
            Sample minibatch $\beta$ of transitions from $D$
            while $(\phi_j, \mathcal{A}_j, \mathcal{R}_j, \phi_{j+1})$ in $\beta$ do
                if episode terminates at step $j + 1$ then
                    Set $y_j = \mathcal{R}_j$
                else
                    Set $y_j = \mathcal{R}_j + \gamma \max_{a'} Q^{'}(\phi_{j+1}, \mathcal{A}^{'}; \bar{\theta})$
                end if
            end while
            Perform a SGD on $(y_j - Q(\phi_j, \mathcal{A}_j; \theta))^2$
            Update parameter $\theta$
        end if
        Notify bias values to S-MECSs and MUEs
        $t = t + 1$
    end while
    $\eta = \eta + 1$
end while
```

2. Action: The action of agent at time t is defined as

$$\mathcal{A}_t = \left\{\mathcal{B}_t^1, \cdots, \mathcal{B}_t^{|\mathbb{S}|}\right\}, \tag{18}$$

where $\mathcal{B}_t^n$ denotes a set of bias values for each S-MECS n at time t, and $\mathcal{B}_t^n$ consists of the bias values of the directional beam i the in S-MECS n at time t, $\text{bias}_{n,t}^i$. In other words, $\mathcal{B}_t^n = \left\{\text{bias}_{n,t}^1, \cdots, \text{bias}_{n,t}^{|\mathbb{D}|}\right\}$.

3. Reward function: In (10), we maximize the utility function that consists of the throughput for all UEs and the handover cost for all UEs. On the other hand, we have one constraint for throughput of the edge UEs in (9), Therefore, in order to maximize the utility while satisfying the constraint, we move the constraint into the objective function as if Lagrangian relaxation. In particular, to reflect the minimum rate constraint in (9) as a penalty term in the reward function, we employ a sigmoid function:
$$\mathbf{S}(x) = \frac{1}{1+e^{-x}} - 1, \tag{19}$$
where, $\mathbf{S}(x) \to 0$ if $x > 0$, $\mathbf{S}(x) \to -1$, if $x < 0$. Consequently, the reward function at time is defined as
$$\mathcal{R}_t = \mathbf{U}_{\mathbf{B}}(t) + w_3\mathbf{S}(\Psi_{\mathbf{B}}(t) - \psi_{min}), \tag{20}$$
where $0 < w_3 < 1$ denotes the weight value to adjust the reward (i.e., $w_1+w_2+w_3 = 1$). We assume that the reward $\mathcal{R}_t$ for the action $\mathcal{A}_t$ at the state $\mathcal{S}_t$ can be observed after the M-MECS receives a status report from all S-MECSs.

At time t, the M-MECS observes a state $\mathcal{S}_t$, takes an action $\mathcal{A}_t$, and achieves a reward $\mathcal{R}_t$. The goal is to determine the optimal directional beam bias for all S-MECSs to maximize long-term returns.

$$R^{long} = \max_{a_t} \mathbb{E}\left[\sum_{t=0}^{\mathbf{T}-1} \gamma^t \mathcal{R}_t\right], \tag{21}$$

where γ^t approaches zero when t is sufficiently high. Algorithm 1 shows the proposed DQN learning process. The M-MECS collects information on the current channel status of all S-MECSs and MUEs. Subsequently, it assembles all the information into a system state and processes it to obtain an action based on an ϵ-greedy strategy (lines 11–16 in Algorithm 1). DQN parameter θ is updated after performing gradient descent on the loss of training Q-values with minibatch samples of experience (lines 23–33 in Algorithm 1).

Remark 2. In this work, we applied experience replay and separated Q-targets, and ϵ-exploration to improve the convergence of the proposed DQN. In particular, we discretely set the candidate bias, and there could be a finite number of actions at time t. This can facilitate faster learning and convergence than by continuously adjusting the bias. In the simulations presented in Section 6, we identified that it converges stably to optimal values and policies through several iterations.

4.3. Decentralized User Association Control

To reduce the communication overhead, computational complexity and latency in the centralized control, we propose a distributed user association control, that is, all S-MECSs independently decide their own bias using a small amount of information exchange between the S-MECS and M-MECS. For distributed user association control, we modified the state, action, and reward function.

Algorithm 2 Distributed User Association Control with DQN

```
Initialize replay memory $D_n$
Initialize $Q_n$ with random parameter $\theta_n$
Initialize $Q_n^{'}$ with $\bar{\theta}_n = \theta_n$
Initialize $\eta = 0$ and $\epsilon = 0.08$
while $\eta < 5000$ do
    Initialize sequence $\mathcal{S}_{n,1} = \{\rho_t^n, m_t, \psi_t^n, \mu_t\}$
    Initialize preprocessed sequence $\phi_{n,1} = \phi(\mathcal{S}_{n,1})$
    Initialize $t = 0$
    while $t < T$ do
        Send average throughput information to M-MECS
        Receive $\mu_t$ from M-MECS
        Sample $z$ uniformly between 0 and 1
        if $z < \epsilon$ then
            Select a random action $\mathcal{A}_{n,t}$
        else
            Select $\mathcal{A}_{n,t} = \max_{\mathcal{A}}(\phi(\mathcal{S}_{n,t}), \mathcal{A}; \theta_n)$
        end if
        Execute $\mathcal{A}_{n,t} = \left\{\text{bias}_{n,t}^1, \cdots, \text{bias}_{n,t}^{|\mathbb{D}|}\right\}$
        Set a reward $\mathcal{R}_{n,t} = \mathbf{U}_{\mathbf{B}}^n(t) + w_3\mathbf{S}(\Psi_{\mathbf{B}}^n(t) - \psi_{min})$
        Set $\mathcal{S}_{n,t+1}$ by $\mathcal{S}_{n,t}, \mathcal{A}_{n,t}, \mathcal{R}_{n,t}$
        Preprocess $\phi_{n,t+1} = \phi(\mathcal{S}_{n,t+1})$
        Store transition $(\phi_{n,t}, \mathcal{A}_{n,t}, \mathcal{R}_{n,t}, \phi_{n,t+1})$ in $D_n$
        if $\eta > 100$ then
            Sample minibatch of transitions $\beta$ from $D_n$
            while $(\phi_{n,j}, \mathcal{A}_{n,j}, \mathcal{R}_{n,j}, \phi_{n,j+1})$ in $\beta$ do
                if episode terminates at step $j + 1$ then
                    Set $y_{n,j} = \mathcal{R}_{n,j}$
                else
                    Set $y_{n,j} = \mathcal{R}_{n,j} + \gamma \max_{a'} Q^{'}(\phi_{n,j+1},$
                    $\mathcal{A}^{'}; \bar{\theta}_n)$
                end if
            end while
            Perform a SGD on $(y_{n,j} - Q(\phi_{n,j}, \mathcal{A}_{n,j}; \theta_n))^2$
            Update parameter $\theta_n$
        end if
        Send bias values to M-MECS
        Receive bias values of other S-MECSs
        Notify bias values to SUEs
        $t = t + 1$
    end while
    $\eta = \eta + 1$
end while
```

1. States: The state of S-MECS n at time t is defined as
$$\mathcal{S}_{n,t} = \{\rho_t^n, m_t, \psi_t^n, \mu_t\}, \tag{22}$$
where ρ_t^n denotes a vector that consists of the number of UE assigned to the S-MECS n; ψ_t^n denotes the average throughput for edge users in the S-MECS n at time t; μ_t

denotes the average throughput for all UEs at time t. This μ_t information is obtained from the M-MECS which calculates the average throughput for all UEs by receiving information from each S-MECS regarding the number of users belonging to it and the average throughput. In fact, this information μ_t implies that how a certain bias control by an S-MECS affects the system. Therefore, by exchange of this information, each S-MECS's bias control can be stabilized in a distributed environment.

2. Action: The action of S-MECS n at time t is defined as

$$\mathcal{A}_{n,t} = \left\{\text{bias}^{1}_{n,t}, \cdots, \text{bias}^{|\mathbb{D}|}_{n,t}\right\}, \tag{23}$$

   where $\text{bias}^{d}_{n,t}$ denotes the bias value of the directional beam d in the S-MECS n at time t.

3. Reward function: The proposed reward function for S-MECS n at time t is defined as

$$\mathcal{R}_{n,t} = \mathbf{U}^{n}_{\mathbf{B}}(t) + w_3\mathbf{S}(\Psi^{n}_{\mathbf{B}}(t) - \psi_{min}), \tag{24}$$

   where $\mathbf{U}^{n}_{\mathbf{B}}(t)$ and $\Psi^{n}_{\mathbf{B}}(t)$ denote the utility function for all UEs allocated the S-MECS n and the edge throughput for all UEs allocated the S-MECS n.

Algorithm 2 shows the proposed DQN learning process for distributed user association control. Unlike Algorithm 1, each S-MECS is responsible for the DQN leaning instead of the M-MECS, respectively.

5. Operation Procedure and Complexity

5.1. Centralized Operation Procedure

In the centralized control model, all UEs report their measured information to their MECS periodically through a radio resource control (RRC) message. After collecting the status reports from all S-MECSs and MUEs, the M-MECS runs the proposed global DQN learning module to determine the bias values for all beams in the system. The newly determined bias values are distributed to all S-MECSs. Again, the S-MECSs notify their own bias values to their SUEs. When the UE detects a stronger pilot signal with the bias value from neighboring MECSs lasting for a given period, the UE immediately reports to its serving MECS. To conduct the handover, the serving MECS sends a request to the target MECS so that a new downlink resource can be allocated between the UE and target MECS. The serving MECS also transfers the status of the UE to the target MECS for continuous data communication between the UE and target MECS. The overall operation procedure for centralized bias control and handover is shown in Fig 3.

Remark 3. Several reports from UEs and S-MECSs could effect on the end-to-end latency, especially when connection is weak, in the network. Therefore, to reduce the latency, (i) some data compression can be applied between M-MECS, S-MECS, and UE, (ii) only S-MECSs or UEs having a large change in state value or having a good communication connection may report their state information, i.e., S-MECSs and UEs update their state information asynchronously, and (iii) a distributed version, which performs the bias control at the local S-MECS (handover decisions and procedures are performed by the M-MECS), of the previously proposed central algorithm can be applied.

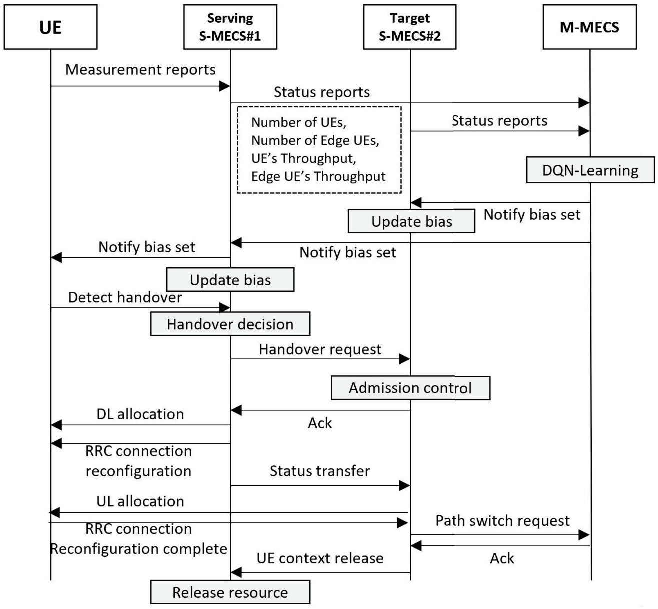


Figure 3: Operation procedures for the bias control and handover in centralized model

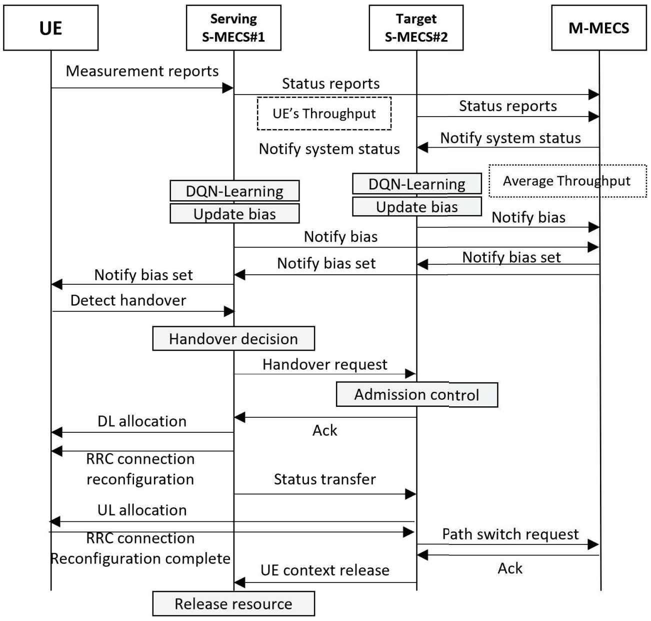


Figure 4: Operation procedures for the bias control and handover in decentralized model

5.2. Decentralized Operation Procedure

In the distributed control model, all UEs periodically report their measured information to the MECS via an RRC message. Unlike the centralized model, the M-MECS sends a global status information, such as overall system throughput to the S-MECSs. After receiving the global and local information from the M-MECS and its serving SUEs, respectively, each S-MECS determines its bias values for its beams using the local DQN learning module. Subsequently, the S-MECSs share their determined bias values with the other S-MECSs through the M-MECS. Then, all UEs can know the bias values of all neighboring MECSs through MECS's broadcast procedure. After that, the user association procedure is the same as the centralized model. The overall operation procedure for distributed bias control and handover is shown in Fig 4.

5.3. *Communication Overhead, Computational Complexity, End-to-End Latency, and Scalability*

For the analysis, we denote the number of S-MECS, number of beams per S-MECS, and total number of UE as $|\mathbb{S}|$, $|\mathbb{D}|$, and $|\mathbb{U}|$, respectively. The number of information bits representing UE ID, beam ID, K discrete bias level, T discrete throughput level, and handover signallings are $\log_2(|\mathbb{U}|)$, $\log_2(|\mathbb{D}|)$, $\log_2(K)$, $\log_2(T)$, and H bits, respectively.

5.3.1. *Communication overhead*

Using the centralized method, each UE sends the following information to its S-MECS: user-id, cell-beam-id, and DQN state information. That is, the total amount of uplink information bits in access links is $|\mathbb{U}| * (\log_2(|\mathbb{U}|) + \log_2(|\mathbb{D}|) + \log_2(T))$ bits. Then, the S-MECSs deliver the collected information to its M-MECS. Therefore, the total amount of uplink information bits in fronthaul links is also $|\mathbb{U}| * (\log_2(|\mathbb{U}|) + \log_2(|\mathbb{D}|) + \log_2(T))$ bits. After calculating the optimal bias, the M-MECS sends the calculated bias information to the S-MECSs, which amounts to $|\mathbb{S}||\mathbb{D}| \log_2(K)$ bits. Then, each S-MECS sends its bias information to its serving SUEs. Moreover, whenever the handover decision is made, the S-MECS sends some handover control messages to the corresponding SUE. These downlink information bits in access links amounts to $|\mathbb{S}||\mathbb{D}| \log_2(K) + \log_2(|\mathbb{S}||\mathbb{D}|) + H$ bits.

On the other hand, with the decentralized method, each UE sends the same information to its S-MECS: user-id, cell-beam-id, and DQN state information. That is, the amount of information uplink bits in access links is $|\mathbb{U}| * (\log_2(|\mathbb{U}|) + \log_2(|\mathbb{D}|) + \log_2(T))$ bits. In the process of sharing the related information for the distributed DQN algorithm among S-MECSs through the M-MECS, the amount of uplink and downlink information bits in fronthaul links is both $|\mathbb{S}||\mathbb{D}| \log_2(K) + |\mathbb{S}| \log_2(T)$ bits. Lastly, each S-MECS sends its determined bias information and handover control messages to its serving SUEs through the access links. These downlink information bits in access links corresponds to $|\mathbb{S}||\mathbb{D}| \log_2(K) + \log_2(|\mathbb{S}||\mathbb{D}|) + H$ bits. These communication overheads are summarized as Table 3 and 4.

Table 3: Communication overhead (Uplink)

Mode	**Fronthaul** **(S/M-MECS)**	**Access** **(UE/S-MECS)**
Proposed (Centralized)	$|\mathbb{U}| \log_2(|\mathbb{U}||\mathbb{D}|T)$	$|\mathbb{U}| \log_2(|\mathbb{U}|\mathbb{D}|T)$
Proposed (Distributed)	$|\mathbb{S}||\mathbb{D}| \log_2(K) + |\mathbb{S}| \log_2(T)$	$|\mathbb{U}| \log_2(|\mathbb{U}||\mathbb{D}|T)$

Table 4: Communication overhead (Downlink)

Mode	**Fronthaul** **(M/S-MECS)**	**Access** **(S-MECS/UE)**
Proposed (Centralized)	$|\mathbb{S}||\mathbb{D}| \log_2(K)$	$|\mathbb{S}||\mathbb{D}| \log_2(K)+\log_2(|\mathbb{S}||\mathbb{D}|)+H$
Proposed (Distributed)	$|\mathbb{S}||\mathbb{D}| \log_2(K) + |\mathbb{S}| \log_2(T)$	$|\mathbb{S}||\mathbb{D}| \log_2(K)+\log_2(|\mathbb{S}||\mathbb{D}|)+H$

5.3.2. *Computational complexity*

We analyzed main computation's complexity: offline DQN learning and online DQN decision. First, regarding the centralized control, DQN-based offline learning and online decision are carried out at the M-MECS. Regarding DQN learning, the input and output size of DQN is $(|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}| + |\mathbb{S}||\mathbb{D}|)$, and $|\mathbb{S}||\mathbb{D}|K$, respectively. Considering the sizes of the input, output, and hidden layers and ReLU activation in the three hidden layers, the complexity of selecting a strategic action using a DQN is $O\Big(H * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}|) + 2H^2 + H|\mathbb{S}||\mathbb{D}|K + 2H\Big)$ with H denoting the size of each hidden layer, which can be simplified to $O(H * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}| + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$. For the learning procedure, each learning step is implemented over C samples; the DRL agent performs gradient descent on the Q-value loss, and utilizes both primary and target DQNs to determine the Q-value loss. Therefore, the complexity of each learning step becomes $O(2CH * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}| + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$. On the other hand, regarding online DQN decisions, the output action is obtained using the primary DQN as $x_k^*(t) = \arg\max_{\mathcal{A}_k} Q_k(\mathcal{S}_k(t), \mathcal{A}_k, \theta_k^*)$. Therefore, the complexity of an online decision is as follows: $O(H * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}| + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$. That is, in the centralized approach, the DQN learning complexity for each learning step and DQN decision complexity become $O(2CH * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}| + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$ and $O(H * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}| + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$, respectively. Second, regarding the distributed approach, the DQN learning and decision process are carried out in parallel at each S-MECSs. In other words, the DQN learning and decision complexities become $O(2CH * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}|_i + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$ and $O(H * (2|\mathbb{S}||\mathbb{D}| + |\mathbb{D}| + |\mathbb{U}|_i + 2H + |\mathbb{S}||\mathbb{D}|K + 2))$, respectively. These computation complexity are summarized as Table 5.

5.3.3. *End-to-end latency*

The latency in 5G can be classified into two types: control latency and data latency, where control latency refers to the time taken to change from the idle state to the active state, and data latency refers to the time taken for data to arrive from the user IP layer to the base station IP layer (one-way). In particular, ultra-reliable low latency communications (URLLC) services in 5G demand 1 msec and 10 msec latency for the data

Table 5: Computational complexity (DQN learning and decision)

Mode	M-MECS	S-MECS
Proposed (Centralized)	DQN learning $O\,(2CH * (2\lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert + \lvert\mathbb{D}\rvert + \lvert\mathbb{U}\rvert + 2H + \lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert K + 2))$	No DQN learning and decision
	DQN decision $O\,(H * (2\lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert + \lvert\mathbb{D}\rvert + \lvert\mathbb{U}\rvert + 2H + \lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert K + 2))$	
Proposed (Distributed)	No bias control at M-MECS (i.e., bias=0 for M-MECS)	DQN learning $O\,(2CH * (2\lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert + \lvert\mathbb{D}\rvert + \lvert\mathbb{U}\rvert_i + 2H + \lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert K + 2))$
		DQN decision $O\,(H * (2\lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert + \lvert\mathbb{D}\rvert + \lvert\mathbb{U}\rvert_i + 2H + \lvert\mathbb{S}\rvert\lvert\mathbb{D}\rvert K + 2))$

and control plane, respectively. On the other hand, in 5G based MEC system, the end-to-end latency (data or control) mainly depends on over-the-air transmission delay (including retransmission delay) τ_t, computational load processing delay τ_p, data holding delay during handover τ_h, caching content retrieval delay τ_c, and control decision delay τ_{cl}. That is, the end-to-end latency for a UE m served by a MECS n can be defined as a combination of

$$\tau(m;n) = \{\tau_t(m;n), \tau_p(m;n), \tau_h(m;n), \tau_c(m;n), \tau_{cl}(m;n)\}.$$

The end-to-end latency is very challenging, owing to the stochastic effects of the wireless fading channel, noise, random queuing delays at the transmitters and edge computing servers, packet retransmissions, and heterogeneity of processing tasks and edge computing resources.

Nevertheless, by using mm-HMEC and the proposed control, we can expect a reduction in the end-to-end latency because of 1) wider transmission bandwidth of the mmWave frequency band in fronthaul links and access links, enhanced spatial frequency reuse by mmWave's short coverage, lower interference by mmWave's directional transmission (i.e., reduced over-the-air round-trip transmission latency, (τ_t)), 2) load balancing using the proposed bias control (i.e, reduced load computing latency, (τ_p)), 3) reduced the number of handovers by the proposed bias control (i.e,, reduced handover latency, (τ_h)), 4) closer caching contents retrieval site using the S-MECS (i.e, reduced caching latency, (τ_c)), and 5) faster control decisions using the proposed DQN-based real-time decision (i.e, reduced control decision latency, (τ_{cl})).

Moreover, if we apply the distributed decision algorithm, that is, control decision by S-MECS instead of M-MECS, the transmission delay (τ_t) and control decision delay (τ_{cl}) will have no difference from the delays in the centralized method. On the other hand, conversely, the distributed approach could bring a increased number of handover (i.e., increased handover latency (τ_h)) and increased load computing latency (τ_p) for some UEs because of the suboptimality of the distributed control. However, the handover latency (τ_h) is expected to be zero in 5G [52], and the proposed distributed algorithm provides near-centralized performance (which is shown through simulations) so that the increased latency in load computing will be negligible. That is, we can expect that the distributed approach will also provide almost the same latency when it is compared to the centralized decision approach.

5.3.4. *Scalability*

the communication overhead of the proposed centralized control is a polynomial overhead $O(|\mathbb{U}| \log |\mathbb{U}|)$ with respect to the total number of UEs $|\mathbb{U}|$ in uplink, and it is also a polynomial overhead $O(|\mathbb{S}|)$ with respect to the total number of S-MECSs in downlink. Second, the computation complexity of the proposed centralized learning and decision algorithm is polynomial with respect to the number of S-MECSs and UEs. Third, the proposed centralized control can provide enhanced end-to-end latency with the help of the characteristics of mmWave transmission the and hierarchical MEC architecture. Fourth, by assigning DQN-based control to the S-MECSs in a distributed manner, the communication overhead can increase a little bit only in the fronthaul link; the computation complexity can be reduced owing to the parallel processing by the multiple S-MECSs that are located closer to the UEs. However, the increased communication overhead in the fronthaul link can be efficiently controlled by the approaches such as (i) asynchronous or partial status information report and (ii) compressed sensing, and this communication overhead disappears at the end of the learning period; the end-to-end latency is almost same as that in the centralized method. Finally, in an environment where the number of UEs increases infinitely, by linearly increasing the number of S-MECSs (that is, keeping the average number of UEs constant for each S-MECS), we can expect that communication overhead, computing complexity, and end-to-end latency can be maintained at least polynomially constant while increasing overall system utility [53]. That is, the proposed network model and control schemes are feasible and scalable for various applications in practical industrial systems.

6. Performance Evaluation

6.1. *Network Models*

To evaluate the proposed scheme, we implemented a system simulator using PyTorch libraries [54]. The system simulator considers a 500 m × 500 m network area, in which one M-MECS and four or 10 S-MECSs are deployed. The M-MECS and S-MECSs use a 28 GHz radio frequency. We assume that all UEs generate a packet every second and transmit it to the serving MECS. Considering the actual handover setting in 5G, when the RSRP of the target MECS is 3 dB higher than the RSRP of the serving MECS for 100 ms, the UE requests a handover to the target MECS; that is, the RSRP offset and time-to-trigger parameters are set as 3 dB and 100 ms, respectively.

6.2. DQN Models

For DQN-based machine learning, we set the experience replay buffer size to 5000 and the minibatch size to 32. For the agent to obtain samples, there must be a sufficient transition in the experience replay buffer. Thus, the agent does not learn until 1000 transitions are stored in the experience replay buffer. We used a learning rate of 0.0005, which can prevent the action-value function from being rapidly updated by a new action, and also used a value of 0.98 for γ, the discount factor for return. For more accurate learning, we ran the learning process over 5000 episodes, and all UE were deployed uniformly and randomly in each episode. Table 6 presents the simulation parameters.

Table 6: Simulation parameters

Parameter	Assumption
UE deployment	Random
Number of M-MECS	1
Number of S-MECS	4, 10
Minimum inter-node distance	10 m
Bias	[5, 10, 15]
Transmit power of M-MECS	25 dBm
Transmit power of S-MECS	15 dBm
Operating frequency	28 GHz
Channel bandwidth	1000 MHz
Traffic load	1024 Mbps
A3 Offset	3 dB
Time-to-trigger	100 ms
Learning rate	0.0005
ϵ	0.08
γ	0.98
Buffer size	5000
Batch size	32
ψ_{min}	700, 900
Node mobility	RWM, RDM, MAN

6.3. User Mobility Models

Initially, all UEs are uniformly deployed, and their performances were evaluated using the following different mobility models.

- *Random waypoint mobility model (RWM)*: In this model, the UE pauses the specific time and decides the next destination. Subsequently, a velocity is randomly selected, and the UE moves to the chosen destination with the specified velocity. When the UE arrives at its destination, the above process is repeated [55].

- *Random direction mobility model (RDM)*: This model is proposed to solve the problem of increasing the node density near the center in RWM. The UE pauses a specific time and decides the next destination, which is located at the boundary of the simulation area [56].

- *Manhattan mobility model (MAN)*: The model uses a map that mimics horizontal/vertical roads within a city. Each road has two opposite lanes, and all UEs can only move over these lanes [57].

6.4. Evaluation Results and Discussions

We compared the following schemes in terms of system utility, throughput, and handover.

- Random CRE (R-CRE): In this approach, each S-MECS sets its beam bias values randomly.

- Standard CRE (S-CRE) [41]: As a static CRE scheme, it sets the static bias values to 5, which provides the best performance on average.

- Dynamic CRE (D-CRE) [26]: As a dynamic CRE scheme, it adjusts the bias values depending on resource balance between M-MECS and S-MECSs.
  When the throughput decreases, the bias value increases by 1 dB; otherwise, it decreases by 1 dB.

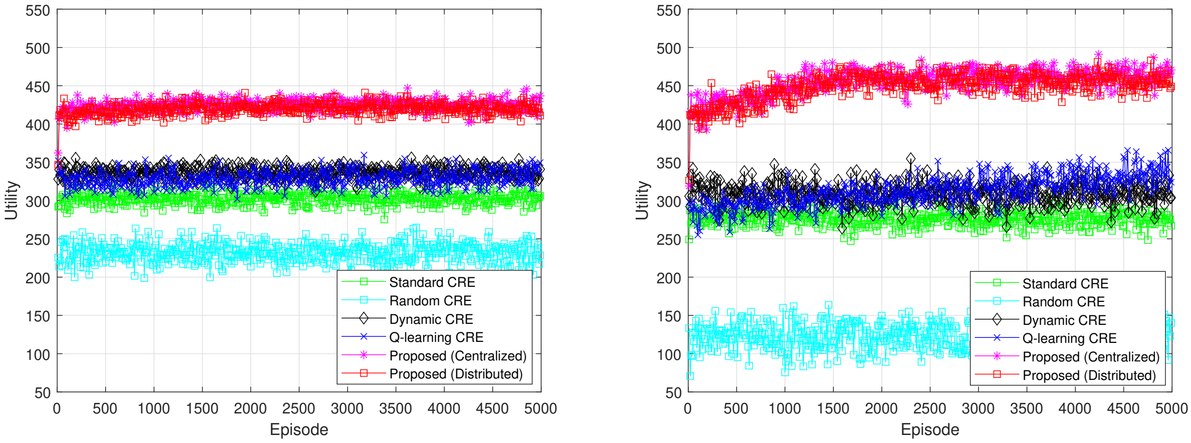
- Q-learning based CRE (Q-CRE) [30]: As a RL-based dynamic CRE scheme, it performs a Q-learning based bias control. All the M/S-MECSs learn their own bias values, respectively.

- Proposed (Centralized): With this scheme, the central M-MECS controls the bias values for all the M/S-MECSs.

- Proposed (Decentralized): With this scheme, all M/S-MECSs independently determine their beam bias values in a distributed manner.

6.4.1. System utility

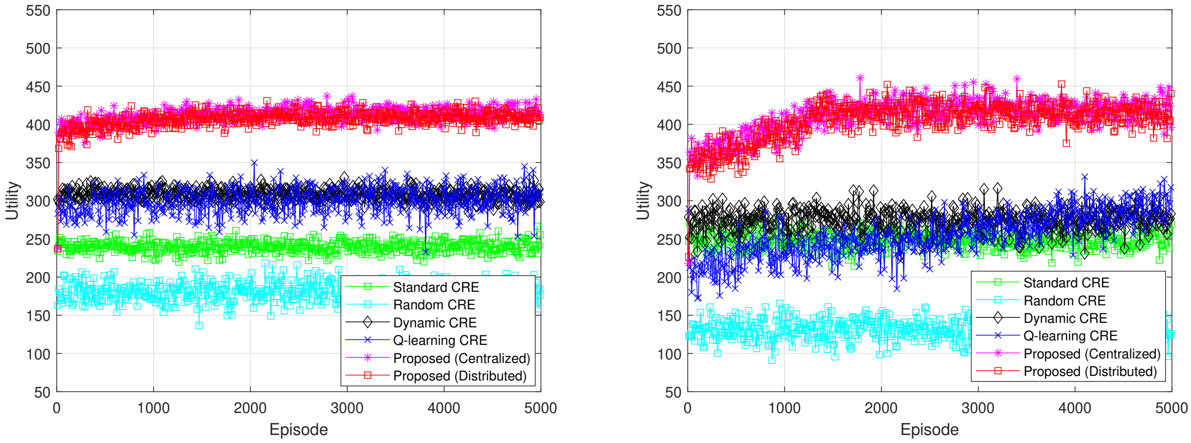
Figs. 5, 6, and 7 show the achieved system utility under RWM, RDM, and MAN mobility model. In Fig. 5(a) for 4 S-MECSs, R-CRE, S-CRE, D-CRE, and Q-CRE have average utilities of 231.18, 303.15, 336.37, and 330.6 for all episodes, respectively. Meanwhile, the proposed centralized and distributed schemes show an average utility of 425.03 and 420.91, respectively. That is, the proposed centralized and distributed control provide 83.85%, 40.20%, 26.35%, and 28.56% and 82.07%, 38.84%, 25.13%, and 27.31% higher utility than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In Fig. 5(b) for 10 S-MECSs, the proposed centralized and distributed control provide 151.37%, 54.68%, 49.49%, and 43.18% and 148.39%, 52.84%, 47.72%, and 41.49% higher utility than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In Fig. 6(a) for 4 S-MECSs, the proposed centralized and distributed control provide 129.65%, 72.65%, 34.81%, and 37.53% and 126.89%, 70.57%, 33.18%, and 35.87% higher utility than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In Fig. 6(b) for 10 S-MECSs, the proposed centralized and distributed control provide 221.38%, 69.25%, 53.49%, and 56.38% and 216.72%, 66.79%, 51.26%, and 54.11% higher



(a) 4 S-MECS

(b) 10 S-MECS

Figure 5: The system utility under Random waypoint mobility model.



(a) 4 S-MECS

(b) 10 S-MECS

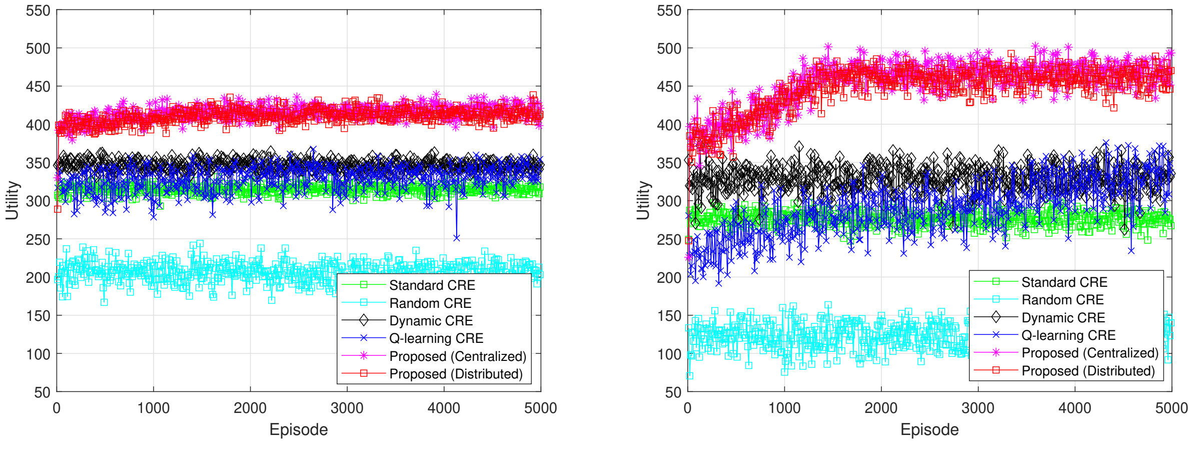
Figure 6: The system utility under Random direction mobility model.

utility than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. Finally, in Fig. 7(a) for 4 S-MECSs, the proposed centralized and distributed control provide 103.5%, 32.35%, 20.85%, and 26.62% and 101.51%, 31.05%, 19.66%, and 24.39% higher utility than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In Fig. 7(b) for 10 S-MECSs, the proposed centralized and distributed control provide 277.44%, 69.28%, 42.38%, and 52.74% and 271.4%, 66.57%, 40.1%, and 50.29% higher utility than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively.

That is, as the average of all moving models, the proposed centralized and distributed control provide 105.67%, 48.4%, 27.33%, and 30.57% and 103.49%, 46.82%, 25.99%, and 29.19% enhanced utility (with 4 S-MECSs), and 216.73%, 64.4%, 48.45%, and 50.77% and 212.17%, 62.07%, 46.36%, and 48.63% enhanced utility (with 10 S-MECSs) when compared to R-CRE, S-CRE, D-CRE and Q-CRE, respectively. We can also see that the proposed distributed control achieves almost the same performance as the proposed centralized control, and the system utility increases as the number of S-MECS increases.

6.4.2. Throughput

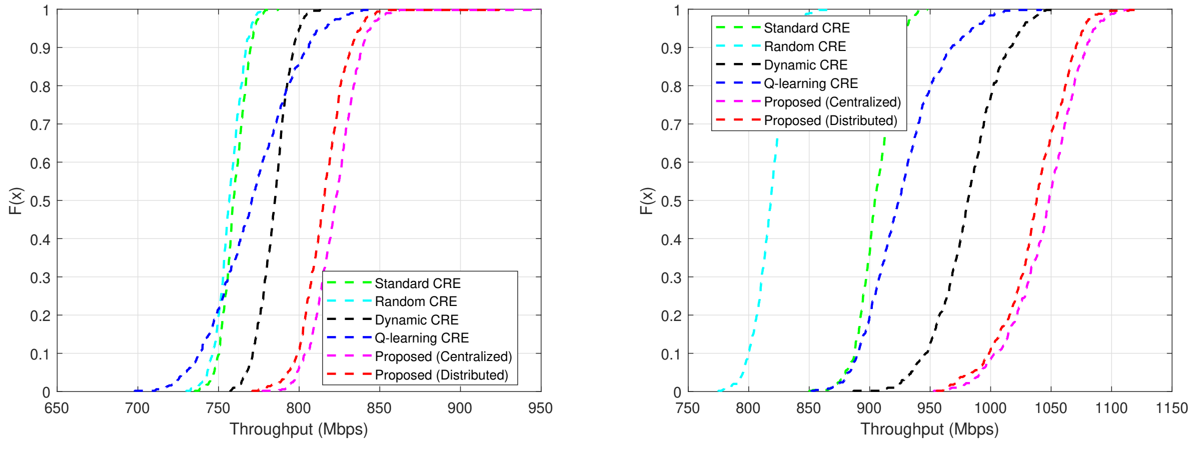
Fig. 8, 9, and 10 compare cumulative distribution functions (CDF) under RWM, RDM, and MAN mobility model, respectively. In Fig. 8(a) with 4 S-MECSs, the proposed centralized and distributed schemes provide 8.5%, 8.23%, 4.84%, and 6.65% and 7.72%, 7.36%, 3.99%, and 5.79% higher average cell throughput than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. On the other hand, they provide 60.25%, 63.34%, 42.31%, and 51.53% and 58.61%, 61.66%, 40.85%, and 49.98% higher average edge throughput than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In Fig. 8(b) with 10 S-MECSs, the proposed centralized and distributed approaches give 33.92%, 16.49%, 7.66%, and 12.48% and 32.75%, 15.47%, 6.72%, and 11.50% higher average cell throughput than R-CRE, S-CRE, D-CRE, and Q-CRE. Also, they give 32.23%, 30.74%, 85%, and 27.41% and 30.88%, 29.40%, 83.10%, and 25.1% higher average edge throughput



(a) 4 S-MECS

(b) 10 S-MECS

Figure 7: The system utility under Manhattan mobility model.



(a) 4 S-MECS

(b) 10 S-MECS

Figure 8: The cumulative distribution function of throughput under Random waypoint mobility model.

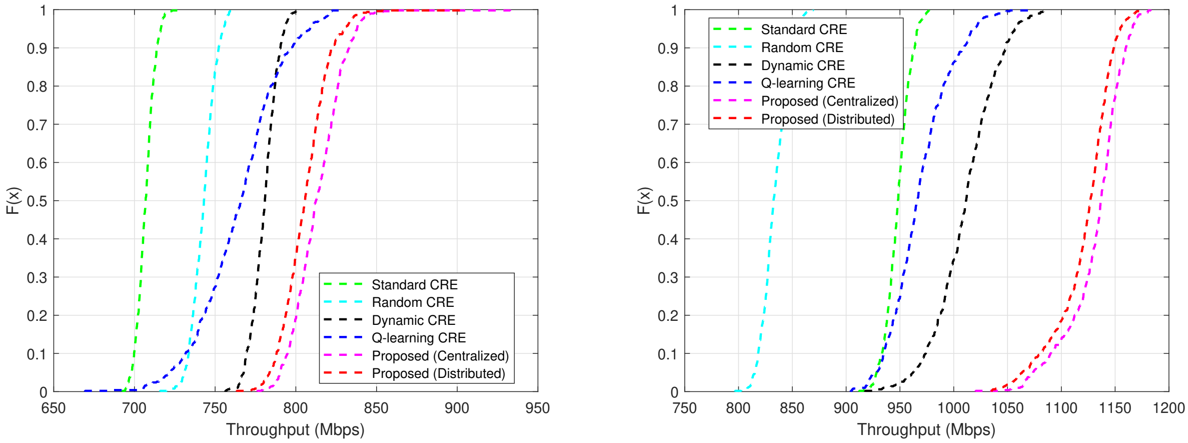
than R-CRE, S-CRE, D-CRE, and Q-CRE. However, for the other mobility models in Fig. 9 and Fig. 10, the proposed centralized and distributed schemes also provide 22.59%, 14.93%, 7.5%, and 10.49% and 21.44%, 13.85%, 6.49%, and 9.46% higher average cell throughput, and 35.05%, 51.56%, 53.91%, and 34.41% and 33.88%, 50.26%, 52.58%, and 33.25% higher average edge throughput than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively.

That is, as the average for all mobility models, the proposed centralized and distributed controls provide 10.18%, 10.51%, 4.64%, and 7.21% and 9.19%, 9.51%, 3.7%, and 6.25% enhanced average throughput (with 4 S-MECSs), and 35%, 19.36%, 10.36%, and 13.77% and 33.68%, 18.19%, 9.28%, and 12.66% enhanced average throughput (with 10 S-MECSs) when compared to R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In addition, the proposed centralized and distributed controls provide 54.79%, 63.64%, 42.51%, and 43.18% and 53.43%, 62.21%, 41.28%, and 41.92% enhanced edge throughput (with 4 S-MECSs), and 15.32%, 39.48%, 65.30%, and 25.64% and 14.32%, 38.32%, 63.87%, and 24.57% enhanced edge throughput (with 10 S-MECSs) compared to R-CRE, S-CRE, D-CRE, and Q-CRE, respectively.

6.4.3. *Handover*

Fig. 11, 12, and 13 show the total number of handovers for RWM, RDM, and MAN mobility models, respectively. In Fig. 11(a) with 4 S-MECSs, we can observe that the proposed centralized and distributed approaches provide 69.26%, 55.27%, 49.11%, and 48.26% and 69.21%, 55.19%, 49.02%, and 48.17% lower handovers than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. In Fig. 11(b) with 10 S-MECSs, the proposed centralized and distributed approaches provide 40.68%, 30.01%, 38.59%, and 28.68% and 40.7%, 30.03%, 38.6%, and 28.7% lower handovers than R-CRE, S-CRE, D-CRE, and Q-CRE, respectively.

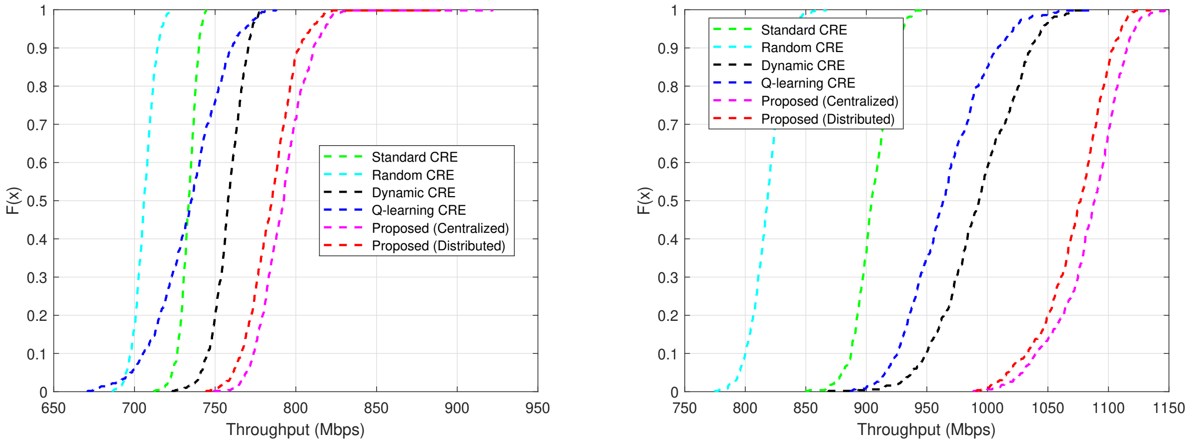
That is, as the average for all mobility models, the proposed centralized and distributed controls provide 71.52%, 56.07%, 49.81%, and 48.82% and 71.51%, 56.06%, 49.81%, and



(a) 4 S-MECS

(b) 10 S-MECS

Figure 9: The cumulative distribution function of throughput under Random direction mobility mode.



(a) 4 S-MECS

(b) 10 S-MECS

Figure 10: The cumulative distribution function of throughput under Manhattan mobility model.

48.81% reduced handover cost (with 4 S-MECSs), and 39.07%, 25.4%, 29.31%, and 25.93% and 39.09%, 25.42%, 29.32%, and 25.95% reduced handover cost (with 10 S-MECSs) compared to R-CRE, S-CRE, D-CRE, and Q-CRE, respectively.

7. Conclusion

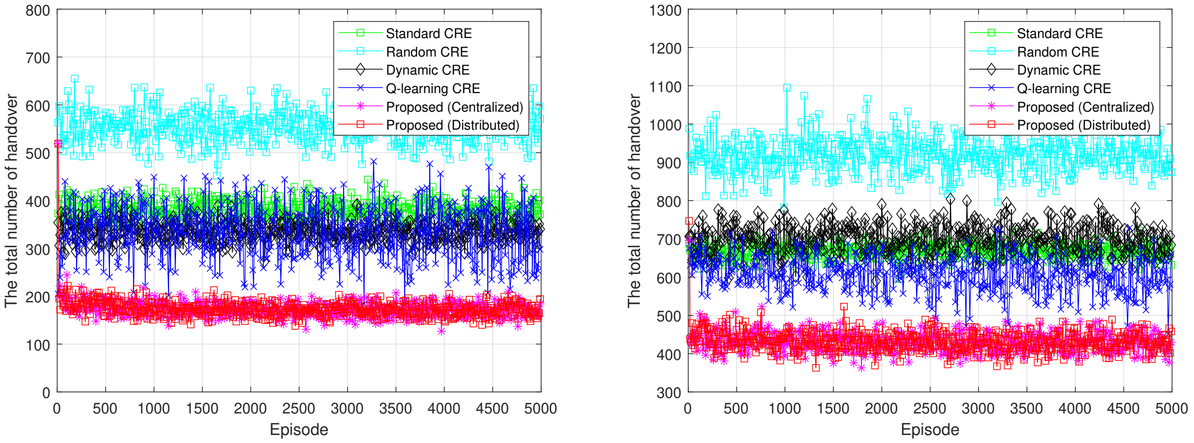
In this paper, we proposed centralized and distributed novel user association control schemes that can maximize UE's utility associated with the system throughput, edge throughput, and handover cost in mm-HMEC systems. To develop the proposed schemes, we first formulated a non-convex MIP programming model and then converted it into DQN-based range-expansion bias control models that can operate in real time. We analyzed that the proposed schemes can provide polynomial communication overhead and computation complexity. Moreover, for various user mobility models such as RWM, RDM, and MAN, the simulation confirmed that the proposed centralized and distributed schemes provide enhanced average throughput, edge throughput, and reduced handover cost when compared to benchmark schemes of R-CRE, S-CRE, D-CRE, and Q-CRE, respectively. As a future work, we will explore a low-complexity joint user association, beamforming control and power control that minimizes end-to-end latency for industrial IoT services in mm-HMEC based 6G system, and quantitatively evaluate the overall end-to-end latency and complexity using real 6G system simulator.

References

[1] Y. Niu, Y. Li, D. Jin, L. Su, A. V. Vasilakos, A survey of millimeter wave communications (mmwave) for 5g: opportunities and challenges, Wireless networks 21 (8) (2015) 2657–2676.

[2] J. Wang, J. Weitzen, O. Bayat, V. Sevindik, M. Li, Interference coordination for millimeter wave communications in 5g networks for performance optimization, EURASIP Journal on Wireless Communications and Networking 2019 (1) (2019) 46.

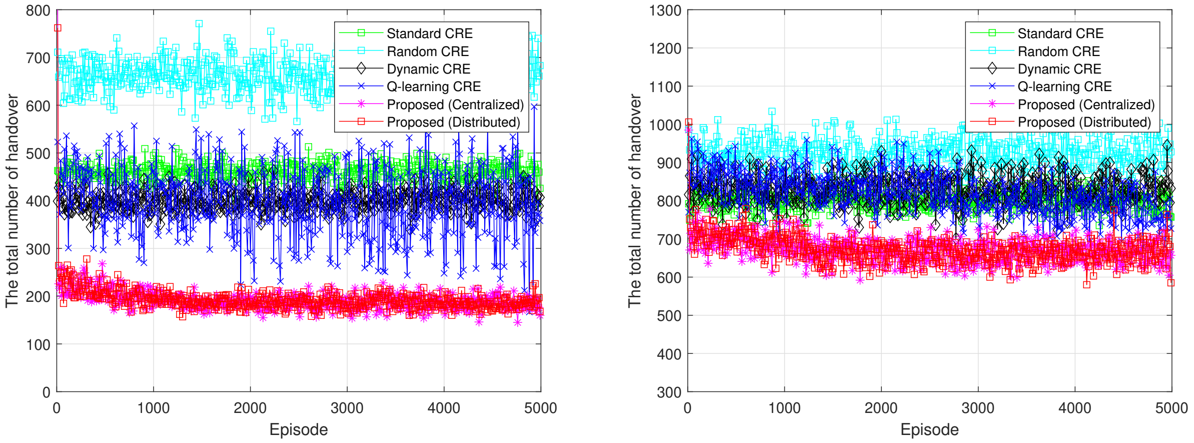
[3] F. Firyaguna, J. Kibilda, C. Galiotto, N. Marchetti, Performance analysis of indoor mmwave networks with ceiling-mounted access points, IEEE Transactions on Mobile Computing 20 (5) (2020) 1940–1950.



(a) 4 S-MECS

(b) 10 S-MECS

Figure 11: The number of handover under Random waypoint mobility model.



(a) 4 S-MECS

(b) 10 S-MECS

Figure 12: The number of handover under Random direction mobility model.

[4] D. Peron, M. Giordani, M. Zorzi, An efficient requirement-aware attachment policy for future millimeter wave vehicular networks, in: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019, pp. 2435–2442.

[5] M. Hadi, R. Ghazizadeh, Joint sub-carrier allocation and 3d beamforming design in oma-noma based mmwave heterogeneous networks under channel uncertainties, AEU-International Journal of Electronics and Communications 137 (2021) 153809.

[6] B. Zhai, A. Tang, C. Huang, C. Han, X. Wang, Antenna subarray management for hybrid beamforming in millimeter-wave mesh backhaul networks, Nano Communication Networks 19 (2019) 92–101.

[7] D. Castanheira, P. Lopes, A. Silva, A. Gameiro, Hybrid beamforming designs for massive mimo millimeter-wave heterogeneous systems, IEEE Access 5 (2017) 21806–21817.

[8] N. W. Sung, Y. S. Choi, Contention based fast beam switching scheme in millimeter-wave cellular systems, in: 2015 17th International Conference on Advanced Communication Technology (ICACT), IEEE, 2015, pp. 521–524.

[9] W. Hao, M. Zeng, G. Sun, P. Xiao, Edge cache-assisted secure low-latency millimeter-wave transmission, IEEE Internet of Things Journal 7 (3) (2019) 1815–1825.

[10] A. M. Nor, E. M. Mohamed, Millimeter wave beamforming training based on li-fi localization in indoor environment, in: GLOBECOM 2017-2017 IEEE Global Communications Conference, IEEE, 2017, pp. 1–6.

[11] A. M. Nor, E. M. Mohamed, Li-fi positioning for efficient millimeter wave beamforming training in indoor environment, Mobile Networks and Applications 24 (2) (2019) 517–531.

[12] A. S. Mubarak, E. M. Mohamed, H. Esmaiel, Millimeter wave beamforming training, discovery and association using wifi positioning in outdoor urban environment, in: 2016 28th International Conference on Microelectronics (ICM), IEEE, 2016, pp. 221–224.

[13] L. Zhou, Y. Ohashi, Fast codebook-based beamforming training for mmwave mimo systems with subarray structures, in: 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), IEEE, 2015, pp. 1–5.

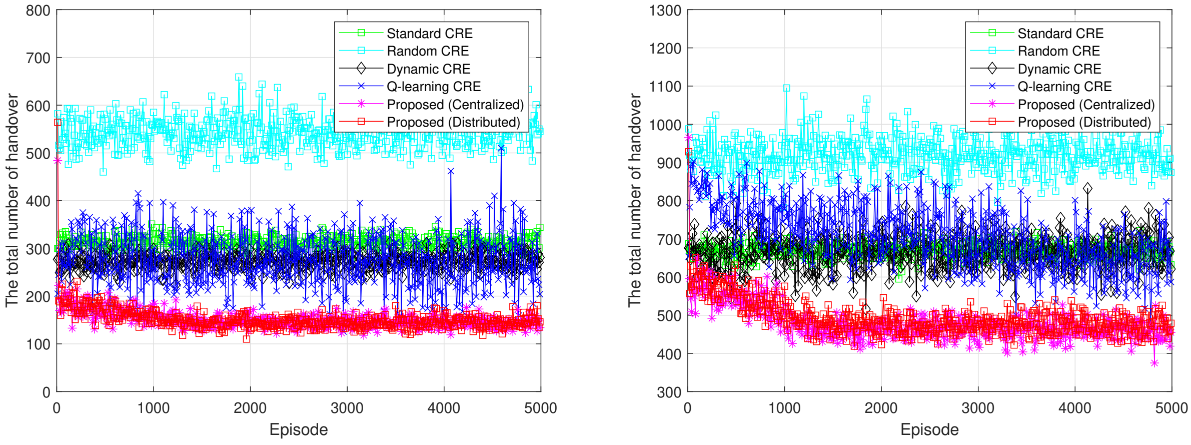
[14] W. Wang, W. Zhang, Joint beam training and positioning for intelligent reflecting surfaces assisted millimeter wave communications, IEEE Transactions on Wireless Communications 20 (10) (2021) 6282–6297.

[15] W. Na, Y. Lee, N.-N. Dao, D. N. Vu, A. Masood, S. Cho, Directional link scheduling for real-time data processing in smart manufacturing system, IEEE Internet of Things Journal 5 (5) (2018) 3661–3671.

[16] W. Ding, Y. Niu, H. Wu, Y. Li, Z. Zhong, Qos-aware full-duplex concurrent scheduling for millimeter wave wireless backhaul networks, IEEE Access 6 (2018) 25313–25322.

[17] O. Semiari, W. Saad, M. Bennis, Z. Dawy, Inter-operator resource management for millimeter wave multi-hop backhaul networks, IEEE Transactions on Wireless Communications 16 (8) (2017) 5258–5272.

[18] Y. Li, J. Luo, R. A. Stirling-Gallacher, G. Caire, Integrated access and backhaul optimization for millimeter wave heterogeneous networks, arXiv preprint arXiv:1901.04959 (2019).



(a) 4 S-MECS

(b) 10 S-MECS

Figure 13: The number of handover under Manhattan mobility model.

[19] G. Yang, M. Xiao, M. Alam, Y. Huang, Low-latency heterogeneous networks with millimeter-wave communications, IEEE Communications Magazine 56 (6) (2018) 124–129.

[20] X. Jia, W. Xu, Y. Chen, L. Yang, Hybrid self-backhaul and cache assisted millimeter wave two-tier heterogeneous networks with mimo equipped backhaul access points, IEEE Access 7 (2019) 59963–59983.

[21] W. Ren, J. Xu, D. Li, Q. Cui, X. Tao, A robust inter beam handover scheme for 5g mmwave mobile communication system in hsr scenario, in: 2019 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2019, pp. 1–6.

[22] M. Mezzavilla, S. Goyal, S. Panwar, S. Rangan, M. Zorzi, An mdp model for optimal handover decisions in mmwave cellular networks, in: 2016 European conference on networks and communications (EuCNC), IEEE, 2016, pp. 100–105.

[23] F. Guidolin, I. Pappalardo, A. Zanella, M. Zorzi, Context-aware handover policies in hetnets, IEEE Transactions on Wireless Communications 15 (3) (2016) 1895–1906.

[24] S. Zang, W. Bao, P. L. Yeoh, H. Chen, Z. Lin, B. Vucetic, Y. Li, Mobility handover optimization in millimeter wave heterogeneous networks, in: 2017 17th International symposium on communications and information technologies (ISCIT), IEEE, 2017, pp. 1–6.

[25] S. Sadr, R. S. Adve, Handoff rate and coverage analysis in multi-tier heterogeneous networks, IEEE Transactions on Wireless Communications 14 (5) (2015) 2626–2638.

[26] M. Al-Rawi, A dynamic approach for cell range expansion in interference coordinated lte-advanced heterogeneous networks, in: 2012 IEEE International Conference on Communication Systems (ICCS), IEEE, 2012, pp. 533–537.

[27] T. M. Shami, D. Grace, A. Burr, J. S. Vardakas, Load balancing and control with interference mitigation in 5g heterogeneous networks, EURASIP Journal on Wireless Communications and Networking 2019 (1) (2019) 1–12.

[28] R. Q. Shaddad, A. A. Neda'a, M. O. Alzylai, T. M. Shami, Biased user association in 5g heterogeneous networks, in: 2021 International Conference of Technology, Science and Administration (ICTSA), IEEE, 2021, pp. 1–4.

[29] H. Jiang, System utility optimization of cell range expansion in heterogeneous cellular networks, in: 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), IEEE, 2016, pp. 412–417.

[30] T. Kudo, T. Ohtsuki, Cell range expansion using distributed q-learning in heterogeneous networks, Eurasip journal on wireless communications and networking 2013 (1) (2013) 61.

[31] Y. Zhang, B. Zhang, H. Wang, T. Zhang, Y. Qian, Deep learning-based coordinated beamforming for massive mimo-enabled heterogeneous networks, in: 2021 IEEE Global Communications Conference (GLOBECOM), IEEE, 2021, pp. 1–6.

[32] R. Kim, Y. Kim, N. Y. Yu, S.-J. Kim, H. Lim, Online learning-based downlink transmission coordination in ultra-dense millimeter wave heterogeneous networks, IEEE Transactions on Wireless Communications 18 (4) (2019) 2200–2214.

[33] K. Ryu, W. Kim, Multi-objective optimization of energy saving and throughput in heterogeneous networks using deep reinforcement learning, Sensors 21 (23) (2021) 7925.

[34] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, M. Latva-Aho, Path selection and rate allocation in self-backhauled mmwave networks, in: 2018 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2018, pp. 1–6.

[35] L. Yan, H. Ding, L. Zhang, J. Liu, X. Fang, Y. Fang, M. Xiao, X. Huang, Machine learning-based handovers for sub-6 ghz and mmwave integrated vehicular networks, IEEE Transactions on Wireless Communications 18 (10) (2019) 4873–4885.

[36] L. Sun, J. Hou, T. Shu, Optimal handover policy for mmwave cellular networks: A multi-armed bandit approach, in: 2019 IEEE Global Communications Conference (GLOBECOM), IEEE, 2019, pp. 1–6.

[37] M. S. Mollel, S. F. Kaijage, K. Michael, Deep reinforcement learning based handover management for millimeter wave communication (2021).

[38] Y. Sun, G. Feng, S. Qin, Y.-C. Liang, T.-S. P. Yum, The smart handoff policy for millimeter wave heterogeneous cellular networks, IEEE Transactions on Mobile Computing 17 (6) (2017) 1456–1468.

[39] Y. Junhong, Y. J. Zhang, Drag: Deep reinforcement learning based base station activation in heterogeneous networks, IEEE Transactions on Mobile Computing (2019).

[40] H. Khan, A. Elgabli, S. Samarakoon, M. Bennis, C. S. Hong, Reinforcement learning-based vehicle-cell association algorithm for highly mobile millimeter wave communication, IEEE Transactions on Cognitive Communications and Networking 5 (4) (2019) 1073–1085.

[41] E. Qualcomm, Range expansion for efficient support of heterogeneous networks, R1-083813 (2008).

[42] H. Sun, R. Q. Hu, Heterogeneous cellular networks, John Wiley & Sons, 2013.

[43] 3GPP, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (2013).

[44] Y. Pochet, L. A. Wolsey, Production planning by mixed integer programming, Springer Science & Business Media, 2006.

[45] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT Press, Cambridge, MA, 2018.

[46] H. Y. Ong, K. Chavez, A. Hong, Distributed deep q-learning, arXiv preprint arXiv:1508.04186 (2015).

[47] V. Mnih et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.

[48] M. Hausknecht, P. Stone, Deep reinforcement learning in parameterized action space, arXiv preprint arXiv:1511.04143 (2015).

[49] S. Gu, T. Lillicrap, I. Sutskever, S. Levine, Continuous deep Q-learning with model-based acceleration, in: Proc. 33rd Int. Conf. Mach. Learn., New York, NY, USA, 2016, pp. 2829–2838.

[50] H. V. Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double Q-learning, in: Proc. 30th AAAI Conf. Artificial Intell., Phoenix, Arizona, USA, 2016, pp. 2094–2100.

[51] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602 (2013).

[52] D. Soldani, Y. J. Guo, B. Barani, P. Mogensen, I. Chih-Lin, S. K. Das, 5g for ultra-reliable low-latency communications, Ieee Network 32 (2) (2018) 6–7.

[53] D. López-Pérez, M. Ding, H. Claussen, A. H. Jafari, Towards 1 gbps/ue in cellular systems: Understanding ultra-dense small cell deployments, IEEE Communications Surveys & Tutorials 17 (4) (2015) 2078–2101.

[54] Pytorch (2019) torchvision.models.
URL `https://pytorch.org/docs/stable/torchvision/models.html`

[55] D. B. Johnson, D. A. Maltz, Dynamic source routing in ad hoc wireless networks, in: Mobile computing, Springer, 1996, pp. 153–181.

[56] P. Nain, D. Towsley, B. Liu, Z. Liu, Properties of random direction models, in: Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies., Vol. 3, IEEE, 2005, pp. 1897–1907.

[57] F. Bai, N. Sadagopan, A. Helmy, Important: A framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks, in: IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428), Vol. 2, IEEE, 2003, pp. 825–835.