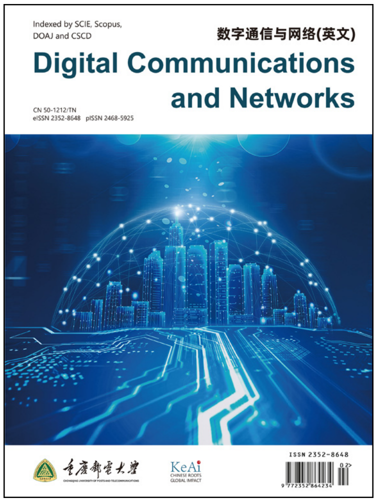

# A novel handover scheme for millimeter wave network: An approach of integrating reinforcement learning and optimization

Ruiyu Wang, Yao Sun, Chao Zhang, Bowen Yang, Muhammad Imran et al.

PII: S2352-8648(23)00136-0
DOI: https://doi.org/10.1016/j.dcan.2023.08.002
Reference: DCAN 685

To appear in: *Digital Communications and Networks*

Received date: 18 December 2021
Revised date: 13 July 2023
Accepted date: 17 August 2023

Please cite this article as: R. Wang, Y. Sun, C. Zhang et al., A novel handover scheme for millimeter wave network: An approach of integrating reinforcement learning and optimization, *Digital Communications and Networks*, doi: https://doi.org/10.1016/j.dcan.2023.08.002.

Digital Communications and Networks(DCN)

HOSTED BY
ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

journal homepage: www.elsevier.com/locate/dcan

# A novel handover scheme for millimeter wave network: An approach of integrating reinforcement learning and optimization

**Ruiyu Wang[a], Yao Sun[a], Chao Zhang[b], Bowen Yang[a], Muhammad Imran[a], Lei Zhang[a]**

[a]**School of Engineering, University of Glasgow, Glasgow, UK**
[b]**Shanghai Shiyue Computer Technology Co., Ltd, Shanghai, China**

## Abstract

The millimeter-Wave (mmWave) communication with the advantages of abundant bandwidth and immunity to interference has been deemed a promising technology to greatly improve network capacity. However, due to such characteristics of mmWave, as short transmission distance, high sensitivity to the blockage, and large propagation path loss, handover issues (including trigger condition and target beam selection) become much complicated. In this paper, we design a novel handover scheme to optimize the overall system throughput as well as the total system delay while guaranteeing the Quality of Service (QoS) of each User Equipment (UE). Specifically, the proposed handover scheme called O-MAPPO integrates the Reinforcement Learning (RL) algorithm and optimization theory. The RL algorithm known as Multi-Agent Proximal Policy Optimization (MAPPO) plays a role in determining handover trigger conditions. Further, we propose an optimization problem in conjunction with MAPPO to select the target base station. The aim is to evaluate and optimize the system performance of total throughput and delay while guaranteeing the QoS of each UE after the handover decision is made. The numerical results show the overall system throughput and delay with our method are slightly worse than that with the exhaustive search method but much better than that using another typical RL algorithm Deep Deterministic Policy Gradient (DDPG).

*KEYWORDS:*
Reinforcement Learning, Optimization theory, MmWave, Cellular network, Handover.

## 1. Introduction

Millimeter-Wave (mmWave), with bandwidth ranging from 30 to 300GHz, is a promising solution to improve the performance of the wireless communication system. However, there are some obvious challenges in mmWave networks, such as short transmission distance, large propagation path loss and high sensitivity to blockage; for example, mmWave can be easily blocked by building materials, even the human body and high oxygen absorption [1]. To overcome the drawbacks of mmWave, beamforming and dense Small Cell Base Stations (SCBSs) architecture [2, 3] play a major role in mmWave communication [4]. Especially, cellular networks with dense SCBSs could improve the efficient propagation of mmWave while beamforming offers a potential solution for mmWave to avoid the blockage.

However, with the increase of SCBS in the cellular network and the propagation becoming directional, there is a great challenge for the Handover (HO) in mmWave cellular network [5, 6]. Specifically, with the SCBS increasing, the inter-cell handover becomes more frequent, leading to higher HO rates. The User equipments (UEs) need to switch from one SCBS (or one beam) to another while moving to maintain the communication quality [7]. In particular, HO mechanisms affect not only the Quality of Service (QoS) on UE side but also the network performance [8]. Since there is a limitation of the resource in BS, growing HO rates usually brings some problems to the network, such as the increase of the HO failures rate and higher signalling overheads, which reduces the system performance [5]. Further, since most beamforming techniques in particular are directional, the HO event also occurs when UEs move from one beamforming covering area to another. In this case, the intra-cell HO also grows significantly compared with the traditional network structure. According to [9, 10], the average handover interval could be lower than 0.75 seconds in the typical mmWave cellular network scenarios and approximately 61% of handovers are unnecessary. Therefore, how to improve the HO efficiency in mmWave cellular network is a key issue to be resolved.

In the traditional communication network, to reduce the redundant handover, the 3rd Generation Partnership Project (3GPP) [11] defines that handover is triggered when the Reference Signal Received Power (RSRP) of current serving BS is lower than the threshold and RSRP of targeting BS is stronger than the current serving BS. However, this method is not adapted to the mmWave cellular network, resulting in the frequent HO and the increase of HO overhand [12]. Therefore, it is crucial to establish an advanced handover mechanism. As the mm Wave scenario becomes more complex, plenty of optimization problems are nonlinear, making the traditional mathematical tool less efficient in solving the problem. In this case, one of the widely-used AI algorithms, Reinforcement Learning (RL), could be designed for a smart handover mechanism in mmWave cellular network, via the interactions with the network environment. However, only in this way, can it not meet the Quality of Service (QoS) of the mmWave network, since RL method focuses on the handover trigger decisions. Further, with the number of UEs and SCBSs increasing, the resource allocation becomes difficult. In other words, resource allocation should be optimized in conjunction with handover decisions. RL method typically estimates the UEs' action through the interaction with the environment, which takes a long time to converge. In this case, after the RL algorithm makes the handover trigger decision, we implement the optimization theory to manage the resource allocation, target BS and beam selection in each SCBS, which not only improves the overall system performance, including total throughput and delay but also guarantees the QoS of each UE. The reason why only applying RL can not achieve the best effect of improving system performance is that RL is an expert to make the decision. In HO scenario, it would be significantly useful for HO trigger decision. However, if we only apply RL to optimize the resource allocation and improve system performance, especially in multi-agents scenario, to achieve the global optimization, RL algorithms would ignore the local optimization for each UE[13]. Therefore, in this case, we only apply RL for decision optimization and use optimal theory to improve the performance of each UE, thereby achieving the best performance of the system.

In this paper, we proposed a novel handover scheme called the optimization-based MAPPO (O-MAPPO) method to help UEs make the optimal handover decision regarding targeting beam and BS and improve the overall system performance, including increasing total system throughput and reducing total system delay. Further, with the assistance of our method, the demand of individual UE, in terms of QoS is met. From the numerical results, we demonstrate our method achieves better performance compared with other typical RL algorithms DDPG and performs slightly worse than the exhaustive search method. The main contributions of this paper are as follows:

1. The O-MAPPO method consists of two parts. HO trigger conditions are learnt by the intelligent handover trigger condition with MAPPO. Meanwhile, the optimal handover decision is to optimize the beams and BSs selection as well as bandwidth allocation based on SINR between each UE and its related BS. An intelligent handover trigger condition scheme based on RL algorithm called MAPPO is implemented in the mmWave cellular network to assist each UE in making the best handover trigger decision. With the help of this method, the reliability of handover in the network is improved, including the reduction of HO rate and HO failures.
2. An optimal handover decision scheme based on optimization theory is designed to manage the resources in each SCBS, such as bandwidth allocation and target beam and BS selection, which optimizes the overall system throughput and delay. In addition, to guarantee the QoS of each UE, we set the constraintt in the optimization function, ensuring that the connecting beams and target beams provide promising service to UEs. Further, the information generated by the optimal handover decision scheme is used as the observation and state of the MAPPO algorithm, making the handover decision more promising.
3. A handover penalty mechanism is applied to reduce the HO rate while avoiding unnecessary handover. In this case, the system is optimized in energy efficiency.

The rest of the paper is organized as follows. In Section 2, the related work is discussed. We propose a system model in Section 3. The basic framework of O-MAPPO is stated in Section 4. The design of intelligent handover trigger condition scheme based on the RL algorithm called MAPPO for handover decision of UEs is discussed in Section 5. The design of optimal handover decision to manage the resources in BSs and improve the system performance and guarantee the QoS of each UE is proposed in Section 6. Simulations results and analysis are given in Section Section 7. Finally, in Section 8 concludes the paper.

## 2. Related work

In order to improve the performance of handover in mmWave system, some research work starts to exploit reinforcement learning with the consideration of different factors, including RSRP, QoS of UEs, UE mobility characteristics, and BS load. In [14], an algorithm with predicted channel information is designed to help UE decide whether to make the handover or not based on UE speed, location, and other information. In [15], a handover based image-to-decision is proposed with Deep Q-learning, which can map pictures to a handover decision of UE. Further, a handover algorithm is proposed in [16], which focuses on context parameters, such as UE velocity, channel gain and cell load information. To maximize the average capacity of UE, Markov Decision Process (MDP) model is applied to help to select BS. Similarly, in [17], the authors provide a handover algorithm with MDP, where they combine the handover overhead, cell load and channel condition in the reward function to achieve high throughput while decreasing the handover rate. Further, in [18], the authors propose Q-learning based handover policy, in which the decision is learnt with optimal policy without prior knowledge of the environment. The results show that the significant quality of experience performance is improved in heterogeneous mmWave network.

However, the above research has focused on a single UE handover scenario. In practice, especially in mmWave system, the handover rate is more frequent, and the cost of handover failure is inestimable with the increase of UE. In this case, the authors in [8] design a smart handover policy for multi-UEs with different UE densities to reduce the handover rate while maintaining the QoS of UE. Further, the authors in [19] propose a multi-agent handover algorithm with actor-critic (A3C) method. Specifically, the handover decisions are made by individual candidates' RSRQs and current connection information with a shred artificial neural network. With handover penalty added in the reward function, redundant handover is significantly reduced. In [7], the authors propose a handover management and power allocation scheme to maximize the overall throughput and reduce the handover frequency. To achieve this, the authors develop a multi-agent RL algorithm based on Proximal Policy Optimization (PPO) method, which separates the learning process to centralized training and decentralized execution. In this case, the global information generated from BS can be used to train the UE at the initial stage. After that, the UEs make their decision and take action based on the local observation.

Inspired by all the research mentioned above, our method takes advantage of the RL to train the multi-UEs in mmWave mesh network to make the proper handover decisions on each movement. However, only applying RL can not achieve the best effect of improving system performance. Therefore, compared with other research, we formulate the system performance with an optimal handover decision scheme based on optimization theory. After the RL makes handover trigger decisions, the optimal handover decision scheme will allocate the resources in related SCBSs, such as target beam and BS selection as well as bandwidth capacity, and generate the system performance, including overall throughput and delay, which is also the states and observations sent back to the RL algorithm as feedback. Further, in multi-UEs scenarios, the RL algorithm tends to be globally optimal and ignores the basic demand of individual UE. Optimal handover decision scheme also can give a hand in solving this problem and achieving the most optimal system performance with the QoS of individual UE guaranteed.

## 3. System model

### 3.1. Network topology

The network topology is shown in Fig. 1. We present the mmWave cellular network, consisting of one Macro Base Station (MBS) and $M$ Small Cell Base Station (SCBS) with $N$ beams in each BS. The set of BSs is denoted as $M = \{0, 1, 2, ..., \mathbf{M} - 1\}$, in which 0 represents the index of MBS while $\{1, 2, ..., \mathbf{M} - 1\}$ is the index of SCBS. We assume that each BS has the same number of beams and the set of beams in each BS is denoted as $N = \{0, 1, 2, ..., \mathbf{N} - 1\}$. Further, the set of UEs is defined as $I = \{0, 1, 2, ..., \mathbf{I} - 1\}$. Each UE is served by either the MBS or one SCBS with only one beam. UEs are located at random positions within the coverage of MBS at the initial stage. The UE mobility model is random walk [20].

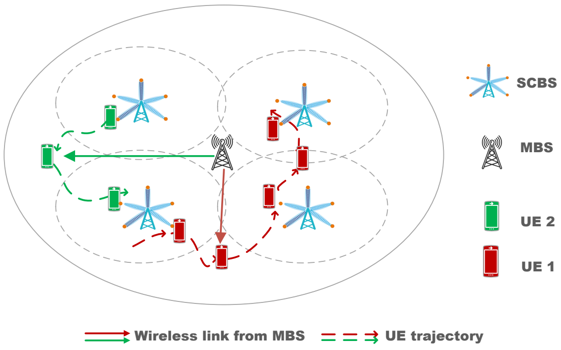

Fig. 1: UEs and BS Distribution.

The channel information of UEs is periodically measured. When UE moves, HO trigger conditions are learnt by RL when either the current SINR cannot meet the demand of UE's service or UE moves to overlapping area. Further, there are two handover cases in our network: inter-cell handover and intra-cell handover. Inter-cell handover occurs among the different BSs, especially when UEs move to the overlapping area and the current SINR is lower than the threshold. Intra-cell handover triggers when UEs moves within the same SCBS, but the current serving SINR cannot meet the demand. When RL decides the handover trigger conditions, the channel information is used to optimize the decisions of beam and BS selection and bandwidth allocation. The UE can either switch to another BS or maintain a different beam in the current BS.

### 3.2. Channel model

The channel models between BS and UE are presented in this subsection. First, the channel model of MBS is introduced. We consider that there is an omnidirectional antenna applied in the MBS to assure the signal coverage. The path loss (in dB) model of the MBS is [21]

$$PL(d)[dB] = \alpha_M + 10\kappa_M \log_{10}(d) + \psi + \xi \qquad (1)$$

where $d$ is the distance in meters, $\kappa_M$ is the path loss exponent representing the slope of the best linear fit to the propagation measurement in the mmWave band, $\alpha_M$ is the path loss factor, $\psi$ is random small-scale fading, and $\xi$ is the random lognormal shadowing.

On the UE $i$ side, we define that $d_i^0$ is the distance between UE and MSB while $p_i^0$ denotes the transmission power from MBS to UE, which satisfies $\sum_i^I p_i^0 = P_M$. Since there is co-channel interference in the mmWave band due to the shared bandwidth, the SINR received by UE from MBS is

$$SINR_i = \frac{PL^{-1}(d_i^0)p_i^0}{\beta_i + N_M \omega_i^0} \qquad (2)$$

where $\beta_i$ is the co-channel interference [1], $N_M$ is the noise power spectral density of MBS, and $\omega_i^0$ represents the bandwidth allocated to UE from MBS.

Second, the channel model of mmWave SCBS is presented. In practice, there are two kinds of channels among different SCBS in mmWave band: Line-Of-Sight (LOS) and Non-Line-Of-Sight (NLOS) channel [22]. We consider a probabilistic LOS-NLOS channel model defined in 3GPP standard [8], which means there are two different channels (LOS and NLOS) for UE in SCBS and the channel can change with its probability. We define that $\nu_i^m$ is the probability of LOS channel adopted from SCBS ($m \in M, m \neq 0$) to UE ($i \in I$). According to [23], where there is an estimation method for LOS channel probability with the building density in the simulation area, the LOS probability from the SBS and UE is:

$$\nu_i^m = \exp(-\frac{2D_B X_B d_i^m}{\pi}), m \neq 0 \qquad (3)$$

where $D_B$ is the building density, $X_B$ is the expectation length of the buildings, and $d_i^m$ is the distance from UE to SCBS. In this case, according to [8], the path loss model of SCBS is:

$$pl(d)[dB] = \alpha_S + 10\kappa_S \log_{10}(d) \qquad (4)$$

where $d$ is the distance in meters, $\alpha_S$ and $\kappa_S$ are the same as those in equation (1), which is path loss factor and exponential decay factor, respectively[2]. The random small-scale fading ($\psi$) and random lognormal shadowing ($\xi$) are ignored since the LOS-NLOS probability mode has already considered.

It is assumed that the directional antennas are equipped on all SCBSs to support beamforming and beam tracking in mmWave system, while there is an omnidirection antenna on UE side to calculate the antenna gain on the SCBS side. In this case, according to [8], the antenna gain is:

$$g(\phi)\begin{cases} g_{max}, & |\phi| < \frac{\phi_S}{2} \\ g_{min}, & otherwise \end{cases} \qquad (5)$$

where $\phi$ is the angle between UE and BS, and $\phi_S$ is the width of the antenna main lobe. In our case, perfect beam tracking is performed, which means the UE is always served by main lobe to obtain the maximum antenna gain.

Since the interference among SCBSs can be ignored in mmWave system, the Signal to Noise Ratio (SNR) is calculated as

$$SNR_i^m = \frac{g_{max} pl^{-1}(d_i^m)p_i^m}{N_S}, m \neq 0 \in M \qquad (6)$$

where $p_i^m$ is the transmission power between UE and SCBS, satisfying $\sum_i^I p_i^m = P_S$, and $N_S$ is the noise power spectral density among SCBSs.

## 4. Framework of O-MAPPO handover scheme

This section proposes an O-MAPPO framework, which contains two parts: intelligent handover trigger condition and optimal handover decision. Specifically, we use MAPPO algorithm to learn the HO trigger condition in the intelligent handover trigger condition. After MAPPO makes the trigger decision, the SINR between UE and BS are calculated based on the channel model and sent to the optimal handover decision. In this part, the beams and BSs selection as well as bandwidth allocation are optimized and evaluated, through which the throughput and delay of all UEs are calculated. The calculation results are then passed to MAPPO as the observation and state to evaluate the handover trigger decision according to the reward function. The basic structure of O-MAPPO framework is shown in Fig. 2.

In more details, there are two handover trigger scenarios applied in our method: handover triggers either in the SCBSs overlapping area, or the serving SINR can not meet the demand. When UEs are moving, MAPPO needs to decide the handover trigger conditions. When the handover trigger occurs, the MAPPO

[1] The interference is the sum power received on the UE side from MBS nearby small cell base station.

[2] $\alpha_S$ and $\kappa_S$ have different values in LOS and NLOS cases.

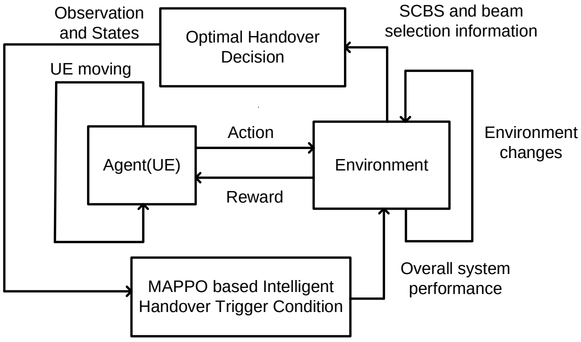
Observation and States
Optimal Handover Decision
SCBS and beam selection information
UE moving
Action
Agent(UE)
Environment
Reward
Environment changes
MAPPO based Intelligent Handover Trigger Condition
Overall system performance

Fig. 2: O-MAPPO framework.

algorithm searches the nearest three target beams in current BS or other BSs with the shortest distance, which can provide the highest SINR to UEs and then send them to the optimal handover decision scheme to make the beams and BSs selection. Further, the optimal handover decision scheme allocates the bandwidth based on the package length of different UEs. It calculates the overall system throughput and delay based on the resource allocation. Meanwhile, during the allocation and calculation, there is a threshold of throughput and delay for individual UE to guarantee the QoS. The allocation and calculation results are then fed back to the MAPPO algorithm as states and observations. According to the reward function in the MAPPO algorithm, each handover trigger decision will be either rewarded or punished. In this way, all UEs learn how to make the best handover trigger decision, which improves the overall system performance and makes the QoS of individual UE promising.

## 5. MAPPO based intelligent handover trigger condition design

This section is based on the design of intelligent handover trigger condition with MAPPO algorithm to learn the handover trigger condition. The MAPPO that we have proposed is a centralized training with MBS but a decentralized execution with BSs the UE connecting framework [7]. The centralized critic and decentralized policy are learnt by the MBS for each UE with our algorithm. Each UE updates its policy based on recent learning results from MBS periodically. Since the UEs in the mmWave system are interactive, we model the problem as a fully cooperative multi-agent task with reinforcement learning. This problem can be described as $\Gamma =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, O, \mathcal{N}, \gamma >$. $\mathcal{S}$ is the state space while $\mathcal{A}$ is the shared action space for each agent. $o_i = O(s; i)$ is the local observation for agent $i$ at global state $s$. $\mathcal{P}(s'|s, A)$ represents the transition probability, while $\mathcal{R}$ is the shared reward function. $\gamma$ is the discount factor, which is $\Sigma_t \gamma^t \mathcal{R}(s^t, a^t)$.

### 5.1. Action

The action of each UE in our system contains handover trigger at each time step $t$. To guarantee the QoS of the UE, MAPPO generates three candidate BSs with the shortest distance and calculates the SINR of UEs. At time step $t$, the action of UE $i$ is expressed as:

$$a_t^i = \{0, 1\} \tag{7}$$

where 1 represents trigger. Since all the UEs in the system are required to be considered, we denote $A$ as the action space of all UEs, which is defined as:

$$A_t = (a_t^1, a_t^2, a_t^3, ..., a_t^i) \tag{8}$$

The reason why we generate the action set of all the UEs' is that there is a trade-off between the single UE reward and the overall system reward. The maximum reward of the single UE not optimal in terms of the overall system reward. The specific statement to solve the trade-off is presented by the end of this sub-section.

### 5.2. State and observation

The current serving beam $n$ in its BS $m$ of the UE $i$ is chosen at the previous time step $t-1$. At the start of each time step, the public information is sent by MBS to each UE. Specifically, for each BS $m \in M$, the total number of served UEs is defined as $I_t^M = \sum_{i\in I} n_{t-1}^i$", where $m$ and $n$ at current time step $t$ is the candidate BSs and beam, which are also parts of state and observation. Therefore, the public information at time step $t$ is $\mathbf{I}_t = (n_t^0, n_t^1, ..., n_t^m)$. At the beginning of each time $t$, the optimal handover decision scheme makes the HO decision from candidate BSs and beams and calculates the handover delay, bandwidth allocation, and overall system throughput based on UEs' actions taken at last time step $t-1$. In this case, for each UE, the observation can be denoted as:

$$s_t^i = (d_{t-1}^{i,m}, r_{t-1}, b_{t-1}^{i,m}, \mathbf{I}_t) \tag{9}$$

where $d_{t-1}^{i,m}$ is the handover delay of each UE at the previous time step $t-1$, $b_{t-1}^{i,m}$ is the bandwidth allocation of each UE at the previous time step, and $r_{t-1}$ is the overall system throughput at the previous time step. Therefore, the global state as the ensemble of observations of all UEs can be defined as

$$S_t = (s_t^1, s_t^1, ..., s_t^I) \in \mathcal{S} \tag{10}$$

where $\mathcal{S}$ is the state space.

### 5.3. Reward

The reward of our algorithm is divided in two parts: overall system throughput and delay evaluation and Handover Rate (HOR). Firstly, since the switch decision leads to the changes of throughput and delay, it is important to evaluate the handover trigger decision based on that. Therefore, we define the system performance reward as:

$$\mathcal{R}^i = \begin{cases} 10\delta, & Rt > Rt_1 \wedge Dt < Dt_1 \\ \delta, & Rt > Rt_2 \wedge Dt < Dt_2 \\ -\delta, & Others \end{cases} \tag{11}$$

where $Rt_1$ is the upper bound of system throughput and $Rt_2$ is the lower bound. Further, $Dt_1$ is the lower bound of system delay while $Dt_2$ is the upper bound. $\delta$ is the reward value. When $R_t$ is higher than its upper band and $D_t$ is lower than its lower bound, we design a great reward to this condition, which is $10\delta$. When $R_t$ and $D_t$ are within their bound, there is a basic reward $\delta$. Otherwise, there would be a penalty $-\delta$.

Secondly, we define the HO penalty, which is to avoid the unnecessary handover trigger decisions:

$$P^i_{HO}(s^u_t, a^u_t) = \varepsilon \mathbb{1}\left\{b^i_t \neq b^i_{t-1}\right\} \tag{12}$$

where $\varepsilon \geq 0$ is the weighting factor. If the target BS $b^i_t$ is different from the current serving BS $b^i_{t-1}$, there will be an HO penalty. After we combine the HO penalty with reward function, the local reward with HO penalty of UE in time step $t$ is expressed as:

$$\mathcal{R}^i_{HO} = \begin{cases} 10\delta P^i_{HO}(s^u_t, a^u_t), & Rt > Rt_1 \wedge Dt < Dt_1 \\ \delta, & Rt > Rt_2 \wedge Dt < Dt_2 \\ -\delta P^i_{HO}(s^u_t, a^u_t), & Others \end{cases} \tag{13}$$

Since the problem is a multi-agent one, we model it as a fully cooperative multi-agent task, where the total reward of UE is

$$\mathcal{R}(S_t, A_t) = \sum_{i=1}^{I} \mathcal{R}^i(s^i_t, A_t) \tag{14}$$

The total reward $\mathcal{R}(S_t, A_t)$ can guild agents to balance the trade-off between SINR trigger condition and HOR with the adjusting weighting factor $\varepsilon$.

### 5.3.1. Q-value and Policy

We define the state-action value function $Q^\pi$, the state value function $V^\pi$, and the advantage function $A^\pi$ as follows

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{s_{t+1}, a_{t+1}}, \ldots [R_t | s_t, a_t] \\ V^\pi(s_t) &= \mathbb{E}_{a_t, s_{t+1}}, \ldots [R_t | s_t] \\ A^\pi(s_t, a_t) &= Q^\pi(s_t, a_t) - V^\pi(s_t) \end{aligned} \tag{15}$$

where $\pi$ is the joint policy. We update the parameters $\omega$ for the critic $V_\omega(s_t)$ by minimizing the loss

$$\begin{aligned} J(\omega) &= \hat{(\mathbb{E})} \left[ (V_\omega(s_t) - y_t) \right] \\ y_t &= \mathcal{R}_t + \gamma V_\omega(s_{t+1}) \end{aligned} \tag{16}$$

where $\mathcal{R}_t$ is the reward in time $t$, and $V_\omega$ is the target state-value function [24].

According to [7], the Independent Proximal Policy Optimization (IPPO) is one of the RL methods that implement the PPO algorithm on each UE independently, where each UE learns the actor and critic on its own. However, this method cannot approach the true overall state value since the state and action information is updated locally on the UE side. In this case, there is no global state information or jointly action information shared on the UE side, which makes the advantage function of IPPO less accurate. In addition, the lack of joint actions makes it more difficult for the UE to learn about cooperation policies and assess the influence of UE action on the reward.

Therefore, we propose the MAPPO algorithm, a centralized training with a decentralized execution framework to improve the performance of the IPPO. In this case, global information is implemented for training the decentralized policies of each UE. More specifically, the global information is supposed to be collected in MBS, and the learning procedure is also processed in MBS.

We implement the decentralized actors and centralized critics framework since the joint advantage function has strong relevance with the policy gradients. In this case, with the global information such as UE action $a_t$ and UE state $s_t$ available, the centralized critic evaluates the joint value function (Q or V) in the training process. At the same time, decentralized actors estimate locally based on UE's observations. When the training process finishes, global information is no longer required, which means the UEs can implement the actions in the decentralized actors. The basic MAPPO structure is shown in Fig. 3, in which there is a neural network in each actor and critic.

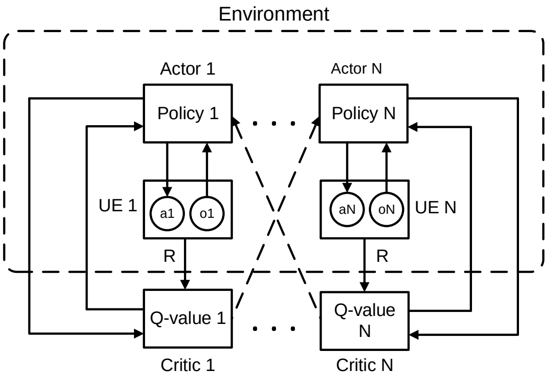

Fig. 3: MPPO structure.

The state-value function $V^i_\omega(s_t)$ is estimated in the centralized critics with the critic parameters $\omega^i$ of the UE. Since the expectation is replaced by sample averages in RL, we update the policy with the gradient:

$$\Delta\theta^i = \nabla^i_\theta \hat{\mathbb{E}}_t \left\{ f(\rho_t(\theta^i), A^i(s_t, a_t)) \right\} \tag{17}$$

where

$$f(\rho_t(\theta^i), A^i(s_t, a_t)) = \min(\rho_t(\theta^i) A^i(s_t, a_t), clip(\upsilon_t(\theta^i), 1 - \epsilon, 1 + \epsilon) A^i(s_t, a_t)) \tag{18}$$

*clip* refers to conservative policy iteration and $\epsilon = 0.2$ is the hyper-parameter, $\rho_t(\theta^i) = \frac{\pi^i(a_t^i|o_t^i)}{\pi_{old}^i(a_t^i|o_t^i)}$, and $\pi$ is the decentralized policy.

$A^i(s_t, a_t)$ is the estimation of joint advantage function, which is calculated by Generalized Advantage Estimation (GAE) [25] with the state-value function $V_\omega^i(s_t)$.

$$A^i(s_t, a_t) = \tau_t + (\gamma\lambda)\tau_{t+1} + ... + (\gamma\lambda)^T \tau_T \tag{19}$$

where $\tau_t = r_t + \gamma V_{\hat{\omega}}^i(s_{t+1}) - V_{\hat{\omega}}^i(s_t)$, and $V_{\hat{\omega}}$ is the local critic of UE $i$, $\gamma \in [0, 1]$ is an estimator of the value function. According to [7], there is a credit assignment problem, since it is not clear how a specific UE action contributes to the reward. In order to solve it, the counterfactual baseline method proposed in [26] is used. In our case, we propose a centralized critic $Q_{\omega^i}(s_t, a_t)$ to evaluate the action-value function. The joint quantities are denoted to UE as $-i$. Therefore, the advantage function for each UE is calculated by comparing the Q-value estimated by the critic for the executed action $a_t^i$ to a counterfactual baseline that marginalizes $a_t^i$, maintaining the actions of other UEs

$$A^i(s_t, a_t) = \hat{Q}^i(s_t, a_t) - b(s_t, a_t^{-i}) \tag{20}$$

where $b(s_t, a_t^{-i})$ is the counterfactual baseline, defined as

$$b(s_t, a_t^{-i}) = \sum_{a^i} \pi_{old}^i(a^i \mid z_t^i) Q_\omega^i(s_t, (a_t^{-i}, a_i)) \tag{21}$$

where $\pi_{old}^i$ is the initial guess of the optimal policy and $\hat{Q}^i$ is the estimation of $Q^{\pi_{old}}$, which is calculated by the Temporal-Difference (TD) [27]. Although each $\hat{Q}^i$ is calculated by separated critics, the joint action-value function $Q^{\pi_{old}}(s_t, a_t)$ is the same.

Further, the discrete action space $\mathcal{A}$ is considered. After the state-action $(s_t, a_t)$ is input into the critic, the scalar $Q_{\omega^i}(s_t, a_t)$ is obtained. However, $|\mathcal{A}|$ times evaluation is necessary when the counterfactual baseline is computed, which is time-consuming when the action space is getting larger. In this case, we use the critic structure in [26]. The input of the neural network of the critic of UE is $Q_{\omega^i}(s_t, a_t)$ while the output is the state-action values of the UE. In order to distinguish whether the specific UE action is marginalized, the critic structure requires that there must be a critic for each UE. The procedure of MAPPO is presented in Algorithm 1.

**Algorithm 1** MAPPO procedure.

```
Initiate critic $Q_{\omega^i}$ and actor $\pi^i$ with $\theta^i$, $\forall i$.
Initiate the initial policies $\pi_{old}^i$ and target critic $Q_{\hat{\omega}^i}$.
initiate state.
for iteration=1,2,...,T do
    initiate state.
    for an episode do
        Executes action for each agent.
        Get reward and the new state.
    end
    Get the movement of each UE.
    Calculate the $\hat{Q}^i(s_t, a_t)$.
    Calculate all UEs' action space $A^i(s_t^i, a_t)$.
    Store the data $\left\{z_t^i, Q^i(s_t, a_t), A^i(s_t^i, a_t)\right\}$ into database $D$.
    for k=1,2,3,...,K do
        Shuffle and relabel the data.
        for j=0,1,2,....,H do
            Select groups of data $D_j$:
            Calculate new action space.
            for l=1,2,...,L do
                Calculate gradient ascent $\Delta\theta^i$.
                Use minibatch Adam [28] to apply gradient ascent $\theta^i$.
                Calculate gradient ascent $\Delta\omega^i$.
                Use minibatch Adam [28] to apply gradient ascent $\omega^i$.
            end
        end
    end
    Update $\theta^i$ and $\omega^i$ for each UE.
    Clear database $D$.
end
```

## 6. Optimal beam selection and bandwidth allocation for handover UEs

After the handover trigger decision is made, the related channel information is sent to optimal handover decision scheme to optimize bandwidth allocation as well as beams and BSs selection. Thereby, it can improve the handover delay as well as the overall system throughput while QoS of each UE is guaranteed. Further, the results of calculation and allocation will be then transferred to the MAPPO algorithm to evaluate the handover trigger decision through reward function. From 3GPP [29], the handover delay in mmWave system is defined as

$$D = T_R + T_I + T_T \tag{22}$$

where $T_R$ is the Radio Resource Control (RRC) procedure delay. It is the time from RRC procedures decided by the communication system; $T_I$ is the handover interruption time which includes target cell searching time, target cell tracking and acquiring time, and interruption uncertainty time, which is the interruption uncertainty in acquiring the first available physical random access channel occasion in the new cell and it is also decided by the network system. Thus, the handover transmission time $T_T$ between UE and BS is the key to optimize the handover delay, which is expressed as:

$$T_T^i = \frac{PL_i}{R_i} \tag{23}$$

where $PL$ is the package length and $R$ is the throughput of the system and it is defined by Shannon Formula:

$$R_i = B_i \log(1 + SINR_i) \tag{24}$$

where $B$ is the bandwidth taken by the UE.

Based on the system model, we denote the optimal handover decision scheme to improve the system performance in terms of throughput and delay while guarantee the QoS of each UE, which can be denoted as:

$$\min \sum_{i\in I} \sum_{n\in N} \sum_{m\in M} x_{m,n}^i \times Delay_i \tag{25}$$

$$\text{s.t.} \quad \sum_{i\in I} \sum_{n\in N} x_{m,n}^i \times B_i \le B_0 \quad MHz, \forall m \in M \tag{26}$$

$$\sum_{m\in M} \sum_{n\in N} x_{m,n}^i \times R_i \ge R_0 \quad Mbps, \forall i \in I \tag{27}$$

$$\sum_{m\in M} \sum_{n\in N} x_{m,n}^i = 1, \forall i \in I \tag{28}$$

$$Delay_i \le D_0 \quad ms, \quad R_i \ge R_{i0} \quad Mbps \quad \forall i \in I \tag{29}$$

where $x_{m,n}^i \in (1, 0)$ describes the UE connection statues. When $x_{m,n}^i = 1$, it means UE $i$ connects with the beam $n$ in BS $m$; on the contrary, $x_{m,n}^i = 0$. Since the maximum bandwidth of each SCBS is fixed, (24) is the constraint showing that the maximum bandwidth for each SCBS, which can provide to the UE. (25) is the constraint, which aims to optimize the total throughput of all UEs. We set a lower bound of the throughput, which formulates the minimum throughput all the UEs can gain from the system. In addition, each UE can only connect with one beam in one BS, which is constrainted by (26). (27) is the minimum delay and throughput that each UE must reach, which is used for guaranteeing the QoS of each UE. $B_0$ is the maximum bandwidth in each SCBS while $R_0$ is the minimum throughput that each SCBS must reach.

From our optimal handover decision scheme i.e., (23)-(27), as can be seen, since $x_{m,n}^i \in (0, 1)$ and the equation (22) is a nonlinear function, the optimization function is a zero-one mix integer nonlinear problem and there are two unknown parameters ($x_{m,n}^i$ and $B_i$) to figure out. In this case, we divide the problem into three parts. Firstly, we utilize the Sequential Quadratic Programming (SQP) algorithm [30] to solve the nonlinear part. In this case, we relax integer $x_{m,n}^i$ as a continuous variable, which ranges from 0 to 1. Secondly, after we obtain the continuous $x_{m,n}^i$, we use tight relaxation algorithm [31] to transfer continuous variable into integer variable. Thirdly, after we solve the integer problem, the rest of the optimization function becomes a linear problem: the function of $B_i$. We solve it with a linear algorithm to obtain the most suitable bandwidth $B_i$ for each UE. With the bandwidth allocation of each UE, the throughput can be calculated; thereby, the delay of different UEs can be obtained.

According to the SQP algorithm, the Lagrangian function of the optimization function is:

$$L = F + \alpha h_1 + \beta h_2 + \gamma h_3 \tag{30}$$

where

$$F = \sum_{i\in I} \sum_{n\in N} \sum_{m\in M} x_{m,n}^i \times \frac{PL}{\log(1 + SINR_i)} \tag{31}$$

$$h_1 = \sum_{i\in I} \sum_{n\in N} x_{m,n}^i \times B_i \tag{32}$$

$$h_2 = \sum_{m\in M} \sum_{n\in N} x_{m,n}^i \times B_i \log(1 + SINR_i) \tag{33}$$

$$h_3 = \sum_{m\in M} \sum_{n\in N} x_{m,n}^i \tag{34}$$

Here, the $PL$ is the package length and $SINR$ is channel state information, which means that the two variables are known in the system. In this case, the optimization function is a nonlinear problem as the function of bandwidth ($B_i$). Then we figure out the first order approximation of the gradient of the Lagrangian function

$$\nabla L = \begin{bmatrix} \frac{dL}{dx} \\ \frac{dL}{d\alpha} \\ \frac{dL}{d\beta} \\ \frac{dL}{d\gamma} \end{bmatrix} = \begin{bmatrix} \nabla F + \alpha \nabla h_1 + \beta \nabla h_2 + \gamma \nabla h_3 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix} \tag{35}$$

Then, the second order approximation of the gradient of the Lagrangian function is:

$$\nabla^2 L = \begin{bmatrix} \nabla_x^2 L & \nabla h_1 & \nabla h_2 & \nabla h_3 \\ \nabla h_1 & 0 & 0 & 0 \\ \nabla h_2 & 0 & 0 & 0 \\ \nabla h_3 & 0 & 0 & 0 \end{bmatrix} \tag{36}$$

We define that $p = \frac{\nabla^2 L}{\nabla L} = \frac{\nabla^2 L(p)}{\nabla L(p)}$. In this case, we can simplify the nonlinear optimization function as:

$$\min(p) F(x_k) + \nabla F(x_k)^T p + \frac{1}{2} p^T \nabla_x^2 L_k p \tag{37}$$

$$\text{s.t.} \quad \nabla h_1 p + h_1 \le B_0 \quad MHz, \forall m \in M \tag{38}$$

$$\nabla h_2 p + h_2 \ge R_0 \quad Mbps, \forall i \in I \tag{39}$$

$$\nabla h_3 p + h_3 = 1, \forall i \in I \tag{40}$$

$$Delay_i \le D_0 \quad ms, \quad R_i \ge R_{i0} \quad Mbps \quad \forall i \in I \tag{41}$$

where $k$ is the number of the iteration. Therefore, the nonlinear function is approximated in the linear function and we can obtain the continuous $x_{m,n}^i$.

We utilize the tight relaxation to transfer continuous $x_{m,n}^i$ into integer. According to [32], the method we use is implicit enumeration. As shown in Fig. 4, the procedure we follow to search the possible $x_{m,n}^i$ is

1. Since we can obtain a set of variable $x_{m,n}^i$ with the help of the SQP algorithm for each UE, the size of which is $m \times n$, we select the top three largest $x_{m,n}^i$ to one of the permissible integer values.

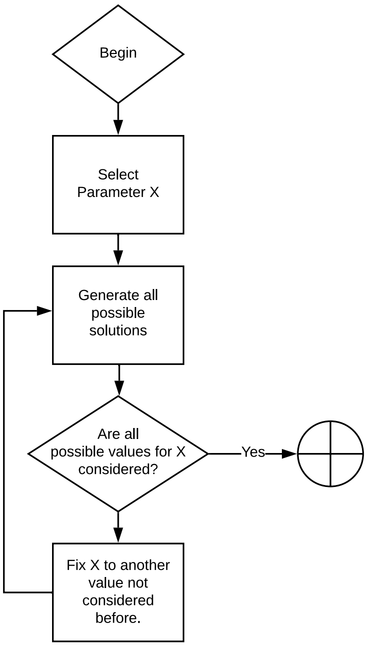
Begin
Select Parameter X
Generate all possible solutions
Are all possible values for X considered?
Yes
Fix X to another value not considered before.

Fig. 4: Zero-One tight relaxation procedure flow diagram.

2. Resolve the problem in the remaining variable.
3. Fix one of three $x^i_{m,n}$ to another permissible value.
4. Repeat (2) and (3) until all possible for $x^i_{m,n}$ are considered.

This algorithm is a basic search that implies a general state of search in which all possible solutions are considered either explicitly or implicitly. Finally, the zero-one nonlinear integer problem is approximated into a linear problem, which is easy to be solved.

## 7. Results and discussions

In this section, the simulation setups are presented and we show some numerical results with discussions and analysis.

### 7.1. Simulation setup

We propose a two-tier heterogeneous mmWave cellular network consisting of one microwave MBS, $M_s$ mmWave SCBSs with the number of UE $I$, and for each SCBS, there are total $N$ beams. Specifically, we donate $M_s = 6$, $N = 8$ and $I = 10$ as default. Cartesian coordinates describe the location of BS and SCBS. We assume that there is an effective propagation coverage with 200 meters radius. The MBS is located in origin $(0, 0)$, and the rest of the six SCBSs are evenly distributed in the considered area. Further, in each SCBS, eight directional beams are equally distributed with 45°. The coverage of each SCBS has an overlapping area with its neighbour. In this case, a handover event occurs either when the current SINR of UEs is lower than the threshold or when the UEs are in an overlapping area for inter-cell handover. On the other hand, intra-cell handover occurs when UEs move from one beam area to another. Furthermore, the UEs that the SCBS cannot cover are served by MBS, which usually occurs on the edge of the effective propagation area. The log-normal shadow fading of MBS $\xi$ has zero mean and 3dB standard deviation, and the small-scale fading in linear value from $10^{\frac{\psi}{10}}$ follows an exponential distribution with unit mean [7]. We summarize the other channel parameters in Table I. The initial location of UEs are randomly distributed, and the movement of them obeys the random walk method [20] with the velocity 2 $m/s$.

Table 1: Channel parameters of mmWave cellular network [8, 21, 33]

| Parameters | Value |
|---|---|
| Bandwidth of SCBSs | 100 MHz |
| Bandwidth of MBS | 20 MHz |
| Pathloss parameters of LOS | $\alpha_S = 70, \kappa_S = 2$ |
| Pathloss parameters of NLOS | $\alpha_S = 70, \kappa_S = 2$ |
| Pathloss parameters of MBS channel | $\alpha_M = 70, \kappa_M = 2$ |
| MmWave noise power density | -163 dBm/Hz |
| Microwave noise power density | -174 dBm/Hz |
| The maximum antenna gain | 10 dB |
| Interval of each timestep | 100 ms |
| Interruption time | 100 ms |
| Building density | $1 \times 10^{-4}/m^2$ |
| Expected length of buildings | 25 m |
| Transmission power of MBS | 46 dBm |
| Transmission power of SCBS | 30 dBm |
| Upper bound of system throughput ($Rt_1$) | 2300 Mbps |
| Lower bound of system throughput ($Rt_2$) | 1800 Mbps |
| Upper bound of system delay ($Dt_2$) | 0.022 s |
| bound of system delay ($Dt_1$) | 0.018 s |

For the hyper-parameters of MAPPO, we apply the Adam optimizer with the learning rate $lr = 5 \times 10^{-4}$. We consider one hidden layer with 64 units using Rectified Linear Unit (ReLU) activation function for the policies and critics. We set the minibatch size to be 128, discount factor to be 0.9, $\gamma = 0.5$, and *clip* loss value 0.2.

### 7.2. Results and discussions

In the following sub-section, to evaluate the system performance in terms of system total throughput and delay, some simulations are implemented with different reward conditions.

**Simulation I:** The number of UE ($I$) is 10 and the configured upper bounds and lower bound of throughput ($Rt_1$), delay ($Dt_1$), ($Rt_2$), and delay ($Dt_2$) can be found in Table I, which means if the throughput or delay cannot reach the lower bound, the training episodes will be done in that round and the training will start again.

First, the training loss is shown in Fig. 5. As can be seen, the training loss decreases with the increasing training episodes, and it eventually converges to 0.1, which means good performance of our method in this simulation and all the numerical results are promising.

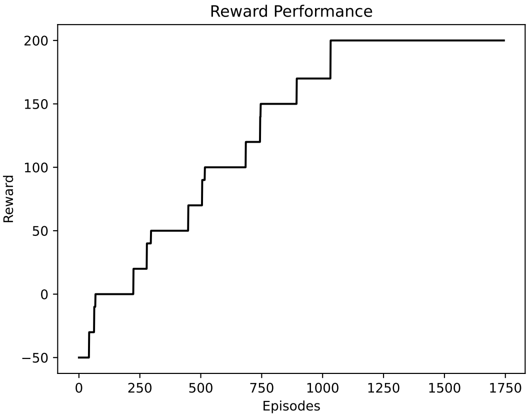

Fig. 5: Training reward in Simulation I.

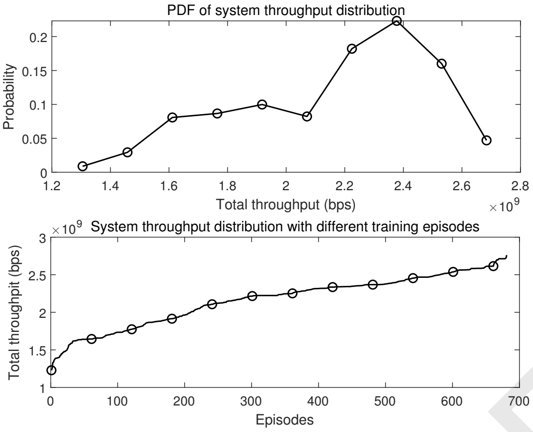

Fig. 6: System throughput performance in Simulation I.

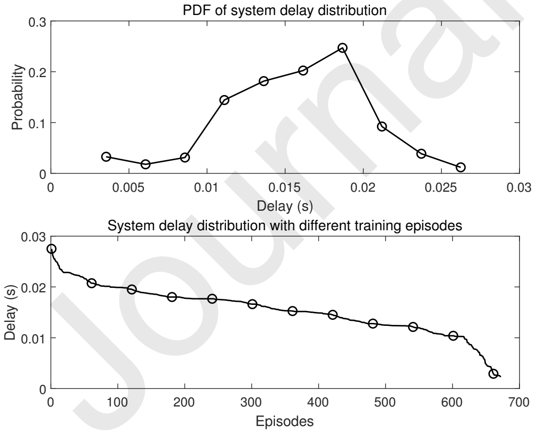

Fig. 7: System delay performance in Simulation I.

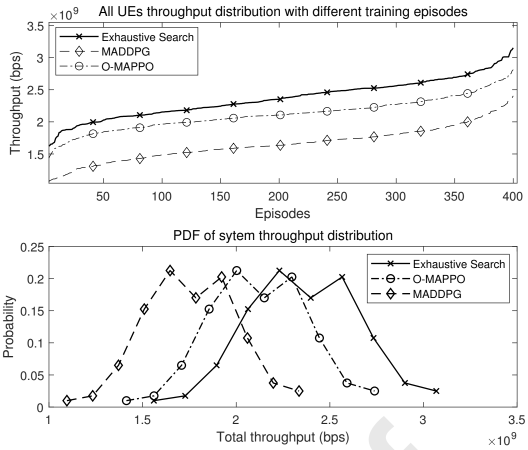

Fig. 8: All UEs throughput performance with different algorithms $UE = 10$ in Simulation II.

Second, the Probability Density Function (PDF) and distribution of system total throughput are shown in Fig. 6. During the training process, most of the system total throughput is distributed from $1800Mbps$ to $2300Mbps$, and the episodes, of which throughput is lower than $1800Mbps$, are 19.5%. Furthermore, from the system throughput distribution curve, the lower throughput occurs in the early stage of the training process. In this stage, the reward is not large enough, which implies that the learning process of UEs is bad. With the training continuing, the reward becomes larger and larger, and the system achieves good control of the SCBSs and beams allocation for each UE. Therefore, the total system throughput increases.

Third, the PDF and distribution of system total delay are shown in Fig. 7. Alone with the training progress, the system total delay mainly distributes from $0.013s$ to $0.023s$. There are 0.16% of episodes, of which the delay is higher than the lower reward bound $Dt_2$, and there are 53% of episodes, of which the delay is lower than the upper reward bound $Dt_1$. Further, the total delay reduces with the training episode increasing, which indicates that UEs achieve a better learning ability. The performance in terms of total system delay is better than total system throughput. The reason is that two conditions (high throughput and low delay) must trigger at the same time to obtain a good reward. Moreover, the upper trigger condition for the system delay is easier than that for the system throughput.

**Simulation II:** In this simulation, we make a comparison between the proposed algorithm with other typical RL algorithms, such as Deep Deterministic Policy Gradient (DDPG) and the exhaustive search method, in terms of system total delay and throughput performance, respectively. The lower bound and upper bound of reward are the same as that in Simulation I, and all the algorithms are compared when the number of UE is 10 ($I = 10$).

From Fig. 8 and Fig. 9, it can be seen that the exhaustive search method has the best performance, especially when the total system throughput is compared. The performance of our method is slightly worse than the exhaustive search method but much better than MADDPG algorithm. For the total sys-

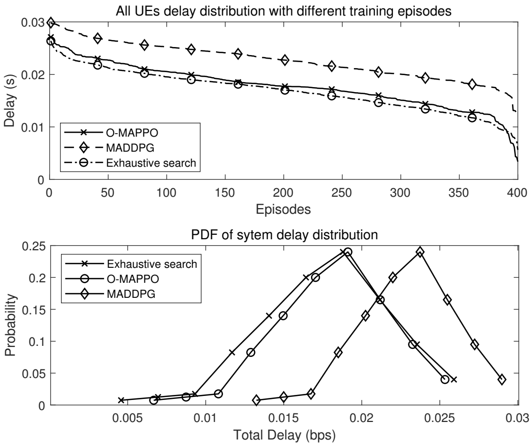

Fig. 9: All UEs delay performance with different algorithms $UE = 10$ in Simulation II.

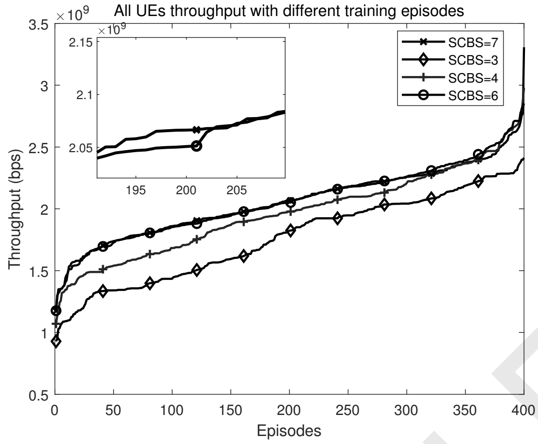

Fig. 10: All UEs throughput performance with different SCBSs in Simulation III.

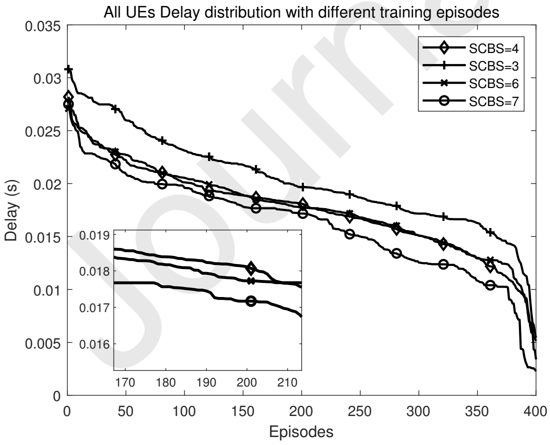

Fig. 11: All UEs delay performance with different SCBSs in Simulation III.

tem delay performance, the result is the same as the comparison of system throughput.

**Simulation III:** This section compares the system performance with different SCBSs to explore the system threshold of SCBS allocation when the number of UEs is 10. The number of SCBSs ranges from 3 to 7 in this simulation. However, from the zoom in part of Fig.10 and Fig. 11, the system throughput and delay performance are almost the same when the number of SCBSs is $M_s = 6$ and $M_s = 7$, which means the maximum capacity of SCBSs in our simulation scenario is $M_s = 6$. As can be seen, there is slight difference between $M_s = 7$, $M_s = 6$, and $M_s = 64$ for the system performance in terms of system delay and throughput. However, there is a significant gap between $M_s = 4$ and $M_s = 3$ for both system performances. This simulation demonstrates that the system total throughput performance reaches its maximum value when the number of SCBSs is 6 in our simulation environment. Further, the performance of $M_s = 3$ is the worst. When it comes to the total system delay performance, shown in Fig. 11, the more SCBSs lead to the better performance until the system reaches its threshold. The reason is that since the MBS effective signal coverage is fixed, when the number of SCBSs is reduced, the radius of each SCBSs increases, making the propagation distance of mmWave in each cell longer. MmWave attenuates as propagation distance increases, which leads to the SINR of each UE reducing in high probability when connecting with SCBS. Thus, the system performance worsens with the number of SCBSs reducing. Meanwhile, when the number of SCBSs increases, the handover rate must be higher. Since each cell's radius is smaller, the overlapping area in the region grows, which means more handover events will occur.

**Simulation IV:** This section evaluates the reliability of our HO scheme in terms of making a comparison between total HO times and HO Failure (HOF) times with different UEs in 700 training episodes. The numerical result is shown in Fig. 12. As can be seen, HOF time grows up with the increase of UE in the system. However, most HO failures occur at the initial training stage, where our algorithm is still learning experience with the environment change. Further, the HOF rate can maintain at a low level when the number of UEs increases, which demonstrates the reliability of our HO scheme.

## 8. Conclusions

An optimization-theory-based on one of the RL methods called O-MAPPO is proposed in this paper to optimize the total system delay and throughput. Specifically, the RL algorithm called MAPPO is applied to improve the handover trigger decision. After the handover triggers, the related channel information is sent to the optimal handover decision scheme to optimize the beams and BSs selection, bandwidth allocation, improving the performance of overall system throughput and delay. Further, to avoid unnecessary HO and lower the HO rate, we implement HO penalty

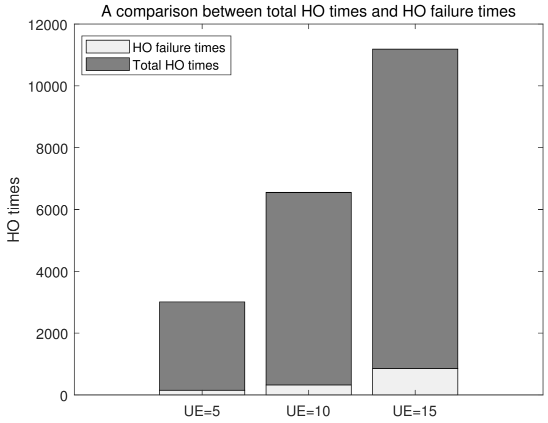

Fig. 12: A comparison between HO times and HOF times with different UEs in Simulation IV.

strategy to improve the efficiency of the system. Simulation results demonstrate that with the training processing, our method can achieve better performance in terms of total system throughput and delay compared with some typical RL algorithms, such as MADDPG and DQN.

## References

[1] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, F. Aryanfar, Millimeter-wave beamforming as an enabling technology for 5g cellular communications: Theoretical feasibility and prototype results, IEEE communications magazine 52 (2) (2014) 106–113.

[2] B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to spatial filtering, IEEE assp magazine 5 (2) (1988) 4–24.

[3] T. Bai, R. W. Heath, Coverage and rate analysis for millimeter-wave cellular networks, IEEE Transactions on Wireless Communications 14 (2) (2014) 1100–1114.

[4] Q. Xue, Y. Sun, J. Wang, G. Feng, L. Yan, S. Ma, User-centric association in ultra-dense mmwave networks via deep reinforcement learning, IEEE Communications Letters 25 (11) (2021) 3594–3598.

[5] M. Tayyab, X. Gelabert, R. Jäntti, A survey on handover management: From lte to nr, IEEE Access 7 (2019) 118907–118930.

[6] Y. Sun, W. Jiang, G. Feng, P. V. Klaine, L. Zhang, M. A. Imran, Y.-C. Liang, Efficient handover mechanism for radio access network slicing by exploiting distributed learning, IEEE Transactions on Network and Service Management 17 (4) (2020) 2620–2633.

[7] D. Guo, L. Tang, X. Zhang, Y.-C. Liang, Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning, IEEE Transactions on Vehicular Technology 69 (11) (2020) 13124–13138.

[8] 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.

[9] B. Van Quang, R. V. Prasad, I. Niemegeers, A survey on handoffs—lessons for 60 ghz based wireless systems, IEEE Communications Surveys & Tutorials 14 (1) (2010) 64–86.

[10] A. Talukdar, M. Cudak, A. Ghosh, Handoff rates for millimeterwave 5g systems, in: 2014 IEEE 79th Vehicular Technology Conference (VTC Spring), IEEE, 2014, pp. 1–5.

[11] A. Barbuzzi, P. H. Perala, G. Boggia, K. Pentikousis, 3gpp radio resource control in practice, IEEE Wireless Communications 19 (6) (2012) 76–83.

[12] C. Shen, M. van der Schaar, A learning approach to frequent handover mitigations in 3gpp mobility protocols, in: 2017 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2017, pp. 1–6.

[13] M. J. Kochenderfer, T. A. Wheeler, Algorithms for optimization, Mit Press, 2019.

[14] Y. Koda, K. Yamamoto, T. Nishio, M. Morikura, Reinforcement learning based predictive handover for pedestrian-aware mmwave networks, in: IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2018, pp. 692–697.

[15] Y. Koda, K. Nakashima, K. Yamamoto, T. Nishio, M. Morikura, Handover management for mmwave networks with proactive performance prediction using camera images and deep reinforcement learning, IEEE Transactions on Cognitive Communications and Networking 6 (2) (2019) 802–816.

[16] F. Guidolin, I. Pappalardo, A. Zanella, M. Zorzi, Context-aware handover policies in hetnets, IEEE Transactions on Wireless Communications 15 (3) (2015) 1895–1906.

[17] 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.

[18] H. Tabrizi, G. Farhadi, J. Cioffi, Dynamic handoff decision in heterogeneous wireless systems: Q-learning approach, in: 2012 IEEE international conference on communications (ICC), IEEE, 2012, pp. 3217–3222.

[19] Z. Wang, L. Li, Y. Xu, H. Tian, S. Cui, Handover control in wireless systems via asynchronous multiuser deep reinforcement learning, IEEE Internet of Things Journal 5 (6) (2018) 4296–4307.

[20] F. Spitzer, Principles of random walk, Vol. 34, Springer Science & Business Media, 2013.

[21] S. Zang, W. Bao, P. L. Yeoh, B. Vucetic, Y. Li, Managing vertical handovers in millimeter wave heterogeneous networks, IEEE Transactions on Communications 67 (2) (2018) 1629–1644.

[22] R. Wang, P. V. Klaine, O. Onireti, Y. Sun, M. A. Imran, L. Zhang, Deep learning enabled beam tracking for non-line of sight millimeter wave communications, IEEE Open Journal of the Communications Society 2 (2021) 1710–1720.

[23] T. Bai, R. Vaze, R. W. Heath, Using random shape theory to model blockage in random cellular networks, in: 2012 International Conference on Signal Processing and Communications (SPCOM), IEEE, 2012, pp. 1–5.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, nature 518 (7540) (2015) 529–533.

[25] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, arXiv preprint arXiv:1506.02438.

[26] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, S. Whiteson, Counterfactual multi-agent policy gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[27] G. Tesauro, Temporal difference learning and td-gammon, Communications of the ACM 38 (3) (1995) 58–68.

[28] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[29] E. U. T. R. Access, Requirements for support of radio resource management (3gpp ts 36.133 version 12.6. 0 release 12), ETSI TS 136 (133) (2015) V11.

[30] S. Wright, J. Nocedal, et al., Numerical optimization, Springer Science 35 (67-68) (1999) 7.

[31] I. Kezurer, S. Z. Kovalsky, R. Basri, Y. Lipman, Tight relaxation of quadratic matching, in: Computer Graphics Forum, Vol. 34, Wiley Online Library, 2015, pp. 115–128.

[32] A. B. Jambekar, D. I. Steinberg, An implicit enumeration algorithm for the all integer programming problem, Computers & Mathematics with Applications 4 (1) (1978) 15–31.

[33] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews,

M. Dohler, Downlink and uplink cell association with traditional macrocells and millimeter wave small cells, IEEE Transactions on Wireless Communications 15 (9) (2016) 6244–6258.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: