Smart Mode Selection Using Online Reinforcement Learning for VR Broadband Broadcasting in D2D Assisted 5G HetNets

Lei Feng, *Member, IEEE*, Zhixiang Yang, Yang Yang, Xiaoyu Que, and Kai Zhang, *Student Member, IEEE*

***Abstract*—As an emerging broadband service pattern in the 5G era, VR broadcasting needs a considerable amount of bandwidth and strict quality of service (QoS) control. The traditional eMBMS or enTV transmission mode in HetNets consisting of macro cells and small cells cannot bring about a good trade-off between broadband performance and resource utilization for VR broadcasting service. D2D multicasting applied to VR broadcasting can improve the performance of edge users and resource utilization. Motivated by the rapid development of AI techniques, this paper proposes a novel hybrid transmission mode selection based on online reinforcement learning to address this problem. Each VR broadband user can be associated by one of the three modes: macrocell broadcasting, mmWave small cell unicasting and D2D multicasting. This paper first models this intelligent mode decision process as a problem to pursue the optimal system throughput. Then, an online machine learning-based method is proposed to solve this problem, which consists of a fast D2D clustering module based on unsupervised learning and a smart mode selection module based on reinforcement learning. The simulation results verify that the WoLF-PHC and Nash Q-learning perform better than other algorithms in large-scale scenarios and small-scale scenarios, respectively. The proposed intelligent transmission mode selection can also achieve larger VR throughput than traditional broadcasting strategies with a good balance between broadband performance and resource utilization.**

***Index Terms*—5G broadcasting, VR broadband service, D2D multicasting, reinforcement learning, transmission mode selection.**

I. Introduction

As THE rapid development of virtual reality (VR), it will be widely used in various fields, such as education, health care, security and industry. And it also will become the most popular multimedia service. In the 5G era, VR is an important broadcasting service pattern. In contrast to other broadcasting services, VR has a greater need for broadband and its QoS control is also strict, owing to the demand to simultaneously support very large capacity, low latency and ultra-reliability. The network throughput requirement may be approximately hundreds of megabits per second by advanced compression technology [1]. However, current multimedia communication technologies are difficult to meet the above-mentioned strict virtual reality service quality requirements.

Work is ongoing in 3GPP Rel-16 to define an LTE-based solution, known as the evolved multimedia broadcast multicast service (eMBMS) and enhanced TV (enTV), suitable for terrestrial broadcast [2]. The eMBMS is typically combined with single-frequency network technology in which different cellular stations broadcast simultaneously over the same channel. In this way, the same frequency resources can be fully utilized to simultaneously deliver content to multiple requesters, thereby effectively utilizing the network resource [3]. However, the function of eMBMS cannot solve all VR broadcast use cases. The reason is that broadcasting service delay tolerance limit requires that eMBMS can only take its advantages in dense networks.

Although the LTE-A network has indisputable advantages, providing group-oriented services poses a challenge to eMBMS design in LTE-A. It is particularly significant when delivering video content with high bitrate requirements, such as VR. In our view, 5G device to device (D2D) communications are beneficial for multicast delivery schemes in eMBMS networks. When communicating in the local range, the efficiency of using a D2D link is better than that of a regular pass through a base station (BS), which helps alleviate network congestion [4]. In addition, it can reduce base station load, expand base station coverage, and share content through multicast [5]. D2D communication in VR further facilitates faster communication and poses less pressure on the network resources.

Technical scene requirements of VR, such as seamless wide-area coverage and low-power large-scale connectivity, pose challenges for 5G. However, the capabilities of traditional single layer cellular network architectures are inadequate [6]. In order to meet the increasing demand for higher data rates in the 5G VR era, while considering that the efficiency of wireless links is approaching its basic limits, 5G HetNet is proposed as a new type of network, which is a multi-layer network and contains different communication technologies. It consists of nodes with different transmission power and coverage. The

Manuscript received November 29, 2019; revised January 13, 2020; accepted January 20, 2020. Date of publication March 16, 2020; date of current version June 5, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61901054, in part by the Fundamental Research Funds for the Central Universities under Grant 2019RC09, and in part by the State Grid Science and Technology Project "Analysis of Power Wireless Private Network Evolution and 4G/5G Technology Application" under Grant 5700-201941235A-0-0-00. *(Corresponding author: Lei Feng.)*

The authors are with the Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: fenglei@bupt.edu.cn; yangzx@bupt.edu.cn; yangyang2018@bupt.edu.cn; xfn@bupt.edu.cn; kaizhang@bupt.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2020.2977577

multitier HetNet is composed of macro cellular and low-power networks such as mmWave small cellular and D2D clusters in its coverage area [7]. Multitier HetNets can greatly improve the overall throughput of HetNet by offloading macro BSs (MBSs) to low-power BSs [8].

However, the transmission mode selection of VR broadcasting in 5G HetNet still faces enormous challenges [9]. One of the challenges is that the problem is complex and difficult to solve with no prior knowledge since the user state always varies randomly. In this article, VR broadcast users switch among three modes: macrocell broadcasting, mmWave small cell unicasting and D2D multicasting. Whenever the number of users in the system changes or the network status varies, the user transmission mode will also change. Traditional network transmission mode selection methods such as Analytic Hierarchy Process (AHP) algorithm and the game theory algorithm are not suitable for complex and changeable environments, and cannot make timely decisions based on real-time changes in the network environment. In addition, in such a real-time changing environment, it is difficult to determine an accurate system state transition model and calculate the optimal transmission mode strategy through a dynamic programming algorithm. Reinforcement learning (RL) has became a popular research method in several fields to find the optimal strategy [10]. And online reinforcement learning does not need to know the user's behavior in advance, and can quickly find the user's optimal transmission strategy in a real-time changing environment [11]. This is why we choose online reinforcement learning to solve this optimization problem.

In this paper, we propose an intelligent mode selection strategy in D2D assisted 5G HetNet to improve the performance of VR broadcasting. Firstly, we do D2D clustering by the fast D2D clustering algorithm based on unsupervised learning. Secondly, smart mode selection based on reinforcement learning is used to find the optimal transmission strategy. Finally, we evaluate the simulation results. The main contributions and content are summarized as follows:

1) A novel hybrid mode selection scheme is proposed for VR broadcasting in 5G HetNet. To the best of our knowledge, it is the first time that D2D multicasting is utilized and analyzed in VR broadband broadcasting. In this D2D assisted 5G HetNet, the VR users can be served by three modes: marco cell broadcasting, mmWave small cell unicasting and D2D multicasting. D2D multicasting reuses the uplink resource of mmWave small cells through controlling the interference. This proposed scheme improves the performance of edge users and resource utilization since the coverage range of marco cell eMBMS is limited by the edge worst channel-condition users' SINR and the unicasting by small cell consumes considerable resources for broadband service.

2) To select the transmission mode for VR broadband service intelligently and dynamically, we use online reinforcement learning to obtain the optimal decisions among the above three modes for each user. Firstly, the theoretic framework of multi-agent learning is given to model this smart mode selection problem under general-sum stochastic games, whose aim is to maximize the total throughput for VR broadband service. We design a reasonable reward function and Q-function for the optimal VR transmission rate and optimize the VR quality over HetNet. Then, two RL polices, Nash-Q-learning and Wolf-PHC are discussed with a consideration of the network scale. To meet the low latency requirement of online learning, WoLF-PHC algorithm is selected in large-scale scenario based on its low complexity and computational space requirement for a large number of agents.

3) A reasonable performance evaluation is given to compare the proposed strategy with other ones, such as traditional eMBMS and normal hybrid transmission by small and macro cell with heuristic Q-learning-based RL. We compare the performance in several aspects, including convergency and VR broadcasting throughput gain evaluation. In the large-scale network scenario, we adopt the WoLF-PHC scheme and compare it with random pick scheme and simple greedy scheme in performance. While in the small-scale one, we use the Nash Q-learning scheme and compare it with other algorithms such as Deep Q-learning, random pick and best policy. Simulation results indicate that the proposed scheme with WoLF-PHC scheme and Nash-Q-learning perform better in large-scale scenarios and small-scale scenarios, respectively.

The remainder of this paper is organized as follows. A review of the related work is presented in Section II. In Section III, we introduce the system architecture and formulate system optimization problems. Section IV presents D2D clustering algorithm and an online reinforcement learning algorithm to solve this problem. Section V shows the analysis of the simulation results, and Section VI presents some conclusions.

II. Related Work

Several works have been performed on 5G broadcasting. It is an opportunity for broadcasting and multicasting with a dramatic increase in multimedia data traffic. First, the authors in reference [12] analyzed the impatient behavior of broadcast users and proposed a scheduling scheme to guarantee eMBMS QoS in LTE environment. Reference [13] proposed a scheme of using cut-off value from the perspective of modulation and coding, which improved the resource efficiency in 5G broadcast network. An emerging hybrid transmission model for sharing broadcast, multicast, and unicast resource in 5G NR was proposed in reference [14], and compared with eMBMS in 4G LTE, 5G NR provided more possibilities for eMBMS. Reference [7] described a broadband broadcasting system based on uhf band multiplexing. This system has certain advantages in transmission performance, anti-interference and energy consumption, which can satisfy the demands of future 5G network development. The work in reference [15] proposed a scheme to improve the resource efficiency of broadcast multicast service by utilizing the redundant channel of transmitting base station in 5G NR. In the future, the same content services will be provided to a large number of users with VR video demands through wireless networks. The combination of VR and broadcasting can be an effective transmission scheme. VR broadcasting in 5G HetNet can improve the quality of service, and some scholars have conducted relevant studies. In reference [16], the authors studied the communication resource management of VR services. Reference [17] proposed a specified color depth packing method for providing 3D video and VR broadcast services. The research on VR broadcasting in 5G network has just started and VR broadcasting is a field worth further study. It can be considered to combine with other key technologies of 5G to create more possibilities.

As an auxiliary communication technology, D2D can promote spectrum resource utilization and realize the whole system performance enhancement in 5G multimedia network. Reference [18] studied the resource allocation of D2D communication in e-band cellular network, and the proposed resource allocation scheme realized the improvement of system throughput, and considered the interference brought by D2D reusable resource blocks. A D2D auxiliary cooperation framework under 5G heterogeneous network was proposed in [19], and the corresponding spectrum access scheme was studied emphatically. The authors in [20] designed a pre-cache algorithm to realize the D2D assisted distribution of VR video and improve the user experience. Reference [21] proposed a resource allocation algorithm based on D2D caching mechanism which achieves higher energy and spectrum efficiency in multimedia service communication system. D2D serves as the edge computing center for relay to unload traffic to light load. Based on this mode, a joint relay selection scheme was designed in [22]. Reference [23] analyzed how the quality of experience (QoE) and backhaul traffic are affected by the combination of D2D and edge computing in video streaming. In addition, the application of mmWave in broadcasting is also emerging. In order to improve the concurrent performance of mmWave links, reference [24] proposed a cluster-based broadcast scheduling scheme, which improved the average packet transfer rate and throughput. A V2V communication broadcast scheme based on mmWave was proposed in [25] to ensure that each vehicle can obtain sensor data of all other vehicles and reduce broadcast delay. In [26], matched filtering (MF) and partial filtering (PMF) were proposed for the transmission of secret information in dual-receiving millimeter-wave system, and their effectiveness is verified. A multicast scheduling scheme of mmWave small cell called CONMD2D was proposed in [27] to optimize network performance by using concurrent transmission and D2D communication. Summarizing the above work, we can find that broadcasting, VR, D2D communication, mmWave and other technologies have made some progress in their respective directions. However, in the context of 5G heterogeneous network, how to combine these technologies to provide better VR video broadcasting services still needs further research. Different from existing works, we focus on VR broadband broadcasting in D2D assisted 5G HetNet, in which transmission modes like Macro-cell eMBMS, D2D multicasting and mmWave small cell unicasting are jointly considered. Through the above three transmission modes, the coverage of Marco base station will be expanded and the channel quality of its edge users will be improved.

Machine learning algorithm as a solution has been applied widely in resource allocation, resource scheduling and decision control of wireless networks. In reference [28] and [29], deep reinforcement learning algorithm was designed to allocate computing resources reasonably and task migration in mobile edge computing network, ensuring the low latency of communication. In reference [30], the authors used role-critics-based reinforcement learning algorithm to make the most reasonable scheduling of downlink broadcast resources. The algorithm not only promoted the system performance, but also guaranteed the fairness between users. Similarly, machine learning algorithm is also applicable to 5G wireless self-organizing network. A scheme based on machine learning to determine the optimal routing path for the terminal was proposed in [31] in terms of distance and capacity. The VR broadcast wireless network scenarios are complex, and it is unrealistic to determine accurate environmental information in real time. Reinforcement learning method works well on this type of problem because it does not require a specific model and can learn how to improve by interacting with the environment.

In contrast to existing works, we focus on VR broadband broadcasting in D2D-assisted 5G HetNet, in which transmission modes such as macro cell eMBMS, D2D multicasting and mmWave small cell unicasting are jointly considered. Through the above three transmission modes, the coverage of the macro base station is expanded, and the channel quality of its edge users improves. In addition, we use reinforcement learning to determine which transmission mode VR users choose to achieve the maximum improvement of system performance.

III. System Model and Problem Formulation

In this section we describe the proposed model in three parts, the first part shows the D2D assisted 5G HetNet system model details. The second part is the specific expansion of the user association principle. The last part analyzes the optimal system throughput problem based on transmission mode selection.

A. System Model

A VR broadband broadcast transmission system is considered in a 5G HetNet. The whole HetNet consists of two tiers: the first tier is represented by the macro cell, and the second tier includes D2D clusters and mmWave small cells, as shown in Fig. 1. We assume that the HetNet works on discrete time slots with constant duration. Let $\mathbf{N} = \{1, 2, \ldots, N\}$ represent the VR UEs in the whole system. There is one macro cell, $\boldsymbol{k}_1$ D2D clusters and $\boldsymbol{k}_2$ mmWave small cells. We chose $\mathbf{M} = \{1\}$, $\mathbf{D} = \{1, 2, \ldots, \boldsymbol{k}_1\}$, $\mathbf{L} = \{1, 2, \ldots, \boldsymbol{k}_2\}$ to designate the sets of macro cell, D2D clusters and mmWave small cells in the system. The macro cell provides eMBMS, and its coverage range is limited by the edge worst-channel-condition users' SINR. D2D multicasting and mmWave small cell unicasting implement the VR broadcasting coverage extension. The UE may simultaneously be in multiple network coverage areas. However, it dose not receive all the VR signals from each network. The UE tunes into the corresponding channel for data reception when the UE establishes an association relationship with a specific network. The transition of the VR network state over time slots is described by $S(t)$ which is defined as the

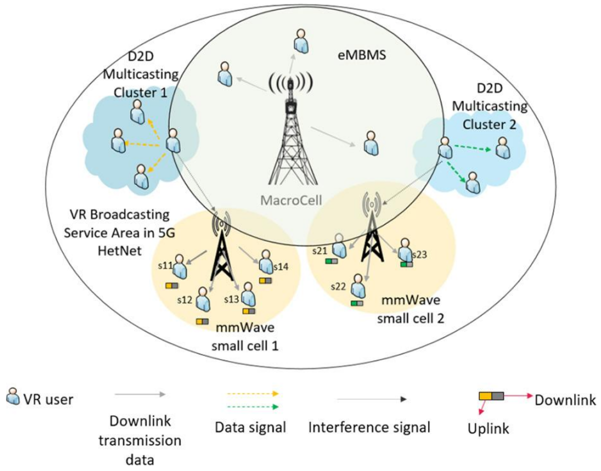

Fig. 1. System model of D2D assisted 5G HetNet.

association. Thus, the network state is represented as follows:

$$S(t) = \{\boldsymbol{s}_1(t), \boldsymbol{s}_2(t), \ldots, \boldsymbol{s}_N(t)\}. \tag{1}$$

Each element in formula (1) represents the association relationship between the user and the three network sets. In the time slot t, we define $\boldsymbol{A}(t) = \{\boldsymbol{a}_1(t), \boldsymbol{a}_2(t), \ldots, \boldsymbol{a}_N(t)\}$ to represent the user selection action in the transmission mode. In the time slot t, the system determines $A(t)$ according to state $S(t)$ and then enters the next state $S(t+1)$.

B. User Association Principle

1) *Macro Cell and mmWave Small Cell:* Macro cells provide eMBMS to UEs in coverage, and UEs associated with macro cells simultaneously receive the same VR broadcast content. Multicast/broadcast services are multi-user services and cannot provide user-specific adaptive parameter configuration, i.e., link adaptive transmission cannot be provided for a single user. Therefore, the coverage of eMBMS, or the data transmission rate that eMBMS can provide depends on the UE with the worst link quality. Among the users who establish an association relationship with the macro cell m, the SINR of the user with the worst link quality is expressed as

$$\boldsymbol{\gamma}_m(s, a) = \min_{n \in \mathbf{N}} \frac{\boldsymbol{G}_{n,m} \boldsymbol{P}_m}{\sigma}, \tag{2}$$

where $\boldsymbol{G}_{n,m}$ denotes the channel gain from the macro BS to the UE n, $\boldsymbol{P}_m$ is the macro BS m broadcast power. Because broadcast channels are only occupied by macro stations, VR users accessing the channels are ideally only affected by additive white Gaussian noise σ in the environment. Further, let $\boldsymbol{B}_m$ denote the broadcast channel bandwidth of the macro cell m, and the data reception rate of each VR UE in the macro cell m is evaluated as

$$\boldsymbol{R}_{n,m}(s, a) = \boldsymbol{B}_m \log_2\bigl(1 + \boldsymbol{\gamma}_m(s, a)\bigr). \tag{3}$$

If the user with inferior channel conditions is also in the coverage range of one or more mmWave cells, the user can consider establishing an association relationship with the nearest mmWave cell, and the mmWave cell base station can provide VR video content to the user through broadband mmWave unicast, which expands the coverage of the macro cell VR broadcast. The SINR of users who choose to access the mmWave small cell unicast network can be calculated as follows:

$$\boldsymbol{\gamma}_{n,l}(s, a) = \frac{\boldsymbol{G}_{n,l} \boldsymbol{P}_l}{\sum_{j \in L/l} \boldsymbol{G}_{n,j} \boldsymbol{P}_j + \sigma}, \tag{4}$$

where $\boldsymbol{G}_{n,l}, \boldsymbol{G}_{n,j}$ is the channel gains from the small cell BS l and other small cell BSs j to the UE n. $\boldsymbol{P}_l, \boldsymbol{P}_j$ separately represents the small cell BS l and other small cell BSs j transmission power. The interference consists of two parts: interference from other mmWave small cell unicast networks and Gaussian white noise interference. The achievable reception rate of VR users in the mmWave small cell l can be expressed as

$$\boldsymbol{R}_{n,l}(s, a) = \boldsymbol{B}_l \log_2\bigl(1 + \boldsymbol{\gamma}_{n,l}(s, a)\bigr), \tag{5}$$

where $\boldsymbol{B}_l$ denotes the transmission bandwidth of the mmWave small cell l.

2) *D2D Cluster:* The SINR of the user who chooses to access the D2D cluster multicast network can be described as,

$$\boldsymbol{\gamma}_{n,d}(s, a) = \frac{\boldsymbol{G}_{n,d} \boldsymbol{P}_d}{\sum_{i \in D/d} \boldsymbol{G}_{n,i} \boldsymbol{P}_i + \sigma}, \tag{6}$$

where $\boldsymbol{G}_{n,d}, \boldsymbol{G}_{n,i}$ are the channel gains of the D2D cluster header to the users. $\boldsymbol{P}_d$ and $\boldsymbol{P}_i$, respectively, represent the transmission power of cluster head d and another D2D cluster head i. Broadband broadcasting can be provided by underlaying all the uplink spectrum resources in the mmWave small cell because the beamforming in massive MIMO can control some interferences. The interference to a D2D cluster mainly comes from other clusters because all the clusters reuse the same uplink resources of the mmWave small cells. We use $\boldsymbol{B}_m$ to describe the D2D clusters' multicast channel bandwidth: therefore, the data reception rate of each VR UE D2D cluster d is evaluated as follows:

$$\boldsymbol{R}_{n,d}(s, a) = \boldsymbol{B}_d \log_2\bigl(1 + \boldsymbol{\gamma}_{n,d}(s, a)\bigr). \tag{7}$$

As a result, the optimal power of D2D headers needs to be calculated because the performance of an uplink in a mmWave small cell cannot degrade much. The main point is to keep the D2D communication from causing exceeding interference. To solve this problem in D2D multicasting user association, the Stackelberg gaming method is introduced as follows:

The mmWave BS is the leader, and the followers are interfering D2D clusters because the former is dominant during the game process. To address interference, the price is set on the received interference, and interference can be coordinated by adjusting the price. The first target is to use the advantage of the mmWave BS most efficiently. Then, D2D pairs compete to maximize their benefits through a non-cooperative game that is based on pricing. The leader and followers are designed to collect the max benefit for themselves owing to their selfishness. By charging followers for interference, the leader gains a benefit. In addition, the leader should limit the interference to a tolerable range because there is a requirement of mmWave small cell users minimum uplink rate $\boldsymbol{R}_{min}^{U}$. Mathematically, the maximization for a leader's utility function can be described as problem P1:

$$\begin{aligned} \text{P1}: \max\ & U_l(c,\boldsymbol{P}_d) \\ s.t.\ & U_l(c,\boldsymbol{P}_d) = c\sum_{d\in D}\boldsymbol{P}_d\boldsymbol{G}_{l,d} \\ & \boldsymbol{R}_{n,l}^{U} \geq \boldsymbol{R}_{min}^{U}, \end{aligned} \tag{8}$$

where c represents the unit interference power pricing factor of, $\boldsymbol{G}_{l,d}$ denotes the channel gain between the BS in mmWave small cell l and the D2D cluster d, $\boldsymbol{R}_{n,l}^{U}$ is the user uplink rate in the mmWave small cell l. According to the price imposed by the leader, each follower regulates its transmission power to pursue the utility value maximization. To fully assess utility, we must consider not only the benefits that followers gain from communication but also the costs of interference to the leader. The optimization problem for each follower is described by P2:

$$\begin{aligned} \text{P2}: \max\ & U_d(c,\boldsymbol{P}_d) \\ s.t.\ & U_d(c,\boldsymbol{P}_d) = \boldsymbol{R}_{n,d}(s,a) - c\boldsymbol{P}_d\boldsymbol{G}_{l,d} \\ & 0 \leq \boldsymbol{P}_d \leq \boldsymbol{P}_{max}^{D}, \end{aligned} \tag{9}$$

where $\boldsymbol{P}_{max}^{D}$ represents the maximum transmission power of D2D cluster header and the utility function $U_d(c,\boldsymbol{P}_d)$ of D2D cluster n as a follower consists of two parts: the cluster broadcasting data rate $\boldsymbol{R}_{n,d}(s,a)$ in (10) and the expense paid to the mmWave small cell BS due to interference.

$$\boldsymbol{R}_{n,d}(s,a) = \boldsymbol{B}_d\log_2\bigl(1+\boldsymbol{\gamma}_{n,d}(s,a)\bigr). \tag{10}$$

Higher transmission power increases the broadcast data rate and the mmBS interference, which increases the cost. Therefore, for the follower, the received data rate and communication cost are considered comprehensively, and the best compromise is achieved through appropriate transmission power planning to maximize the utility value. From above, in the process of the game, the benefits of mmWave small cell and D2D clusters are considered, not only unilateral benefit but also to eventually obtain a better benefit balance. The Stackelberg equilibrium (SE) is a steady result, and there is no motivation for any participant to diverge. The SE in this pattern can be expressed as follows:

$$\begin{aligned} \boldsymbol{U}_l\bigl(c^*,\boldsymbol{P}_d^*\bigr) &\geq \boldsymbol{U}_d\bigl(c,\boldsymbol{P}_d^*\bigr) \\ \boldsymbol{U}_d\bigl(c^*,\boldsymbol{P}_d^*\bigr) &\geq \boldsymbol{U}_d\bigl(c^*,\boldsymbol{P}_d\bigr), \end{aligned} \tag{11}$$

The inverse induction can solve this type of Stackelberg game. Assuming c is known, the ultimate result of followers via a non-cooperative game is a Nash equilibrium (NE). Theorem 1 is proposed and proven with regard to the followers' game.

Theorem 1: The followers in P2 have a unique NE, and the optimal transmission power $\boldsymbol{P}_d^*$ in the NE follows:

$$\boldsymbol{P}_d = \frac{1}{\ln 2\cdot c\cdot\boldsymbol{G}_{l,d}} - \frac{\sum_{i\in D/d}\boldsymbol{P}_i\boldsymbol{G}_{n,i}+I_n}{\boldsymbol{G}_{n,d}}. \tag{12}$$

Proof: According to game theory, if the pure policy space of each participant in Euclidean space is a nonempty compact convex set and the utility function is consecutive, then there must be an NE in an n-person policy game. In the HetNet introduced in this paper, each D2D cluster transmission power satisfies the maximum transmission power constraint $0 \leq \boldsymbol{P}_d \leq \boldsymbol{P}_{max}^{D}$. Clearly, the policy space is a compactly convex nonempty set. The continuity of the utility function is satisfied. In the game policy space of all followers, there must be an NE policy.

First, we solve the second derivative of the objective function $\boldsymbol{U}_d(c,\boldsymbol{P}_d)$ and find that the second derivative is constant negative. Then, we can conclude that the objective utility function is strictly concave. A unique maximum exists in a bounded closed convex set $[0,\boldsymbol{P}_{max}^{D}]$. Hence, the NE, as described by the theorem exists and is unique. The first-order condition is used to obtain the optimal reaction function, as shown in the formula (12). ■

Then, the convergence of (12) is tested using Banach the contraction theorem. Substituting (12) into P1, we can also obtain the optimal price factor c^*. From above, we can use an easy iterative adjustment scheme to find the final converged optimal solution $\boldsymbol{P}_d^*$.

C. Problem Formulation

In different transmission modes, VR users have different signal receiving rates, which affects the overall throughput of the system. Therefore, the system should select the appropriate transmission mode for the user. Specifically, the user chooses an action $a(t)$ and enters the next state $s(t+1)$ according to the entire system state in the time slot t. In this section, we formulate the total throughput of the VR broadcast system in the above 5G heterogeneous network as follows:

$$\begin{aligned} \boldsymbol{R}_{sum}(s,a) = \sum_{n=1}^{N}\Biggl\{&\sum_{m\in\mathbf{M}}\boldsymbol{R}_{n,m}(s,a) + \sum_{d\in\mathbf{D}}\boldsymbol{R}_{n,d}(s,a) \\ &+ \sum_{l\in\mathbf{L}}\boldsymbol{R}_{n,l}(s,a)\Biggr\}. \end{aligned} \tag{13}$$

The total broadcast throughput of the system consists of three parts: macro cell eMBMS, D2D multicasting and mmWave small cell unicasting. The problem in this paper is to solve the overall maximum throughput of the VR broadcast system by employing an optimal transmission mode selection strategy. The formula is as follows:

$$\begin{aligned} \max_{\pi\in\Pi}\ & \boldsymbol{R}_{sum}(\pi(s,a)) \\ s.t.\ & \boldsymbol{\gamma}_{n,m},\boldsymbol{\gamma}_{n,d},\boldsymbol{\gamma}_{n,l}(s,a) \geq \gamma_{min} \\ & \boldsymbol{P}_m \leq P_{m,max} \\ & \boldsymbol{P}_d \leq P_{d,max} \\ & \boldsymbol{P}_l \leq P_{l,max}, \end{aligned} \tag{14}$$

where $\boldsymbol{\gamma}_{min}$ represents the minimum SNR requirement of the VR users' QoS. The transmission power of the macro BS, the D2D cluster head, and the mmWave small cell BS cannot exceed the maximum transmission power limit.

IV. Transmission Mode Selection Using Online Reinforcement Learning

This section describes the process of achieving overall maximum throughput through the proposed online reinforcement learning method. We divide whole process into two parts, D2D clustering is the first part, the second is the transmission mode selection via online reinforcement learning.

A. D2D Clustering

This section describes the clustering process of D2D by using the rapid clustering method based on speedy search and discovery of density peaks. First, we need to select the D2D cluster centre and then classify the remaining D2D devices according to the D2D cluster centre.

Consider the set of D2D waiting to cluster $\mathcal{D} = \{d_i\}_{i=1}^{D}$, $l_{i,j} = ||d_i - d_j||_2$ represents the distance between VR UEs d_i and d_j. For any point d_i in $\mathcal{D}$, ρ_i and δ_i can be defined to describe the characteristics of the point. Finally, the preliminary clustering results are divided into D2D-cluster cores and D2D-cluster halos.

Definition 1: Local Density ρ_i.

$$\rho_i = \sum_{j \neq i} e^{-\left(\frac{l_{i,j}}{l_c}\right)^2}, \tag{15}$$

where $l_{i,j}$ represents the Euclidean distance between VR UEs d_i and d_j, and $l_c > 0$ represents the cutoff distance.

Definition 2: Minimum Distance δ_i.

$$\delta_{\kappa_i} = \begin{cases} \min_{\kappa_j, j<i}\left\{l_{\kappa_i,\kappa_j}\right\}, & i \geq 2, \\ \max_{j\geq 2}\left\{\delta_{\kappa_j}\right\}, & i = 1, \end{cases} \tag{16}$$

where $\{\kappa_i\}_{i=1}^{D}$ denote a descending subscript of $\{\rho_i\}_{i=1}^{D}$, i.e., $\rho_{\kappa_1} \geq \rho_{\kappa_2} \geq \cdots \geq \rho_{\kappa_D}$.

To ensure that the different D2D clusters do not overlap, the D2D cluster centres should be as far apart as possible. Therefore, we believe that the D2D-cluster centre has a relatively larger density and is surrounded by lower density neighbours. In addition, the D2D-cluster centre should be far from other points with higher density.

According to (16), when d_i has the highest local density, δ_i represents the maximum distance between other VR UEs and d_i, otherwise δ_i represents the minimum distance between d_i and the D2D devices whose local density is greater than d_i. The local density ρ_i and minimum distance δ_i are simultaneously maximized to ensure that d_i is selected as a D2D-cluster centre. Note that δ_i is much larger than the typical nearest neighbour distance merely for points that are local or global maximum in the density.

To automatically select the D2D-cluster centre, considering the local density ρ value and the minimum distance δ value, we definite a target γ_i for determining the number of D2D-cluster centres.

$$\gamma_i = \hat{\rho}_i \cdot \hat{\delta}_i, \quad i = 1, 2, \ldots, D, \tag{17}$$

where $\hat{\rho}$ represents the z-score normalization of ρ and $\hat{\delta}$ is similar.

Obviously, the larger γ_i, the more likely d_i is the D2D-cluster centrum. Thus, we only need to sort $\{\gamma_i\}_{i=1}^{D}$ in descending order and then select several points from the beginning as the D2D-cluster centre.

Calculate the mean value μ and the variance σ^2 of γ_i according to the (18).

$$\mu = \frac{\sum_{i=1}^{D} \gamma_i}{D}, \quad \sigma^2 = \frac{\sum_{i=1}^{D} (\gamma_i - \mu)^2}{D}. \tag{18}$$

Then, for each γ_i, their respective probability densities are calculated according to the (19).

$$P(\gamma_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\gamma_i - \mu)^2}{2\sigma^2}\right). \tag{19}$$

Finally, if $P(\gamma_i) < \epsilon$ and $\gamma_i > \mu$, then the point n_i is considered to be the D2D-cluster centre, where ϵ is a very small positive integer.

To distinguish outliers and noises, we introduce the concept of local density within clusters to divide the D2D-cluster core and the D2D-cluster halo from the preliminary D2D-cluster.

$$\varrho_i = \sum_{j,\varphi_i=\varphi_j}^{D} I\left(l_c - l_{i,j}\right), \tag{20}$$

where $I(\bullet)$ define an indicator function. ϱ_i indicates the number of points in the same cluster, which is located less than l_c from point i. If ϱ_i is less than a fixed constant, then d_i is considered as an outlier.

The pseudo-code for the D2D clustering strategy is shown in Algorithm 1.

B. Online Reinforcement Learning

In the reinforcement learning model, each user is actually an intelligent agent taking actions based on state and reward in time slot t. After taking actions according to policy, system moves to time slot $t + 1$ and user changes its state. Our model is listed as below.

State: The whole VR network state set is defined in (1), because each user in the VR network takes strategy independently.

Action: The whole VR user action set is defined as $\boldsymbol{A}(t) = \{\boldsymbol{a}_1(t), \boldsymbol{a}_2(t), \ldots, \boldsymbol{a}_N(t)\}$ in time slot t, where a_i is the action referring to state.

Policy: The strategy selection probability of the user i in the time slot t is defined as $\pi_i(t)$. The policy set $\pi(t) = \pi_1(t), \ldots, \pi_N(t)$. The probability policy is used to adopt actions at the start of each iteration. The transition function is defined on the basis of the strategy selection probability given as follows:

$$P\left(s, s', a\right) = P\left(s(t+1) = s' | s(t) = s, a(t) = a\right). \tag{21}$$

Reward: The reward function is interrelated to the state and action of all the VR users in the network and is defined in (13). Furthermore, we define $r(s^t, a^t)$ as the $\boldsymbol{R}_{sum}(s, a)$ of the user i in the time slot t.

Algorithm 1: Improved Fast Search and Density Peaks Identification Algorithm

```
Input: Given the parameter $t \in (0, 1)$ used to determine
       the cutoff distance $l_c$, constant $\eta$ and the
       coordinates of all D2D devices.
Output: The serial number of D2D-cluster which point
        belongs to.
Calculate the distance $l_{i,j}$ and let $l_{i,j} = l_{j,i}, i < j$.
Sort the distance set $\{l_{i,j}\}$ in ascending order to obtain the
sequence $l_1 \leq l_2 \leq \cdots \leq l_M, M = \frac{D(D-1)}{2}$. Calculate the
cutoff distance $l_c = l_{\lfloor M \cdot t \rfloor}$.
Calculate $\{\rho_i\}_{i=1}^{D}$ from (15) and generate its descending
order subscript $\{\kappa_i\}_{i=1}^{D}$.
for $i \leftarrow 2 : D$ do
    $\delta_{\kappa_i} \leftarrow \max_{i<j}\{l_{i,j}\}$.
    for $j \leftarrow 1 : i - 1$ do
        if $l_{\kappa_i,\kappa_j} < \delta_{\kappa_i}$ then
            $\delta_{\kappa_i} \leftarrow d_{\kappa_i,\kappa_j}$.
            $\psi_{\kappa_i} \leftarrow \kappa_j$.
$\delta_{\kappa_1} \leftarrow \max_{j\geq 2}\{\delta_{\kappa_j}\}$.
Select the D2D-cluster centre $\{\phi_j\}_{j=1}^{g}$ and set $f \leftarrow 1$.
for $i \leftarrow 1 : D$ do
    if $x_i \in \{\phi_j\}_{j=1}^{g}$ then
        $\varphi_i \leftarrow f$.
        $f \leftarrow f + 1$.
    else
        $\varphi_i \leftarrow -1$.
for $i \leftarrow 1 : D$ do
    Calculate $\varrho_i$ based on (20).
    if $\varphi_i = -1$ then
        if $\varrho_i >= \eta$ then
            $\varphi_i = \varphi_{\phi_{\kappa_i}}$.
        else
            $\varphi_i = 0$.
```

In multi-agent reinforcement learning, we consider a joint reward maximization for player i as

$$v_i(s, \pi) = \sum_{t=0}^{\infty} \beta^t E(r_i|\pi, s), \tag{22}$$

where β is the discount factor, s' is the next time $t + 1$ state s^{t+1}. In Q-learning, function Q is defined as

$$Q(s, a) = r\big(s^t, a\big) + \beta \sum_{s'} P\big(s, s', a\big)v\big(s', \pi\big). \tag{23}$$

Q-learning provides us with a simple Q value update as follows:

$$\begin{aligned} Q^{t+1}(s, a) = (1 - \alpha_t)Q^t\big(s^t, a^t\big) \\ + \alpha_t\Big[r\big(s^t, a^t\big) + \beta \max_a Q^t\Big(s^{t+1}, a\Big)\Big], \end{aligned} \tag{24}$$

Algorithm 2: Nash Q-Learning Algorithm

```
Require: initial state $s_0, t, rw_{min}, ru_{min}, N_u$
Ensure: performance of the mmWave small cell uplink,
unicasting resources
if $(N_u > 0)$ then
    $t \leftarrow 0$ for $s \in S$ do
        for $j \leftarrow 1 : N$ do
            $Q_j^t(s, a_1, ..., a_N) \leftarrow 0$
while true do
    $a_i(t) \leftarrow \Pi_i(s)$,  $s_{t+1} \leftarrow P(s, a_i(t))$ update
    $Q_j^{t+1}(s, a_1, ..., a_N)$  according to (27) $t \leftarrow t + 1$
```

where α_t is the learning rate in the time slot t. We assume that all agents in our model are rational and convergent in the game. A Nash equilibrium π^* is a united strategy in which the strategy of each agent is the optimum response to other agents. The Nash equilibrium meets the following requirements:

$$\begin{aligned} v^i\big(s, \pi_1^*, \ldots, \pi_N^*\big) \geq v^i\big(s, \pi_1^*, \ldots, \pi_{i-1}^*, \pi_i, \pi_{i+1}^*, \ldots, \pi_N^*\big) \\ \forall \pi_i \in \Pi, \end{aligned} \tag{25}$$

where Π is the available policy set of agent i.

To meet the low latency requirement of online learning, WoLF-PHC algorithm is selected based on its low complexity and computational space requirement. However, in small-scale scenario, Nash Q-learning converges faster than WoLF-PHC and achieves almost as good performance as WoLF-PHC. Therefore, our online reinforcement learning is based on these two algorithms and we adopt WoLF-PHC in large-scale scenario and Nash Q-learning in small-scale scenario.

1) *Nash Q-learning:* In Nash Q-learning, agents observe not only their own reward but also other agents' in the model. The Nash Q-learning algorithm is shown in Algorithm 2. The Q value function is defined as $(s, a_1, \ldots, a_N)$. All agents are assumed to follow the joint NE strategy. Agent i's Nash Q-function is defined as follows:

$$\begin{aligned} Q_i^*(s, a_1, \ldots, a_N) = r(s, a_1, \ldots, a_N) + \beta \sum_{s' \in S} P\big(s', a_1, \ldots, a_N\big) \\ \times v_i\big(s, \pi_1^*, \ldots, \pi_N^*\big), \end{aligned} \tag{26}$$

where $(\pi_1^*, \ldots, \pi_N^*)$ is the joint Nash equilibrium strategy, $r(s, a_1, \ldots, a_N)$ is reward in state s based on joint action $a_1, \ldots, a_N$. $v_i(s, \pi_1^*, \ldots, \pi_N^*)$ is the total discount reward defined in (25). Nash Q-learning updates Q-value according to

$$\begin{aligned} Q_i^{t+1}(s, a_1, \ldots, a_N) = (1 - \alpha_t)Q_i^t(s, a_1, \ldots, a_N) \\ + \alpha_t\big[r(s, a_1, \ldots, a_N) + \beta NashQ_i^t\big(s'\big)\big]. \end{aligned} \tag{27}$$

We have

$$NashQ_i^t\big(s'\big) = \pi_1\big(s'\big) \cdots \pi_N\big(s'\big) \cdot Q_i^t\big(s'\big). \tag{28}$$

Theorem 2: Nash Q-learning updates Q-value updated by formula (27):

$$Q_i^{t+1}(s, a_1, \ldots, a_N) = (1-\alpha_t)Q_i^t(s, a_1, \ldots, a_N) + \alpha_t\big[r(s_i, a_1, \ldots, a_N) + \beta NashQ_i^t(s')\big], \tag{29}$$

converges to Q_* with probability 1.

Proof: The learning rate α satisfies $\sum_t \alpha_t(s,a) = \infty, \sum_t \alpha_t^2(s,a) < \infty$. We define the mapping $H^t : R \rightarrow R$, where

$$H^t(s, a_1, \ldots, a_N)Q_i^t(s, a_1, \ldots, a_N) = r(s, a_1, \ldots, a_N) + \beta NashQ_i^t(s'). \tag{30}$$

Then, the iteration is define by,

$$Q_i^{t+1} = (1-\alpha_t)Q_i^t + \alpha_t\big(H^t Q_i^t\big). \tag{31}$$

According to [32], [33], if there is a number $0 < \beta < 1$, $\|H^t Q_i^t - H^t Q_i^*\| \le \beta\|Q_i^t - Q_i^*\|$ and $Q_i^* = E[H^t Q_i^*]$, the formula (27) converges to Q^* w.p. 1. Now, we proceed to prove the two conditions. Based on formula (23), we have

$$\begin{aligned} Q_i^*(s, a_1, \ldots, a_N) &= r(s, a_1, \ldots, a_N) + \beta \sum_{s' \in S} P\big(s', a_1, \ldots, a_N\big) v_i\big(s, \pi_1^*, \ldots, \pi_N^*\big) \\ &= \sum_{s' \in S} P\big(s', a_1, \ldots, a_N\big)\big[r(s_k, a_1, \ldots, a_n) + \beta v_i\big(s, \pi_1^*, \ldots, \pi_N^*\big)\big] \\ &= E\big[H^t Q_i^*(s, a_1, \ldots, a_N)\big], \end{aligned} \tag{32}$$

for all s, a. Thus $Q_i^* = E[H^t Q_i^*]$.

We assume that we can find a global optima or saddle point in every state game. For $\|H_i^t Q_i^t - H_i^t Q_i^*\| \le \beta\|Q_i^t - Q_i^*\|$, we have

$$\begin{aligned} \big\|H_i^t Q_i - H^t Q_i^*\big\| &= \max_i \Big|H^t Q_i - H^t Q_j^*\Big|_i \\ &= \max_i \max_s \big|\beta\pi_1(s)\cdots\pi_N(s)Q_i(s) - \beta\pi_1^*(s)\cdots\pi_N^*(s)Q_i^*(s)\big| \\ &= \max_i \beta\big|\pi_1(s)\cdots\pi_N(s)Q_i(s) - \pi_1^*(s)\cdots\pi_N^*(s)Q_i^*(s)\big|. \end{aligned} \tag{33}$$

We proceed to prove that

$$\big|\pi_1(s)\cdots\pi_N(s)Q_i(s) - \pi_1(s)\cdots\pi_N^*(s)Q_i^*(s)\big| \le \big\|Q_i(s) - Q_i^*(s)\big\|. \tag{34}$$

We replaced the symbol in the original formula with a new symbol to make the formula more concise. The above formula can be expressed as

$$\Big|\sigma_i\sigma_{-i}Q_i(s) - \sigma_j^*\sigma_{-i}^*Q_i^*(s)\Big| \le \big\|Q_i(s) - Q_i^*(s)\big\|. \tag{35}$$

Case 1: Assume $(\sigma_1, \ldots, \sigma_N)$ and $(\sigma_1^*, \ldots, \sigma_N^*)$ are global optimal points.

If $\sigma_i\sigma_{-i}Q_i(s) \le \sigma_i^*\sigma_{-i}^*Q_i^*(s)$, we have

$$\begin{aligned} &\sigma_i\sigma_{-i}Q_i(s) - \sigma_i^*\sigma_{-i}^*Q_i^*(s) \\ &\quad\le \sigma_i\sigma_{-i}Q_i(s) - \sigma_i\sigma_{-i}Q_i^*(s) \end{aligned}$$

$$\begin{aligned} &= \sum_{a_1 \ldots a_N} \sigma_1(a_1)\ldots\sigma_N(a_N)\big(Q_i(s, a_1, \ldots, a_N) - Q_i^*(s, a_1, \ldots, a_N)\big) \\ &\le \sum_{a_1 \ldots a_N} \sigma_1(a_1)\ldots\sigma_n(a_N)\big\|Q_i(s) - Q_i^*(s)\big\| \\ &= \big\|Q_i(s) - Q_i^*(s)\big\|. \end{aligned} \tag{36}$$

If $\sigma_i\sigma_{-i}Q_i(s) \le \sigma_i^*\sigma_{-i}^*Q_i^*(s)$, then

$$\sigma_i^*\sigma_{-i}^*Q_i^*(s) - \sigma_i\sigma_{-i}Q_i(s) \le \sigma_i^*\sigma_{-i}^*Q_i(s) - \sigma_i^*\sigma_{-i}^*Q_i(s), \tag{37}$$

and the rest of the proof is similar to the above.

Case 2: Suppose NE are saddle points, if $\sigma_i\sigma_{-i}Q_i(s) \ge \sigma_i^*\sigma_{-i}^*Q_i^*(s)$, we have

$$\begin{aligned} \sigma_i\sigma_{-i}Q_i(s) - \sigma_i^*\sigma_{-i}^*Q_i^*(s) &\le \sigma_i\sigma_{-i}Q_i(s) - \sigma_i\sigma_{-i}^*Q_i^*(s) \\ &\le \sigma_i\sigma_{-i}^*Q_i(s) - \sigma_i\sigma_{-i}^*Q_i^*(s) \\ &\le \big\|Q_i(s) - Q_i^*(s)\big\|. \end{aligned} \tag{38}$$

If $\sigma_i\sigma_{-i}Q_i(s) \le \sigma_i^*\sigma_{-i}^*Q_i^*(s)$, a similar proof applies. Thus

$$\begin{aligned} \big\|H^t Q - H^t Q^*\big\| &\le \max_i \max_s \beta\big|\pi_1^*(s)\cdots\pi_N(s)Q_i(s) - \pi_N^*(s)Q_i^*(s)\big| \\ &\le \beta\big\|Q - Q^*\big\|. \end{aligned} \tag{39}$$

Both conditions have been proven; therefore, the updated formula (27) converges to Nash Q-Values Q^*. ■

(2) *WoLF-PHC:* Due to the large number of agents, the computational complexity grows exponentially. Therefore, we adopt the WoLF-PHC scheme. The Q-value update formula of WoLF-PHC is defined as follows,

$$\begin{aligned} Q_i^{t+1}(s_i, a_i) &= \big(1-\alpha^t\big)Q_i^t(s_i, a_i) + \alpha^t\Big(r\big(s_i^t, a_i\big) + \gamma \max_{a'} Q_i\big(s', a'\big)\Big), \\ &\quad \textit{for} \quad i = 1, \ldots, \mathbf{N}, \end{aligned} \tag{40}$$

where α is the learning rate, ranging from 0 to 1. γ is the discount coefficient indicating the importance of reward in the following time slots.

To further improve the joint cooperation efficiency, WoLF-PHC defines $C(s)$ as the number of states s that has appeared during the Q-value update and applies estimate policy π' to record and adjust policy π, which is

$$\pi_i'(s, a_i) \leftarrow \pi_i'(s, a_i) + \frac{1}{C(s)}\big[\pi_i(s, a_i) - \pi_i'(s, a_i)\big]. \tag{41}$$

After the estimate policy π' updates, it is compared with π. If $\sum_{a_i \in \mathbf{A}_i} \pi_i(s, a_i)Q_i(s, a_i) > \sum_{a_i \in \mathbf{A}_i} \pi_i'(s, a_i)Q_i(s, a_i)$, π is regarded as the winner. Otherwise, π' is regarded as the winner. As is shown in Algorithm 3, the agent learns quickly when losing and slowly when winning.

The online reinforcement learning for mode selection based on VR broadcasting is shown in Algorithm 3. N_u is the active user number in the model. rw_{min} reflects the mmWave small cell uplink performance threshold and ru_{min} reflects the unicasting resources wasting performance threshold. r_i under these two values is regarded as 0. s' is the state when the agent in the state s takes the action a_i.

However, as far as we know, the convergence of WoLF-PHC hasn't been proved theoretically. WoLF-PHC learns by only

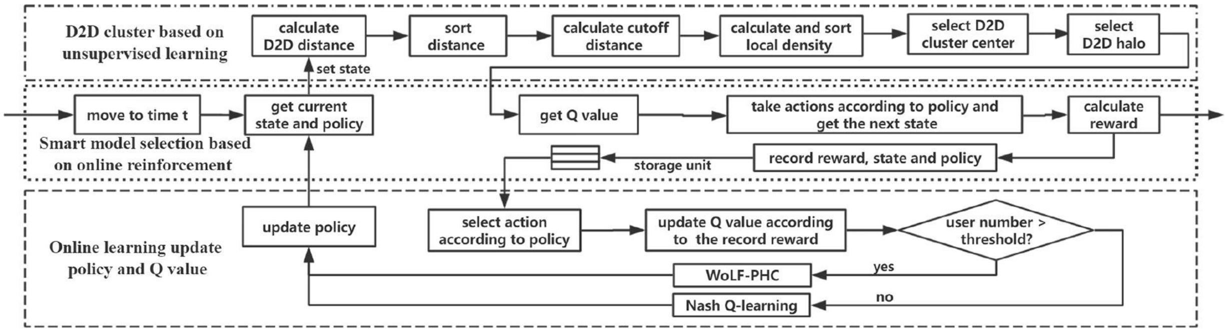
D2D cluster based on unsupervised learning
calculate D2D distance
sort distance
calculate cutoff distance
calculate and sort local density
select D2D cluster center
select D2D halo
set state
move to time t
get current state and policy
get Q value
take actions according to policy and get the next state
calculate reward
Smart model selection based on online reinforcement
storage unit
record reward, state and policy
Online learning update policy and Q value
update policy
select action according to policy
update Q value according to the record reward
user number > threshold?
yes
no
WoLF-PHC
Nash Q-learning

Fig. 2. D2D cluster and hybrid transmission mode selection strategy.

Algorithm 3: Win or Learn Fast Policy Hill Climbing Algorithm

```
Require: $\delta_l, \delta_w, s, t, rw_{min}, ru_{min}, N_u$
Ensure: performance of the mmWave small cell uplink,
unicasting resource wasting.
if $(N_u > 0)$ then
    for $i \leftarrow 1 : N$ do
        $Q_i(s, a_i) \leftarrow 0,\ \ \pi_i(s, a_i) \leftarrow \frac{1}{|\mathbf{A}|}\ \pi'(s, a_i) \leftarrow \frac{1}{|\mathbf{A}|},$
        $\delta_l < \delta_w,\ \ C(s) \leftarrow 0$
while true do
    $a_i(t) \leftarrow \Pi_i(s),\ r_i \leftarrow r(s, a_i(t)),\ s' \leftarrow P(s, a_i(t))$
    $Q_i(s, a_i(t)) \leftarrow$
    $Q_i(s, a_i(t)) + \alpha[r_i + \gamma \max_{a'} Q_i(s', a') - Q_i(s, a_c)]$ if
    $(r_i < \max(rw_{min}, ru_{min}))$ then
        $r_i \leftarrow 0$
    for $i \leftarrow 1 :\ \mid A \mid$ do
        $C(s) \leftarrow C(s) + 1$
        $\pi_i'(s, a_i) \leftarrow \pi_i'(s, a_i) + \frac{1}{C(s)}[\pi_i(s, a_i) - \pi_i'(s, a_i)]$ if
        $(\sum_{a_i \in \boldsymbol{A}} \pi_i(s, a_i) Q_i(s, a_i) >$
        $\sum_{a_i \in \boldsymbol{A}_i} \pi_i'(s, a_i) Q_i(s, a_i))$ then
            $\delta \leftarrow \delta_w$
        else
            $\delta \leftarrow \delta_l\ \delta_{sa} \leftarrow min(\pi_i(s, a_i),\ \frac{\delta}{|\mathbf{A}|-1})$
        if $(a_i \neq argmax_{a'} Q(s, a'))$ then
            $\xi_{sa} \leftarrow -\delta_{sa}$
        else
            $\xi_{sa} \leftarrow \sum_{a' \neq a} \delta_{sa'}$
        $\pi_i(s, a_i) \leftarrow \pi_i(s, a_i) + \xi_{sa}$
    $t \leftarrow t + 1\ s \leftarrow s'$
```

changing learning rate, indicating that the algorithm remains rational. In the next section, empirical results show that WoLF-PHC converges to an equilibrium in self-play with multi-agent model.

TABLE I
ENVIRONMENT SIMULATION PARAMETERS

Simulation Parameters	Value
Number of user	50 : 2 : 200
Number of D2D cluster $k1$	5
Number of small cell $k2$	5
Broadcasting bandwidth B_m	100 MHz
mmWave small cell total bandwidth B_l	1 GHz
Power of macro cell P_m	43 dBm
Power of D2D cluster P_d	26 dBm
Power of small cell P_l	33 dBm
Power of AWGN σ	-174 dBm
Unicasting resource performance rw_{min}	0.05 Gbps
Small cell uplink performance ru_{min}	0.05 Gbps

TABLE II
LEARNING APPROACH PARAMETERS

Algorithm	Learning approach Parameters	Value
Deep Q-Network	Learning rate α	0.1
	Greedy factor ϵ	0.1
	Discount factor γ	0.9
	Batch size	20
	Experience replay memory number	5000
Nash Q-Learning	Learning rate α	0.1
	Greedy factor ϵ	0.1
	Discount factor γ	0.9
WoLF-PHC	Learning rate α	0.1
	Win rate δ_w	0.01
	Lose rate δ_l	0.001

V. SIMULATION RESULTS AND ANALYSIS

We simulate our proposed smart mode selection according to Fig. 2 to prove its ability to improve the system throughput. The simulation environment is a 5G HetNet, which consists of macro cell, small cells and D2D clusters. The macro BS locates in the centre of the region, while the small BSs and D2D clusters are distributed at the edge of the macro cell. The channel gains are determined by the physical distance between the VR user and the broadcast VR signal transmitting point. Table I lists the values of the main simulation parameters. We compare the performance of our WoLF-PHC and Nash Q-learning scheme with other algorithms, such as simple greedy algorithm, stochastic algorithm and Deep Q-learning. Offline training number is assumed to be 2000, constraining that the convergence of WoLF-PHC cannot be too slow. Learning rate of these algorithms is all set to the same value of 0.1. Therefore, learning approach parameters are shown in Table II. In the simple greedy algorithm, each user chooses its best policy itself. In the stochastic algorithm, each user randomly choose an action. The Nash Q-learning and deep Q Network all need $|S| \cdot |A|^N$ space to maintain a joint Q-value

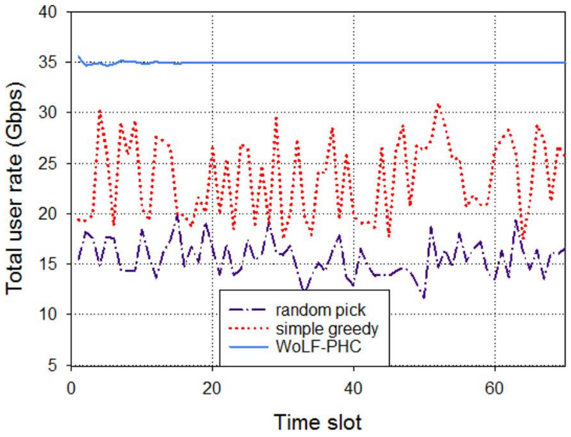

Fig. 3. Online optimization performance on total user rate of different algorithms vs. time slot (200 agents).

table, which means that these two algorithms are not suitable in large-scale scenario. WoLF-PHC costs just $N \cdot |S| \cdot |A|$, indicating that it is more efficient in large-scale user scenarios. Therefore, WoLF-PHC, simple greedy and stochastic algorithm are simulated when user number is 200 and WoLF-PHC, simple greedy, stochastic, Nash Q-learning and Deep Q-learning are simulated when user number is 5.

The scheme is evaluated from four aspects: online optimization performance on total user rate, empirical cumulative distribution function (CDF) of user rate, training performance on total user rate of different learning approach parameters and system bandwidth efficiency. In broadcasting model, we define bandwidth efficiency ε as, $\varepsilon = S/(B_m + B_l)$, where S means the system throughput, B_m means the broadcasting bandwidth, B_l means the total small cell bandwidth.

As shown in Fig. 3, in large-scale scenario with 200 agents, the WoLF-PHC algorithm achieves higher system throughput compared with the simple greedy and stochastic algorithms. This shows that WoLF-PHC can improve the service quality of VR users and create a better user experience. In our simulation model, WoLF-PHC achieves performance nearly 50% higher than simple greedy algorithm. Because the proposed scheme can dynamically select a better transmission scheme for users according to the state of the system. Compared with WoLF-PHC, user in simple greedy algorithm chooses mode only based on its own information.

Fig. 4 presents the system throughput under different policies in small-scale scenarios, where there are only 5 users. When the number of users is small, WoLF-PHC, deep Q network and Nash Q-learning can also achieve good performance while the random algorithm fluctuates greatly. However, Deep Q Network's convergence is still poor. Simple greedy algorithm often performs well in small-scale scenario due to the reason that the impact between users is relatively small. However, in simple greedy algorithm, bad user may cause significant system performance degradation. The empirical CDF plots in Fig. 5 presents the data rate for our proposed hybrid transmission mode with different algorithms in large-scale scenario. In Fig. 5, the user rate of WoLF-PHC is

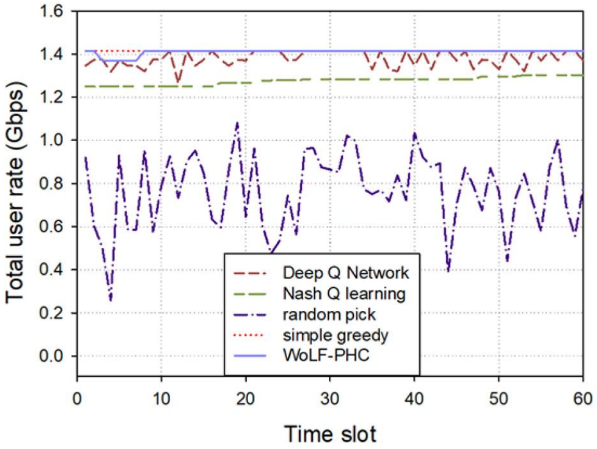

Fig. 4. Online optimization performance on total user rate of different algorithms vs. time slot (5 agents).

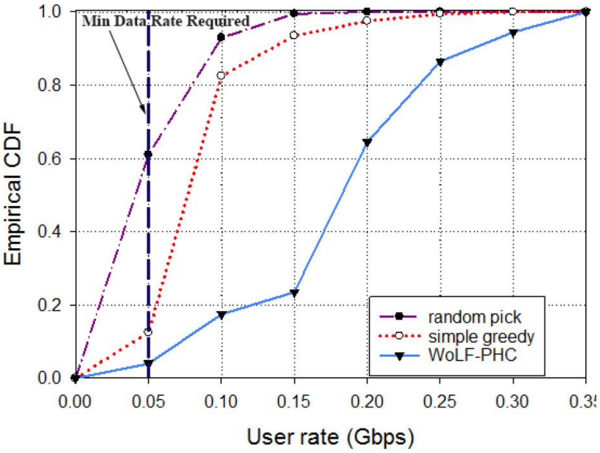

Fig. 5. Empirical CDF on user rate of different algorithms (200 agents).

generally higher than random pick and simple greedy algorithms. The minimum data rate requirements considered in this evaluation is 0.05 Gbps for users. We discover that WoLF-PHC algorithm keep the ratio of user rate lower than min data rate required is less than 5%. However, this ratio in random pick and simple greedy are 61% and 14%. This is due to the fact that WoLF-PHC can find a more suitable and optimal transmission mode for each user, thereby improving the user rate and reducing the number of users with low rate.

The hyper parameters evaluation results of the proposed WoLF-PHC algorithm for large-scale user scenarios is given in Fig. 6. First of all, we can find that the total user rate eventually convergence after a certain number of iterations. Obviously, the size of the convergence value and convergence rate are influenced by the win rate and lost rate. The convergence speed increases with the increase of the two hyper parameters, while the convergence value of the total user rate decreases. This is consistent with the actual situation. When the learning rate decreases, more times of learning are needed to converge. Therefore, we can adjust the hyper parameters value according to the actual requirements of the scenario to improve the user experience.

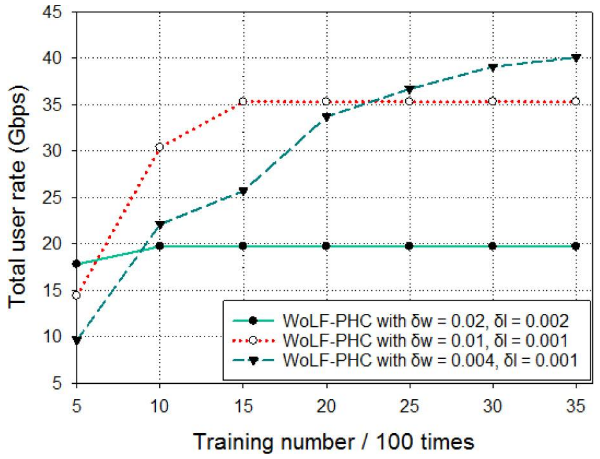

Fig. 6. Training performance on total user rate of different learning approach parameters vs. training number (200 agents).

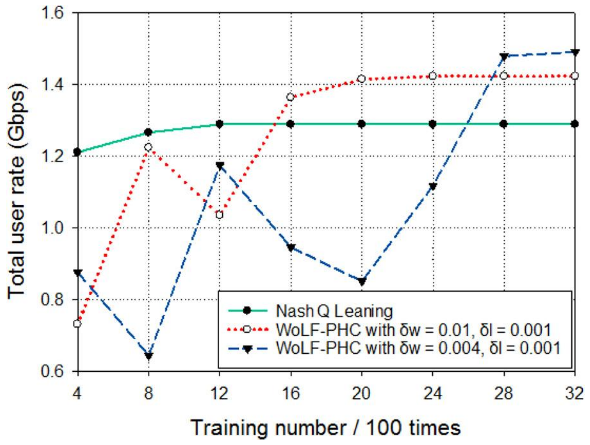

Fig. 7. Training performance on total user rate of different learning approach parameters vs. training number (5 agents).

In addition, Fig. 7 presents the training performance of Nash Q-learning with WoLF-PHC in small-scale scenario. WoLF-PHC ultimately realizes a higher total user rate value since that WoLF-PHC tends to converge to a optimal point in Nash equilibrium. Meanwhile, Nash Q-learning performs more stably on the convergence curve and reaches the convergence state faster because each user knows the whole system information in the algorithm and is more likely to choose the appropriate mode, which also means that Nash Q-learning may converge to a saddle point in nash equilibrium.

Fig. 8 shows the average user performance under different policies. In our proposed scheme, there are three strategies: macro cell broadcasting, macro cell broadcasting with mmWave small cell unicasting and our proposed smart mode selection. We consider the average user rate with different user numbers ranging from 50 to 200. First, as seen from the Fig. 8, because the user rate is related to the worst user's rate in the system and more users would bring more interference, the average user rate of all the policies declines with the number of users increases. Second, the average user rate of our proposed smart mode selection is higher than others. When the user number increases, the average user rate of our smart mode selection is more than 200% of others', which suggests that we put forward the D2D assisted 5G HetNet perform excellent.

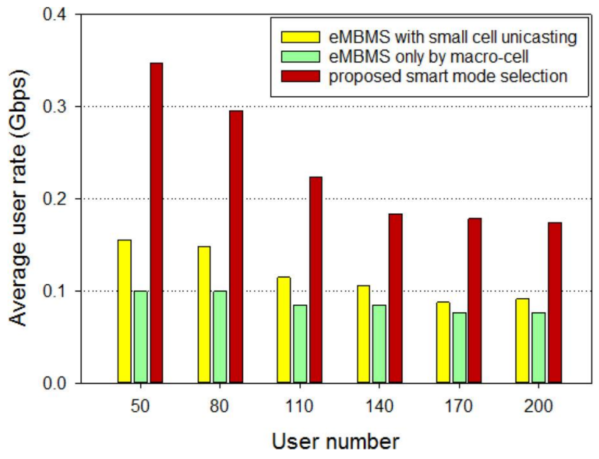

Fig. 8. Average user rate of different policies vs. user number.

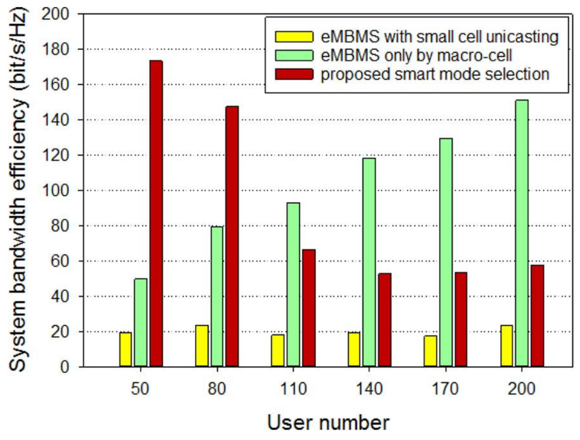

Fig. 9. System bandwidth efficiency of different policies vs. user number.

In Fig. 9, we compare the system spectrum efficiency of only macro cell broadcasting, small cell unicasting assisted macro cell broadcasting and our proposed hybrid transmission mode selection. When user number is small, our proposed smart mode selection's bandwidth efficiency is almost 300% of that of only macro cell broadcasting. As the user number increases, the bandwidth efficiency of our proposed smart mode selection is still almost 200% of that of mmWave small cells unicasting assisted eMBMS. The spectrum efficiency of macro cell broadcasting increases while that of mmWave small cells remains almost the same. Due to the reason that each user adopting a small cell policy has its own bandwidth, the spectrum efficiency of strategies, including a small cell is lower than that of other strategies when the user number increases.

The evaluation results in this section indicate that our proposed smart mode selection performs better both at the aspect of online performance and convergence speed. Moreover, the results show that the throughput and bandwidth efficiency benefit from our proposed smart hybrid transmission mode selection.

VI. Conclusion

To achieve the efficient transmission of 5G VR broadcasting, an intelligent mode selection scheme based on reinforcement learning has been proposed in this paper. First, a novel hybrid transmission mode selection framework of D2D multicasting, mmWave small unicasting, macro cell broadcasting is established to support the VR broadband broadcasting in 5G HetNets. Second, the principle of user association is discussed for each transmission mode. To maximize the system throughput, we formulate the best mode decision process as an optimization problem. Then, we propose a scheme based on online reinforcement learning to address it. State, action, and reward functions are elements of RL that are designed to adapt the proposed problem. Two RL policies, Nash Q-learning and Wolf-PHC, are discussed with a convergence analysis. Finally, the simulation results verify that our proposed smart mode selection scheme enables better system throughput for VR broadband services with a moderate resource cost than traditional broadcasting schemes in 5G HetNets. With the promotion and application of 5G technology, new technical challenges will constantly emerge in the field of 5G VR broadcasting. In the future, we will focus on a more general algorithm for multiple scenarios on the basis of this paper.

References

[1] A. Prasad, M. A. Uusitalo, D. Navrátil, and M. Säily, “Challenges for enabling virtual reality broadcast using 5G small cell network,” in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Barcelona, Spain, Apr. 2018, pp. 220–225.

[2] J. J. Gimenez *et al.*, “5G new radio for terrestrial broadcast: A forward-looking approach for NR-MBMS,” *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 356–368, Jun. 2019.

[3] J. J. Gimenez, P. Renka, S. Elliott, D. Vargas, and D. Gomez-Barquero, “Enhanced TV delivery with EMBMS: Coverage evaluation for rooftop reception,” in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, pp. 1–5, Jun. 2018.

[4] E. Bastug, M. Bennis, M. Médard, and M. Debbah, “Toward interconnected virtual reality: Opportunities, challenges, and enablers,” *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 110–117, Jun. 2017.

[5] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G. Muntean, “Single frequency-based device-to-device-enhanced video delivery for evolved multimedia broadcast and multicast services,” *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 263–278, Jun. 2015.

[6] J.Calabuig, J. F. Monserrat, and D. Gómez-Barquero, “5th generation mobile networks: A new opportunity for the convergence of mobile broadband and broadcast services,” *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 198–205, Feb. 2015.

[7] J. J. Gimenez, D. Gomez-Barquero, J. Morgade, and E. Stare, “Wideband broadcasting: A power-efficient approach to 5G broadcasting,” *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 119–125, Mar. 2018.

[8] C. Wei, G. Feng, S. Qin, and M. Yan, “Cellular offloading in heterogeneous mobile networks with D2D communication assistance,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4245–4255, May 2017.

[9] C. Bo, J. Yang, S. Wang, and J. Chen, “Adaptive video transmission control system based on reinforcement learning approach over heterogeneous networks,” *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 3, pp. 1104–1113, Jul. 2015.

[10] C. Zhang, Z. Liu, B. Gu, K. Yamori, and Y. Tanaka, “A deep reinforcement learning based approach for cost- and energy-aware multi-flow mobile data offloading,” *IEICE Trans. Commun.*, vol. E101.B, no. 7, pp. 1625–1634, 2018.

[11] X. Wang, X. Su, and B. Liu, “A novel network selection approach in 5G heterogeneous networks using *Q*-learning,” in *Proc. 26th Int. Conf. Telecommun. (ICT)*, Hanoi, Vietnam, Apr. 2019, pp. 309–313.

[12] R. Sachan, N. Saxena, and A. Roy, “An efficient hybrid scheduling scheme for impatience user in eMBMS over LTE,” in *Proc. Int. Conf. Comput. Commun. Informat.*, Coimbatore, India, Jan. 2013, pp. 1–5.

[13] A. Awada, D. S. Michalopoulos, and A. Ali, “An improved method for on-demand system information broadcast in 5G networks,” in *Proc. IEEE Conf. Stand. Commun. Netw. (CSCN)*, Helsinki, Finland, Sep. 2017, pp. 18–23.

[14] W. Guo, M. Fuentes, L. Christodoulou, and B. Mouhouche, “Roads to multimedia broadcast multicast services in 5G new radio,” in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 1–5.

[15] W. Guo and B. Mouhouche, “A method to tailor broadcasting and multicasting transmission in 5G new radio,” in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Valencia, Spain, Jun. 2019, pp. 364–368.

[16] M. Chen, W. Saad, and C. Yin, “Virtual reality over wireless networks: Quality-of-service model and learning-based resource management,” *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5621–5635, Nov. 2018.

[17] W. Yang *et al.*, “An assigned color depth packing method with centralized texture depth packing formats for 3D VR broadcasting services,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 122–132, Mar. 2019.

[18] Z. Guizani and N. Hamdi, “mmWave E-band D2D communications for 5G-underlay networks: Effect of power allocation on D2D and cellular users throughputs,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Messina, Italy, Jun. 2016, pp. 114–118.

[19] G. I. Tsiropoulos, A. Yadav, M. Zeng, and O. A. Dobre, “Cooperation in 5G HetNets: Advanced spectrum access and D2D assisted communications,” *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 110–117, Oct. 2017.

[20] H. Huang, B. Liu, L. Chen, W. Xiang, M. Hu, and Y. Tao, “D2D-assisted VR video pre-caching strategy,” *IEEE Access*, vol. 6, pp. 61886–61895, 2018.

[21] X. Zhang and J. Wang, “Heterogeneous statistical QoS-driven resource allocation for D2D cluster-caching based 5G multimedia mobile wireless networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[22] A. Omran, L. Sboui, B. Rong, H. Rutagemwa, and M. Kadoch, “Joint relay selection and load balancing using D2D communications for 5G HetNet MEC,” in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Shanghai, China, May 2019, pp. 1–5.

[23] A. Mehrabi, M. Siekkinen, G. Illahi, and A. Ylä-Jääski, “D2D-enabled collaborative edge caching and processing with adaptive mobile video streaming,” in *Proc. IEEE 20th Int. Symp. World Wireless Mobile Multimedia Netw. (WoWMoM)*, Washington, DC, USA, Jun. 2019, pp. 1–10,

[24] X. Zhang, Y. Li, and Q. Miao, “A cluster-based broadcast scheduling scheme for mmWave vehicular communication,” *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1202–1206, Jul. 2019.

[25] X. Chen, S. Leng, Z. Tang, K. Xiong, and G. Qiao, “A millimeter wave based sensor data broadcasting scheme for vehicular communications,” *IEEE Access*, vol. 7, pp. 149387–149397, 2019.

[26] W. Mei, Z. Chen, and S. Li, “Confidential broadcasting and service integration in millimeter wave systems,” *IEEE Syst. J.*, vol. 13, no. 1, pp. 147–158, Mar. 2019.

[27] Y. Niu, L. Yu, Y. Li, Z. Zhong, and B. Ai, “Device-to-device communications enabled multicast scheduling for mmWave small cells using multi-level codebooks,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2724–2738, Mar. 2019.

[28] T. Yang, Y. Hu, M. C. Gursoy, A. Schmeink, and R. Mathar, “Deep reinforcement learning based resource allocation in low latency edge computing networks,” in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Lisbon, Portugal, Aug. 2018, pp. 1–5.

[29] C. Zhang and Z. Zheng, “Task migration for mobile edge computing using deep reinforcement learning,” *Future Gener. Comput. Syst.*, vol. 96, pp. 111–118, Jul. 2019.

[30] P. K. Tathe and M. Sharma, “Dynamic actor-critic: Reinforcement learning based radio resource scheduling for LTE-advanced,” in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Pune, India, Aug. 2018, pp. 1–4.

[31] C. V. Murudkar and R. D. Gitlin, “Optimal-capacity, shortest path routing in self-organizing 5G networks using machine learning,” in *Proc. IEEE 20th Wireless Microwave Technol. Conf. (WAMICON)*, Cocoa Beach, FL, USA, Apr. 2019, pp. 1–5.

[32] C. J. Watkins and P. Dayan, “*Q*-learning,” *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[33] J. Hu and M. P. Wellman, “Nash *Q*-learning for general-sum stochastic games,” *J. Mach. Learn. Res.*, vol. 4, pp. 1039–1069, Dec. 2003.