Hindawi
Mobile Information Systems
Volume 2020, Article ID 7607316, 17 pages
https://doi.org/10.1155/2020/7607316

Research Article

Stabilizing Transmission Capacity in Millimeter Wave Links by Q-Learning-Based Scheme

Jinsong Gui, Xiangwen Dai, and Xiaoheng Deng

School of Computer Science and Engineering, Central South University, South Road of LuShan, Changsha, Hunan 410083, China

Correspondence should be addressed to Jinsong Gui; jsgui06@163.com

Received 21 October 2019; Accepted 22 January 2020; Published 11 February 2020

Academic Editor: Carlos T. Calafate

Due to uncontrollable factors (e.g., radio channel quality, wireless terminal mobility, and unpredictable obstacle emergence), a millimeter wave (mmWave) link may encounter some problems like unstable transmission capacity and low energy efficiency. In this paper, we propose a new transmission capacity stabilization scheme based on the Q-learning mechanism with the aid of edge computing facilities in an integrated mmWave/sub-6 GHz system. With aid of the proposed scheme, an integrated mmWave/sub-6 GHz user equipment (UE) can adjust its transmission power and angle, even choose a relaying UE to stabilize its transmission capacity. Differing from traditional schemes, the proposed scheme is run in edge computing facilities, where any UE only needs to provide its personalized information (e.g., base station discovery, neighboring UEs, working status (i.e., busy and idle), position coordinates, and residual energy level), and then it will receive intelligent and adaptive guidance from edge computing facilities. This facilitates each UE to maintain its transmission capacity stability by adjusting its radio parameters. The simulation results show that any UE with aid of the proposed scheme can achieve more stable transmission capacity and higher energy efficiency.

1. Introduction

Although the current cellular networks can maintain Quality of Service (QoS) provisioning and provide good user experience [1–5], current techniques in these networks hardly meet the increasing capacity demands of future wireless users [6–9]. The wireless network traffic and the number of connected user equipments (UEs) are predicted to have rapid growth in the next generation cellular network [10–12]. Richer spectrum resources, higher spectral efficiency of physical layer techniques, and denser network deployment are key enablers to cope with this growth [13, 14].

Although the sub-6 GHz frequency bands have favorable propagation characteristics, the total available bandwidth is insufficient to support the rapid growth of traffic demand. Therefore, millimeter wave (mmWave) frequency bands with an abundant amount of bandwidth can be exploited [15–18]. However, although mmWave bands have extremely rich spectrum resources, the users who are using mmWave devices may suffer from poor service due to the very high signal blocking loss.

Employing relay stations can circumvent obstacles thereby avoiding blockages [19]. However, in order to reduce the probability that mmWave links are blocked, mmWave base stations should be densely deployed [20]. In a mmWave network infrastructure, a distance of 75–100 meters between mmWave base stations is required for full coverage [21], which will obviously increase capital and operating expenditures for network operators. Therefore, ultradense deployment for achieving full mmWave coverage may be impracticable. Combining a mmWave network with an existing sub-6 GHz cellular network is a feasible way to exploit the mmWave technique [22].

The new mmWave antenna module (i.e., QTM052) has been developed by Qualcomm [23], which may coexist with a sub-6 GHz antenna module in a wireless device. Therefore, the sub-6 GHz antenna module can be used to connect cellular infrastructure when there is not any mmWave base station nearby. Furthermore, if a mmWave link is blocked, a sub-6 GHz cellular link is also an alternative. However, it is difficult to guarantee a consistent user experience, which is especially true for mobile users.

To maintain the consistency of user experience, it should be ensured that transmission capacity is as stable as possible. This will be a challenging task. Although millimeter wave links with line-of-sight (LOS) conditions have very high throughput, the throughput drops sharply when such links are blocked. Since a sub-6 GHz cellular link cannot reach the original throughput level, it results in poor user experience and thus it is not a qualified alternative. Moreover, the mobility of wireless devices makes it difficult to maintain stable transmission capacity.

Since some wireless environment factors (e.g., radio channel quality, wireless terminal mobility, and unpredictable obstacle emergence), which will affect the stability of transmission capacity, may be uncontrollable, it is necessary to adjust some controllable parameters (e.g., transmission power, antenna transmitting and receiving beam width, and alternative path) in an intelligent way to suppress the instability caused by uncontrollable factors. Machine learning methods can be used to design such an intelligent regulation scheme.

Compared to typical machine learning methods (e.g., support vector machine, linear discriminant analysis, k-nearest neighbor, and backpropagation neural network) [24], reinforcement learning is typically characterized by trial and error learning and delayed return. Multiarmed bandit (MAB) is a theoretical model of single-step reinforcement learning. Some wireless communication problems have been treated by using the MAB tool [25]. In MAB problems, each decision maker needs to select a subset of actions of unknown expected rewards, where its goal is to get the maximum reward over time [26]. Therefore, it must firstly solve the exploration and exploitation dilemma. On the one hand, all actions should be sufficiently explored for the purpose of learning their rewards. On the other hand, those actions that are known to produce high rewards should be sufficiently exploited.

Q-learning is a model-free reinforcement learning technique, which consists of a set of agents, a set of states, and a set of actions. By performing an action in a given state, an agent gets a reward, where the goal is to maximize its accumulated reward. Usually, a Q-function is used to illustrate such a reward, which is regarded as an action utility function. Q table is adapted to store the values of such a Q-function, which is used to evaluate the pros and cons of taking an action in a specific state. Firstly, each entry in Q-table is initialized to the zero value, and then it is updated in an iterative manner after an agent executes an action and gets the corresponding reward as well as the resultant next state at each time instant.

The method based on the Q value has high data utilization and stable convergence, especially when the state space is small or the number of actions is small. Therefore, we introduce the Q-learning method into this paper and combine the mentioned controllable parameters with the uncontrollable factors to construct the state space. The actual state of a wireless device will be mapped to this system-maintained state space. When the system-maintained state space is larger, the mapped state point will be closer to the actual state. However, when a target state point needs to be searched in a larger state space, it may take longer (or pay more) to reach this target state point. Therefore, the training process of the Q-learning-based scheme should be run in a resource-rich cloud infrastructure, while the trained Q table should be sent to an edge computing facility that makes decisions for UEs based on this Q table.

When any uncontrollable factor changes, the current state will be converted to a new state, which may make the transmission capacity either better or worse. If the former happens, the user experience is usually unaffected, but network resources can be saved by adjusting the controllable parameters. If the latter occurs, the controllable parameters should be adjusted to increase resource supply in order to suppress the capacity downtrend. Based on the above analysis, we propose a Q-learning-based scheme to maintain the mmWave link capacity stability, which mainly includes the following contributions:

(1) To the best of our knowledge, we are the first to combine device-to-device (D2D) communication mode with transmission power control, beam width adjustment, and the other environmental factors (e.g., link blockage and communication distance) to construct a unified state space, which lays a foundation for stabilizing transmission capacity in millimeter wave links.

(2) Any edge computing facility can offload the Q table training task to the cloud facility, while it makes decisions for each UE based on the trained Q table. Also, in order to ensure the timeliness of Q table information, we combine a cloud computing facility with multiple edge computing facilities to train each Q table. Each UE only needs to periodically report its personalized state information to the edge computing facility in its vicinity. Therefore, it is not frequently involved in the maintenance of its access link quality, which has a small maintenance overhead.

(3) Unlike the existing performance optimization methods based on Q-learning, which always optimizes system performance at all costs, while the proposed scheme only maintains the performance to meet the requirements of user application experience, which is beneficial to save resources as much as possible under the premise of meeting the user application experience.

(4) Based on cloud training, edge decision, and end-user parameter adjustment, a closed loop process is built to form a virtuous cycle of learning and doing repeatedly, which is advantageous to suppress the oscillation of transmission capacity in mmWave links caused by dynamic environmental factors.

The remainders of this paper are organized as follows. The related works are introduced in Section 2, while the system model is described in Section 3. The Q-learning solution is detailed in Section 4. Experiment setup and simulation results are presented in Section 5. The conclusions and further works are summarized in Section 6.

2. Related Work

Due to model-free characteristics, Q-learning has been widely applied in various wireless networks for an intelligent decision. The literature [27] proposed a heterogeneous distributed multiobjective strategy based on Q-learning, which was established for the self-configuration and optimization of femtocells. The literature [28] constituted a Q-learning-based scheme for dense small cell networks for the purpose of managing their cell outage.

To address interference management problems in a type of small cells (e.g., femtocell), the literature [29] proposed a set of distributed and hybrid Q-learning-based power allocation algorithms, where the goal is to improve network throughput, energy efficiency, and user experience. To effectively utilize limited resources of sensor nodes while meeting quality of service requirements, an effective task scheduling scheme should be adopted to address this problem. Based on a cooperative Q-learning model, the literature [30] designed a task scheduling algorithm, which can help sensor nodes intelligently to determine its next executing task.

The literature [31] designed a multistate reinforcement learning method to improve p-persistent carrier sense multiple access protocol. Based on Q-learning and deep learning, the literature [32] explored a transmission scheduling mechanism to improve packet transmission efficiency in cognitive radio-based Internet of Things (IoT). To improve the throughput and energy efficiency of a directional hybrid cognitive radio media access control protocol, the literature [33] proposed a channel selection algorithm based on Q-learning and a directional transmission power control scheme, respectively.

The integration of multiple Radio Access Technologies (RATs) is helpful to improve the network capacity. To maximize the long-term network throughput while meeting diverse traffic requirements, the literature [34] proposed a smart aggregated RAT access strategy and constructed a semi-Markov decision process (SMDP), where some Q-learning-based schemes are used to solve this SMDP problem.

To improve the throughput and energy efficiency of a directional hybrid cognitive radio media access control protocol, the literature [33] proposed a channel selection algorithm based on Q-learning and a directional transmission power control scheme, respectively. The former can try to select the best channel based on secondary users' observation of primary users' traffic and channel characteristics (e.g., achieved throughput and lost packets), while the latter can allow nodes to reuse the channels subject to interference constraints for the purpose of achieving maximum concurrent transmissions.

None of the above works focuses on the problem of transmission capacity stability in mmWave Links. However, the literature [35] proposed a single state Q-learning algorithm to improve the reliability of a mmWave backhaul system. When the non-line-of-sight (NLOS) operation is unavoidable, the propagation by diffraction will be employed. Based on the learning results coming from this algorithm, the system can select a suitable propagation path in a predefined time and switch to it. Although the literature [35] addresses the research on the performance reliability of mmWave links, it focuses on the improvement of the capacity of backhaul links. Moreover, in the literature [35], each agent only models its own state-action space from a partial view of the operating environment instead of a global one. Therefore, it is difficult to provide more accurate decision-making basis.

3. System Model

3.1. Network Architecture. As shown in Figure 1, we consider a single macrocell, which includes an integrated mmWave/sub-6 GHz macro base station (MBS), a certain number (denoted as m) of mmWave small base stations (SBSs), and a large number (denoted as n) of integrated mmWave/sub-6 GHz UEs. Due to the scarcity of sub-6 GHz bands, all the SBSs are only enabled on mmWave bands. As is known to all, sub-6 GHz bands have smaller bandwidth resources when compared with mmWave bands, but they have better coverage and higher stability. Therefore, sub-6 GHz bands are mainly used for signaling information and network control for the purpose of managing mmWave links.

For the sub-6 GHz frequency bands, the propagation characteristics of wireless signals can be approximated by the free space model or the two ray ground model. However, as summarized in [18], mmWave signals have the higher path loss, higher penetration loss, severe atmospheric absorption, and more attenuation due to rain and thus exhibit markedly different propagation features from those of the sub-6 GHz frequency bands. Therefore, a reasonable principle should be followed to develop mathematical frameworks, which can model realistic mmWave propagation by fully considering path loss and blockage.

The mmWave propagation model based on stochastic geometry involves three types of links (i.e., LOS, NLOS, and outage (OUT) links), where the probability of a link in any of the three states is considered as a function of distance according to [20]. If a direct path with good channel propagation conditions exists between two communication nodes, it is in the LOS state. Otherwise, when this direct path is blocked but there are other paths between them (e.g., via reflections), it is in the NLOS state. Furthermore, if the path loss between two communication nodes is so large that any link cannot be established between them, the OUT state occurs.

Data information can also be transmitted in sub-6 GHz links if necessary (e.g., for UEs with poor signal quality in mmWave bands). As mentioned above, it is difficult to reach the capacity level of mmWave links in LOS conditions. Therefore, only in the case of small traffic demand, an integrated mmWave/sub-6 GHz UE can ignore the access channel state of the nearby mmWave SBSs and select to communicate directly with the MBS on the sub-6 GHz frequency band. However, when access traffic demand is huge, it must rely on a nearby mmWave SBS to establish a high-throughput access link. Usually, the control signals and simple service request information are small, and thus they

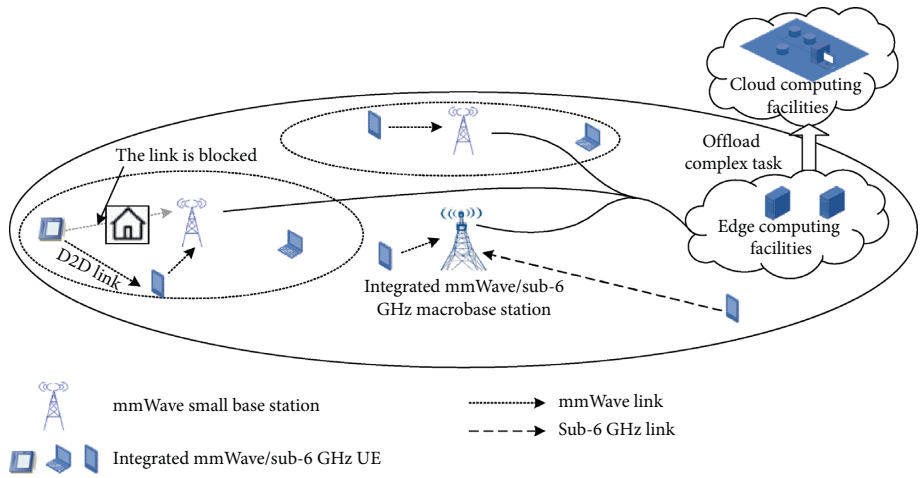

Figure 1: A macrocell architecture with the integrated mmWave/sub-6 GHz.

are handled by the long-range sub-6 GHz macrocell. Short videos or photos that are shot and transmitted instantly have a lot of data, and thus, they are handled by the mmWave small cells for high capacity.

Any UE can find the desired SBS by actively initiating a probe packet or passively receiving beacon information from its nearby SBSs. When this UE needs to transmit large amounts of short video or photo data and only finds the SBS to which it can connect in NLOS conditions, that is, there only exists a NLOS direct communication link between the UE and the SBS, we consider constructing the path composed of multisegment LOS links by using the D2D communication mode, which is used to replace this NLOS mmWave link. If there is a good incentive mechanism, we believe that any UE is usually willing to act as a D2D relaying UE if it is in idle state and also requested. Such an incentive mechanism is beyond the scope of this paper, but similar schemes have been explored in some literature [13, 36].

The destination UEs and selected relaying UEs can fully exploit the wideband mmWave bands if they have LOS conditions. In addition, we assume that the edge computing facilities are deployed within the macrocell, which may take on computing tasks offloaded by UEs and provide real-time responses. Also, since the edge computing facilities only have limited computing resources, they offload the complex training task of the Q-learning scheme to the cloud infrastructure.

3.2. Problem Statement. In this paper, our goal is to stabilize transmission capacity in mmWave links by designing a Q-learning-based scheme. To this end, we first need to build a state-space table for each UE, which should take full account of the factors that affect transmission capacity of mmWave links.

For any UE i and $1 \le i \le n$, we denote the *set* of the associating states as $B_i = \{b_{i,j} \mid 1 \le j \le m\}$, where $b_{i,j} \in$ {"Unassociated," "Associated"}. "Unassociated" means that UE i cannot discover SBS j, while "Associated" means that UE i can discover SBS j with which it may be associated.

The *set* of the distances to SBSs from UE i is denoted as $D_i = \{d_{i,j} \mid 1 \le j \le m\}$, where $d_{i,j} \in$ {"Near," "Medium," "Far"}. "Near" means that the distance to SBS j from UE i is less than half of the maximum coverage distance of SBS j. If the distance to SBS j from UE i is more than the maximum coverage distance of SBS j, it falls under the category "Far." Except for the above two cases, it belongs to the category "Medium."

The *set* of the transmission powers to SBSs from UE i is denoted as $P_i = \{p_{i,j} \mid 1 \le j \le m\}$, where $p_{i,j} \in$ {"Invalid," "Low," "Medium," "High"}. "Invalid" means that UE i cannot communicate with SBS j even if it adopts its maximum transmission power. In the cases of "Low," "Medium," "High," UE i can communicate with SBS j. If the transmission power to SBS j from UE i is not more than one-third of the maximum transmission power of UE i, it belongs to "Low." If the transmission power to SBS j from UE i is more than one-third but less than two-thirds of the maximum transmission power of UE i, it belongs to "Medium." If the transmission power to SBS j from UE i is not less than two-thirds of the maximum transmission power of UE i, it belongs to "High."

The *set* of the transmission angles (or transmitting beam widths) to SBSs from UE i is denoted as $G_i = \{g_{i,j} \mid 1 \le j \le m\}$, where $g_{i,j} \in$ {"Very small," "Small," "Medium," "Big"}. "Very small" means that the transmission angle to an SBS from UE i is less than 30°. "Small" means that the transmission angle to an SBS from UE i ranges from 30° to 90°. "Medium" means that the transmission angle to an SBS from UE i ranges from 90° to 180°. "Big" means that the transmission angle to an SBS from UE i is more than 180°.

We denote the *set* of the relaying states as $L_i = \{l_{i,j} \mid 1 \le j \le m\}$, where $l_{i,j} \in$ {"Unselected," "Selected"}. "Unselected" means that UE i does not select a relaying UE to forward its data to SBS j, while "Selected" means that it has done so.

When a UE cannot be associated with any SBS in the current, this state of disconnect from any SBS may change if it is moving. If it stays stationary for a period of time, it had better maintain communication with an SBS through the

D2D communication mode. After broadcasting a D2D relaying request packet, a UE can select one of its neighbors that send D2D relaying response packets to act as its D2D relay. Such neighbors need to meet the following conditions: (1) they currently have no need to transmit data through an SBS but can be connected to it; (2) they can stay stationary for a period of time; and (3) they have received a D2D relaying request packet. By a D2D relaying response packet, a responder will report the link capacity between it and its associated SBS and its own position coordinates, so as to facilitate the corresponding requester to make a reasonable decision.

The theoretical free space path loss (FSPL) in the D2D relaying link from the requester (e.g., UE i) to the responder (e.g., UE i') is given by Friis' Law [37] as follows, which is used in the LOS scenario:

$$\mathrm{FSPL} = 20 \cdot \log_{10}(d) + 20 \cdot \log_{10}(f_c) + 32.45. \quad (1)$$

In (1), FSPL is measured in dB; f_c is the carrier frequency in GHz. When the requester adopts its maximum transmission power to broadcast its D2D relaying request, the RSS value under the omnidirectional path loss is estimated by the following formula:

$$P_{\mathrm{RSS}} = P_{\mathrm{TX\_MAX}} + G_{\mathrm{TX}} + G_{\mathrm{RX}} - PL\,(\text{or FSPL}) - 30. \quad (2)$$

In (2), P_{RSS} and $P_{\mathrm{TX\_MAX}}$ are the RSS value of the responder and the maximum transmission power of the requester in dBm, respectively, while G_{RX} and G_{TX} are the receiving gain of horn antenna and transmitting gain of horn antenna in dBi, respectively, where G_{RX} and G_{TX} may take the value 24.5 dBi. If the UE i knows the acceptable level of receiving bit error rate (BER) of the UE i', the corresponding receiving power threshold for the UE $i'$is estimated by the following formula:

$$p_{i'}^{\mathrm{th}} = 10 \log_{10}\left(-2\sigma^2 \ln \mathrm{BE}_{\mathrm{th}}\right). \quad (3)$$

In (3), $p_{i'}^{\mathrm{th}}$ is the receiving power threshold for the UE i', which is measured in dB; $\mathrm{BE}_{\mathrm{th}}$ is the receiving BER threshold in the D2D link from the UE i to the UE i', where the desired BER value is usually 10^{-8}~10^{-10}; and σ^2 is the ambient noise power in the receiving-end which is measured in Watt. The corresponding strategy to determine the actual adopted transmission power $p_{i,i'}$ is as follows:

$$\begin{cases} p_{i,i'}^{t} = \mathrm{FSPL} + p_{i'}^{\mathrm{th}}, \\ p_{i,i'} = \min\left\{p_i^{\max}, p_{i,i'}^{t}\right\}. \end{cases} \quad (4)$$

In (4), $p_{i,i'}^{t}$ is the desired transmission power in the D2D link from the UE i to the UE i', which is measured in dB; $p_i^{\max}$ is the maximum transmission power of the UE i, which is measured in dB; and FSPL is the path loss value in the D2D link from the UE i to the UE i'. When the UE i knows its own position coordinates and the position coordinates for the UE i', the distance from the UE i to the UE i' can easily be determined by the formula for the distance between two points. Based on the distance from the UE i to the UE i', we can estimate FSPL according to formula (1).

Based on the above, when UE i only expects to access SBS j, its state space is defined as follows:

$$s_{i,j} = b_{i,j} \times d_{i,j} \times p_{i,j} \times g_{i,j} \times l_{i,j}. \quad (5)$$

In (5), there are 192 composite states for UE i, where each composite state contains five single state dimensions. In theory, the number of composite states will increase exponentially with the refinement of the granularity of single state values.

For any UE i, the adoptable actions that promote the state transitions are generated by changing the values of $p_{i,j}$, $g_{i,j}$, and $l_{i,j}$ since the values of $b_{i,j}$ and $d_{i,j}$ are hardly controlled by UE i. Therefore, the set of actions that UE i can take is defined below:

$$a_{i,j} = p_{i,j} \times g_{i,j} \times l_{i,j}. \quad (6)$$

In (6), there are 32 actions for UE i. Moreover, when UE i wants to access anyone of all the SBSs, its state space is defined as follows:

$$S_i = \bigcup_{j=1}^{m} s_{i,j}. \quad (7)$$

Accordingly, the corresponding action space of UE i is defined below:

$$A_i = \bigcup_{j=1}^{m} a_{i,j}. \quad (8)$$

In order to focus on network applications instead of spending resources on network performance maintenance, an UE delegates the edge computing facility in its vicinity to maintain its state space and make decisions for it. In order to ensure that Q tables are updated in a timely manner, we use the combination of the cloud computing facility and the multiple edge computing facilities to train them. These Q tables are established and then trained in the cloud computing facility, which will be sent to the edge computing facilities after a certain level of training. Based on each trained Q table, the edge computing facilities can make decisions for each UE to guide it in making parameter adjustments to maintain transmission capacity stability.

Also, each computing facility updates its own Q table after each decision. When a computing facility receives the trained Q table from the cloud computing facility, it updates its Q table by computing the Q table coming from the cloud computing facility and its own current Q table. Since each value of a Q table will eventually converge to its corresponding fixed value from any initial value, the cloud computing facility or each edge computing facility can continue to train each Q table based on the current information. Therefore, although the more frequent change of the environment may result in the longer training time, it only has a slight impact on the performance of the trained Q-tables as long as the training process is ongoing. On the contrary, when a Q table is trained for a certain period of time, the cloud facility can stop training it to save resources.

Each UE needs to periodically report its personalized information to the edge computing facility in its vicinity. For example, based on the information from UE i, the edge

computing facility will map the actual state of UE i to a state on the trained Q table. Starting with this mapped state, the trained Q table will indicate a feasible path for state transitions, where the corresponding Q value is updated after the corresponding state transition. For each composite state, the corresponding transmission capacity level is estimated through a set of predetermined policy rules. In a feasible path for state transitions, each state transition should improve this transmission capacity. The state transition operations will be stopped if this transmission capacity meets the application requirements or the feasible path has been traversed. The details of the solution are detailed below.

4. The Q-Learning Scheme

4.1. Reward Table. The reward values in different states can be stored in a set of reward tables, where each reward table is a two-dimensional matrix with the states as the rows and the actions as the columns, as shown in Figure 2. In this paper, the reward value of UE i ($1 \le i \le n$) is defined as the transmission capacity that it can acquire to access SBS j ($1 \le j \le m$) in an energy efficient way after taking an action a_y ($1 \le y \le 32$) under a state s_x ($1 \le x \le 192$), which is denoted as $s_{i,j} : r(s_x, a_y)$ and estimated by the following formula:

$$s_{i,j} : \; r\left(s_x, a_y\right) = \begin{cases} 0, & s_x = (\text{“0”}, *, *, *, \text{“0”}), \\ \dfrac{bw}{p_{i,j}^t} \cdot \log_2\left(1 + \dfrac{\varphi_{i,j}^t \cdot p_{i,j}^t \cdot \partial_{i,j}}{\sigma^2}\right), & s_x = (\text{“1”}, *, *, *, \text{“0”}). \end{cases} \tag{9}$$

In (9), s_x is a five-tuple composite state, and its first term is assigned a value from $b_{i,j}$, where “Unassociated” and “Associated” are replaced as “0” and “1,” respectively, for simplifying the representation. Also, the fifth term of s_x is assigned a value from $l_{i,j}$, where “Unselected” and “Selected” are replaced as “0” and “1,” respectively, for the same reason. The other three terms are assigned a value from $d_{i,j}$, $p_{i,j}$, and $g_{i,j}$, respectively, where the symbol $*$ means that any value in the corresponding *set* can be taken. Also, a_y in (5) is a three-tuple, where the three terms are assigned a value from $p_{i,j}$, $g_{i,j}$, and $l_{i,j}$, respectively.

In addition, bw is the bandwidth of UE i sending data to SBS j. σ^2 is the ambient noise power. $\partial_{i,j}$ is the channel attenuation coefficient of UE i sending data to SBS j, which can be detected and measured by the receiving end. Also, $\partial_{i,j}$ can be estimated by some empirical formulas. In this respect, the many mmWave channel propagation models are summarized in the literature [38], which will be the basis for building such empirical formulas. $p_{i,j}^t$ is the power of UE i sending data to SBS j, which is the first term of a_y. $\varphi_{i,j}^t$ is a receiving power enhancement coefficient, which is approximately expressed as follows:

$$\varphi_{i,j}^t = \ln\left(1 + \frac{360}{g_{i,j}^t}\right). \tag{10}$$

In (10), $g_{i,j}^t$ is the angle of UE i sending data to SBS j, which is just the second term of a_y. The larger transmitting angle will lead to the weaker receiving power, since the transmission power is more dispersed. Formula (10) roughly characterizes this feature.

4.2. Time Slot Structure. For each UE, the communication time is divided into even intervals with a constant length T_L, and each is denoted by $t \in T = \{1, 2, \ldots\}$, which is shown in Figure 3. Each UE executes the following tasks in turn during each interval. Firstly, in the information report slot, each UE reports its personalized information to the edge computing facility, for example, SBS discovery, neighboring UEs, working status (i.e., busy and idle), position coordinates, and residual energy level.

Then, each UE will wait for feedback from the edge computing facility in the waiting for the feedback slot, during which the edge computing facility entrusts the cloud facility to perform Q table training tasks and then makes decisions for each UE based on the corresponding trained Q table. To speed up the response, the edge computing facility may make decisions by using the previous trained Q table, which was trained by the cloud computing facility based on the collected information in the previous interval, while the trained Q table based on the collected information in the current interval will be used in the next interval.

Next, based on the decision results coming from the edge computing facility, each UE will adjust its transmission parameters in the parameter adjustment slot, where the goal is to achieve the stability of access capacity as far as possible. Finally, each UE will use the adjusted parameters for data transmission in the data communication slot. If the first three time slots can be shortened as much as possible, the fourth time slot will be long enough, which is conducive to network access capacity enhancement.

4.3. Q Table. Like a reward table, a Q table is also a two-dimensional matrix with the states as the rows and the actions as the columns. Each value in a Q table means the knowledge obtained by an agent from the network environments. Each agent keeps learned values from the network environments in its Q table for all its possible actions.

At the beginning of an interval t, an agent (e.g., UE i) under a state (e.g., s_x) chooses an action (e.g., a_y) and receives a reward (e.g., $s_{i,j} : r(s_x, a_y)$) at the beginning of the next interval $t + 1$. Initially, because there is no experience to learn, each entry of a Q table is initialized to 0. The Q value is then updated based on the following Q-function:

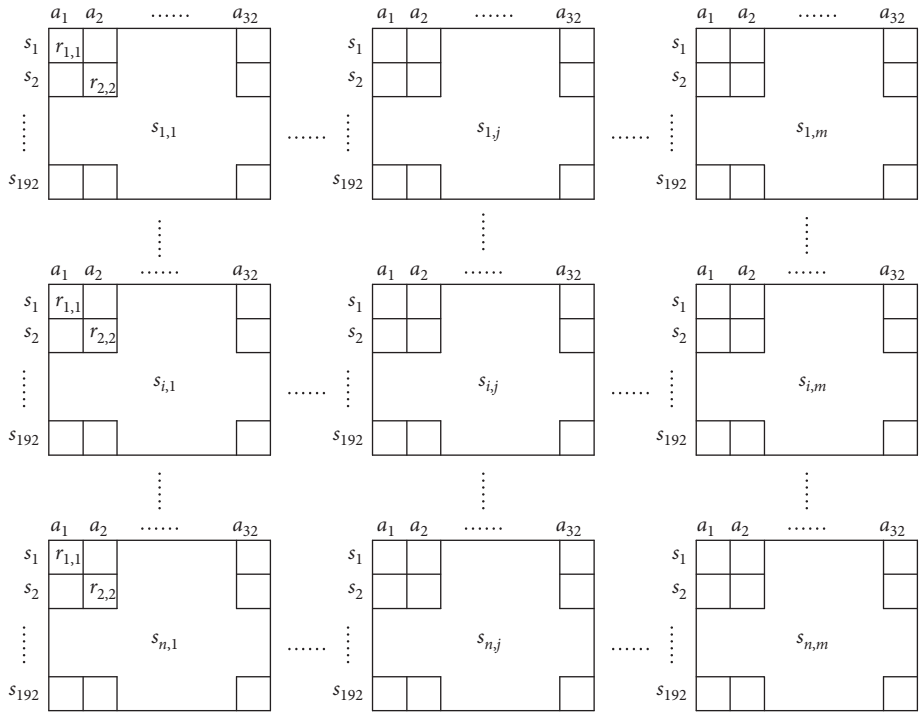

Figure 2: An example for a set of reward table.

$$s_{i,j} : Q_{t+1}\left(s_x, a_y\right) = (1-\alpha)\cdot s_{i,j} : Q_t\left(s_x, a_y\right) + \alpha \cdot \left(\begin{array}{c} s_{i,j} : r_{t+1}\left(s_x, a_y\right) + \\ \beta \cdot \max\limits_{s_{\widehat{x}} \in S_i, a_{\widehat{y}} \in A_i} \left(s_{i,j} : Q_{t+1}\left(s_{\widehat{x}}, a_{\widehat{y}}\right)\right) \end{array} \right). \tag{11}$$

In (11), α denotes a learning rate and $0 \le \alpha \le 1$, where a higher value of α gives more weight to current reward than past knowledge. β is a discount factor and $0 \le \beta < 1$, where a higher value of β means that an agent attaches greater importance to future rewards. $s_{\widehat{x}}$ is the next state of s_x after a_y is executed. $a_{\widehat{y}}$ is the action that maximizes Q value under the state $s_{\widehat{x}}$.

In a Q-learning algorithm, a learning rate and a discount factor are two important parameters, where the former is used to measure the speed of the learning process, while the latter is used to measure the proportion of future rewards.

4.4. State Transition Strategy. The strategy for an agent (e.g., UE i) to choose an action from A_i under a given state $s_x \in S_i$ is modeled as follows:

$$V_{i,j} = \max_{a_y \in A_i}\left(s_{i,j} : Q_t\left(s_x, a_y\right)\right). \tag{12}$$

In (12), $V_{i,j}$ denotes a strategy function, where the action that maximizes the value of $V_{i,j}$ under a given state $s_x \in S_i$ will be selected.

For each UE, the edge computing facility will keep a Q table for it, which records each Q value for each environment state and each possible action. Exploitation was performed by the edge computing facility using an epsilon-greed algorithm, which randomly selects one of the other actions except for the best-known action to enhance the estimates of all the Q values at the exploration probability ϵ and selects the best-known action at the exploration probability $1-\epsilon$.

4.5. Description of Q-Learning Algorithm. The pseudocode description for reward table initialization is shown as Algorithm 1, where the specific meaning of line 7 is to adopt the allowed maximum transmission power, the allowed maximum transmission angle, and the allowed the furthest distance to SBS in a specific state to compute reward value under this state.

Since the initialization of reward tables does not require any UE's personalized information, it can be handled independently by the cloud facility. However, due to the dynamic wireless communication environment, this kind of personalized information is required to periodically update reward tables. In order to meet the timeliness requirement of the interaction, this update process should be handled by the edge computing facility.

In the information report slot of each interval with a constant length T_L, each UE (e.g., i) will report its personalized information to the edge computing facility. Usually, it reports relevant information through its associated SBS (e.g., j) on a mmWave channel, which facilitates SBS j to estimate the parameter $\partial_{i,j}$ based on the channel state information it perceives. If UE i cannot be associated with any SBS, it reports relevant information through MBS on a sub-6 GHz cellular channel. In the subsequent waiting

Run at the cloud computing facility
Input: S_i and A_i for UE i
Output: the initialized reward table for UE i

(1) **For** each SBS $j \in \{1, 2, \ldots, m\}$ **do**
(2) **For** each $s_x \in \{s_1, s_2, \ldots, s_{192}\}$ **do**
(3) **For** each $a_y \in \{a_1, a_2, \ldots, a_{32}\}$ **do**
(4) **If** UE i is not associated with SBS j **then**
(5) $s_{i,j} : r(s_x, a_y) = 0$
(6) **Else**
(7) Determine $g_{i,j}^t$, $p_{i,j}^t$, and $\partial_{i,j}$ according to s_x and a_y
(8) $\varphi_{i,j}^t = \ln(1 + (360/g_{i,j}^t))$
(9) $s_{i,j} : r(s_x, a_y) = (bw/p_{i,j}^t) \cdot \log_2(1 + (\varphi_{i,j}^t \cdot p_{i,j}^t \cdot \partial_{i,j}/\sigma^2))$
(10) **End if**
(11) **End for**
(12) **End for**
(13) **End for**

Algorithm 1: The reward table initialization process for UE i.

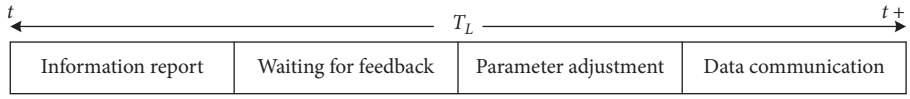

Figure 3: Time slot structure for UEs.

for the feedback slot, the edge computing facility updates the reward table, entrusts the cloud computing facility to train the Q table, and makes decisions based on the trained Q-table for each UE.

Once the personalized information is received from UE i, the edge computing facility firstly determines whether UE i is associated with an SBS according to the SBS discovery information reported by UE i. If it is true, the edge computing facility will abandon updating the reward table for UE i. Otherwise, for each neighboring UE of UE i, if the edge computing facility finds that it can be associated with an SBS according to the SBS discovery information reported by it and also it is idle according to the working status information reported by it, it is regarded as a candidate relaying UE of UE i. If UE i has multiple candidates relaying UEs, the candidates who are in the same SBS coverage area as UE i will be given priority, where it is easy to determine whether they are in the same SBS coverage area based on the information of position coordinates reported by them. After the above screening, if UE i still has multiple candidates relaying UEs, the candidate with the highest energy reserve level becomes the resulting relaying UE of the UE i. The pseudocode description for reward table update is shown as Algorithm 2.

The updated reward table for UE i is submitted to the cloud computing facility for the purpose of training the Q table. The pseudocode description for the Q table training process is shown in Algorithm 3.

Although Algorithm 3 can train Q table successfully, it needs more capital and operating expenditures. Also, it spends longer time to obtain the trained Q table. Since a real state is rarely mapped to any initial state that contains "$d_{i,j}$ = "Far,"" we can define a screening condition to filter all the initial states with "$d_{i,j}$ = "Far."" That is, if any initial state does not contain "$d_{i,j}$ = "Far,"" the Q table training process is allowed (i.e., the lines 4~24 in Algorithm 3 will be executed).

Moreover, we can ignore the actions that are rarely adopted in the search process of action space. For example, when the current state contains "$b_{i,j}$ = "Associated,"" we can ignore the actions that contain "$l_{i,j}$ = "Selected."" Also, when the current state contains "$b_{i,j}$ = "Unassociated,"" we can ignore the actions that contain "$l_{i,j}$ = "Unselected."" The above constraints in terms of selecting actions are used to avoid the execution of the lines 6~11 in Algorithm 3, so it can speed up the Q table training process.

The trained Q table for UE i will be sent to the edge computing facility, which will be used to make decisions for UE i. After each decision, the corresponding Q value will be updated. Prior to receipt of a new trained Q table, the existing Q table will be both used and updated repeatedly by the edge computing facility. The pseudocode description for the Q-table-based decision and update process is shown in Algorithm 4.

In Algorithm 4, we set a threshold for energy efficiency in mmWave links, which is denoted as r_{th}. On the one hand, when the threshold value is large, it can maintain high energy efficiency of the system, but the average decision time based on the Q table will be longer. On the other hand, when the threshold is small, the reverse is true. Therefore, a reasonable setting of this threshold will achieve the appropriate tradeoff between the energy efficiency of the system and the decision-making speed of the proposed scheme.

In order to make readers more intuitive understanding of the collaborative relationship between the four algorithms, an example diagram of algorithms deployment and interaction is given in Figure 4. For any UE, it collects its personalized information and then reports it to the edge

Run at the edge computing facility
Input: the initialized reward table for UE i and the personalized information reported by all the UEs
Output: the updated reward table for UE i

(1) Find the SBS associated by UE i according to the personalized information reported by UE i
(2) **If** there is not any SBS associated by UE i **then**
(3) Determine the set of neighboring UEs according to the personalized information reported by UE i
(4) **For** each neighboring UE i' **do**
(5) Determine its associating state and working state according to the personalized information reported by UE i'
(6) **If** UE i' is both associated with an SBS and idle **then**
(7) Record it as a candidate relaying UE of UE i and store it in the *set* R_i
(8) **End if**
(9) **End for**
(10) Extract each candidate from the *set* R_i that is in the same coverage area as the UE i and then store it in the *set* SR_i
(11) **If** the *set* SR_i is not *empty* **then**
(12) Select the candidate with the highest energy reserve level from the *set* SR_i, which is denoted as UE i' and associated with SBS j
(13) **For** each $s_x \in \{s_1, s_2, \ldots, s_{192}\}$ **do**
(14) **For** each $a_y \in \{a_1, a_2, \ldots, a_{32}\}$ **do**
(15) Determine $g^t_{i',j}$, $p^t_{i',j}$, and $\partial_{i',j}$ according to s_x and a_y
(16) $\varphi^t_{i',j} = \ln(1 + (360/g^t_{i',j}))$
(17) $s_{i,j} : r(s_x, a_y) = (bw/p^t_{i',j}) \cdot \log_2(1 + ((\varphi^t_{i',j} \cdot p^t_{i',j} \cdot \partial_{i',j})/\sigma^2))$
(18) **End for**
(19) **End for**
(20) **Else if** the *set* R_i is not *empty* **then**
(21) Select the candidate with the highest energy reserve level from the *set* R_i, which is denoted as UE i' and associated with SBS j'
(22) **For** each $s_x \in \{s_1, s_2, \ldots, s_{192}\}$ **do**
(23) **For** each $a_y \in \{a_1, a_2, \ldots, a_{32}\}$ **do**
(24) Determine $g^t_{i',j'}$, $p^t_{i',j'}$ and $\partial_{i',j'}$ according to s_x and a_y
(25) $\varphi^t_{i',j'} = \ln(1 + 360/g^t_{i',j'})$
(26) $s_{i,j} : r(s_x, a_y) = (bw/p^t_{i',j'}) \cdot \log_2(1 + ((\varphi^t_{i',j'} \cdot p^t_{i',j'} \cdot \partial_{i',j'})/\sigma^2))$
(27) **End for**
(28) **End for**
(29) **End if**
(30) **End if**

Algorithm 2: The reward table update process for UE i.

computing facility. Also, when it receives the feedback of the decision result, it will adjust its transmission parameters for purpose of stabilizing transmission capacity.

For the edge computing facility, on the one hand, when it receives the personalized information reported by each UE and the initialized R table coming from the cloud computing facility, it will invoke Algorithm 2 to update the R table and then send the updated R table to the cloud computing facility; on the other hand, when it receives the trained Q table coming from the cloud computing facility, it will invoke Algorithm 4 to make decision for each UE and then feedback the decision result.

For the cloud computing facility, on the one hand, it invokes Algorithm 1 to initialize the R table for each UE and then sends the initialized R table to the edge computing facility; on the other hand, when it receives the R table updated by the edge computing facility, it invokes Algorithm 3 to train Q table and then feedbacks the trained Q table to the edge computing facility.

5. Performance Evaluation

5.1. Simulation Metrics and Deployment Settings. In our simulations, we evaluate the performance of the proposed Q-learning-based scheme for stabilizing transmission capacity. We will observe the stability of transmission capacity from the three performance indexes (i.e., the number of UEs connected with SBS, the average number of state transitions, and the average energy efficiency of working UEs). The number of UEs connected with SBS is defined as the number of working UEs that can communicate with SBSs in a direct or indirect manner. The energy efficiency is defined as the ratio of data rate to power consumption, while the average energy efficiency is an average value of all the working UEs' energy efficiency. The simulation scenario is shown in Figure 1, where the macrocell coverage radius is 1000 meters and the mmWave small cell coverage radius is 100 meters. The number of mmWave SBSs is set to a fixed value (i.e., $m = 70$), and these SBSs are placed inside the macrocell in a nonoverlap way. A large number of UEs are randomly distributed in the macrocell.

At the beginning of each interval, each UE determines to move with probability p_{mov} (e.g., a value from 0.1 to 0.5), while it decides to stay still with probability $1 - p_{\text{mov}}$ until the beginning of the next interval. When determining to move, it adopts the random walk model, where a UE selects a direction randomly from 0 to 2π, a speed randomly from 0.1 to

Run at the cloud computing facility
Input: α, β, the updated reward table for UE i
Output: the trained Q table for UE i
(1) Initialize each entry of Q table to 0
(2) **For** each episode **do**
(3) Randomly select an initial state $s_x \in \{s_1, s_2, \ldots, s_{192}\}$
(4) $Q_{max} = 0$
(5) **For** each $a_y \in \{a_1, a_2, \ldots, a_{32}\}$ **do**
(6) Compute $s_{i,j} : Q_{t+1}(s_x, a_y)$ according to formula (7)
(7) Update the corresponding entry of Q table
(8) **If** $s_{i,j} : Q_{t+1}(s_x, a_y) > Q_{max}$ **then**
(9) $Q_{max} = s_{i,j} : Q_{t+1}(s_x, a_y)$
(10) $a_{max} = a_y$
(11) **End if**
(12) **End for**
(13) Determine the exploration probability ϵ (e.g., 0.1) based on exploration-exploitation policy
(14) Generate a random number ε from 0 to 1
(15) **If** $\varepsilon > \epsilon$ **then**
(16) **If** a_{max} can transfer s_x to the next state (e.g., $\hat{s}_x$) **then**
(17) $s_x = \hat{s}_x$ and go to 4
(18) **End if**
(19) **Else**
(20) Randomly select an action from $\{a_1, a_2, \ldots, a_{32}\}/a_{max}$
(21) **If** the selected action can transfer s_x to the next state $\hat{s}_x$ **then**
(22) $s_x = \hat{s}_x$ and go to 4
(23) **End if**
(24) **End if**
(25) **End for**

Algorithm 3: The Q table training process for UE i.

Run at the edge computing facility
Input: r_{th}, the trained Q table for UE i, and the current state $s_x \in (s_{i,j} \subset S_i)$
Output: the target state s_d
(1) $s_d = s_x$ and $V_{i,j} = 0$
(2) Compute the real energy efficiency value for UE i according to the personalized information reported by UE i and formula (5), and then save it in r_i
(3) **If** $r_i < r_{th}$ **then**
(4) **For** each $a_y \in \{a_1, a_2, \ldots, a_{32}\}$ **do**
(5) Get $s_{i,j} : Q_t(s_d, a_y)$ by Q table according to s_d and a_y
(6) **If** $V_{i,j} < s_{i,j} : Q_t(s_d, a_y)$ **then**
(7) $V_{i,j} = s_{i,j} : Q_t(s_d, a_y)$
(8) $a_t = a_y$
(9) **End if**
(10) **End for**
(11) Compute $s_{i,j} : r_{t+1}(s_d, a_t)$ according to formula (5)
(12) **If** $s_{i,j} : r_{t+1}(s_d, a_t) < r_{th}$ and the action a_t can transfer s_d to the next state $\hat{s}_d$ **then**
(13) Compute $s_{i,j} : Q_{t+1}(s_d, a_t)$ according to formula (7)
(14) Update the corresponding entry of Q table
(15) $s_d = \hat{s}_d$ and go to 4
(16) **End if**
(17) **End if**

Algorithm 4: The Q-table-based decision and update process.

1 m/s, and then moves until the beginning of the next interval according to the selected direction and speed. When the mobile UE reaches the simulation boundary, it will bounce back from the simulation boundary and then continue to move, where the rebound angle is determined by the incident direction. We adopt the mmWave channel model of the 73 GHz band and consider the three types of mmWave link states (i.e., OUT, LOS, and NLOS) described

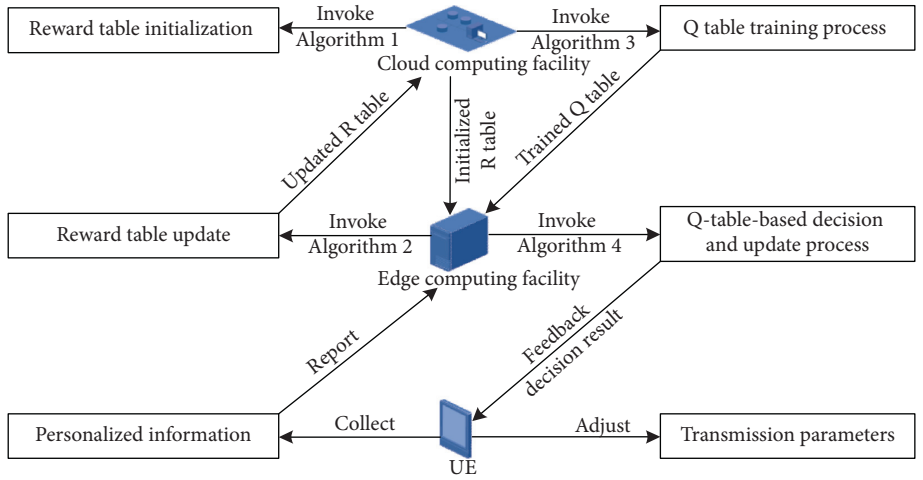

Figure 4: Example diagram of algorithms deployment and interaction.

in [20]. The probability of a UE in the outage state is formulated as the following function in terms of this UE's transmitting angle to the SBS and the distance between this UE and the SBS:

$$p_{ls} = \frac{1}{\left(1 + (d)^{-1}\right)} \cdot \cos\frac{g_{i,j}^{t}}{4}, \quad d > 60. \tag{13}$$

In (13), d is the distance in meters between the UE and the SBS. When $d > 60$, the outage state occurs with the probability p_{ls}, while the NLOS state happens with the probability $1 - p_{ls}$. Also, we assume that the emergence probability of NLOS and LOS states is equal, but no outage state occurs when $30 \le d \le 60$, while the LOS state always occurs when $d < 30$. The bandwidth of a mmWave channel is set to 1 GHz, and the maximum transmission power of each UE is set to 20 dBm. Furthermore, in our simulation scenarios, the path loss values of LOS links are estimated by formula (1), while those of NLOS links are estimated by the following formula [20]:

$$\begin{aligned} PL &= \eta + \chi \cdot 10 \cdot \log_{10}(d) + \xi, \\ \xi &\sim \mathcal{N}\left(0, \omega^2\right). \end{aligned} \tag{14}$$

In (14), PL is measured in dB; ω^2 is the lognormal shadowing variance, which is measured in dB; η and χ are the best fit floating intercept and slope over the measured distances (from 30 to 200 meters), respectively; and ω^2, η, and χ take 8, 86.6, and 2.45, respectively, in the 73 GHz mmWave band.

5.2. Simulation Results and Analysis. We compare the proposed scheme with a set of schemes in which the fixed transmission power and angle are adopted. For convenience, the compared schemes are subdivided into the compared scheme one (i.e., the transmission power and transmission angle are fixed as one-third of the maximum value), the compared scheme two (i.e., the transmission power and transmission angle are fixed as two-thirds of the maximum value), and the compared scheme three (i.e., the transmission power and transmission angle are fixed as the maximum value).

For the above four schemes, the proportion of working UEs (i.e., the UEs that have data be transmitted through SBSs) in the total number of UEs is fixed as 50%. When the channel noise power and the threshold r_{th} for energy efficiency are set as −100 dBm and 700 Mbps/W, respectively, the performance variation trend with the number of UEs is shown in Figure 5. From Figure 5(a), we see that the number of UEs connected with SBS increases with the number of UEs. The reason behind is that the number of working UEs increases with the number of UEs since the ratio of working UEs is fixed. Also, from Figure 5(b), we observe that the average energy efficiency of working UEs hardly varies with the number of UEs, which shows that the increase in network size has no obvious effect on energy efficiency. This is because that the mmWave frequency band is rich in resources and can support high concurrent communication.

The simulation results in Figure 5 also demonstrate that the proposed scheme is obviously superior to the others in terms of both the number of UEs connected with SBS and the average energy efficiency of working UEs. On the one hand, for each working UE that cannot communicate directly with SBS, the proposed scheme may establish a communication path by selecting a relaying UE, which is beneficial to increase the number of UEs connected with SBS. On the other hand, there are fewer working UEs that cannot communicate with SBS in the proposed scheme and thus less energy is wasted when they all try to connect to SBSs.

When the number of UEs and the threshold r_{th} for energy efficiency are set as 1000 and 700 Mbps/W, respectively, the performance variation trend with the channel noise power is shown in Figure 6, which shows that the number of UEs connected with SBS hardly varies with the channel noise power, while the average energy efficiency decreases with the channel noise power. This is because the problem of broken links caused by high noise power may be restored by increasing the transmission power within a

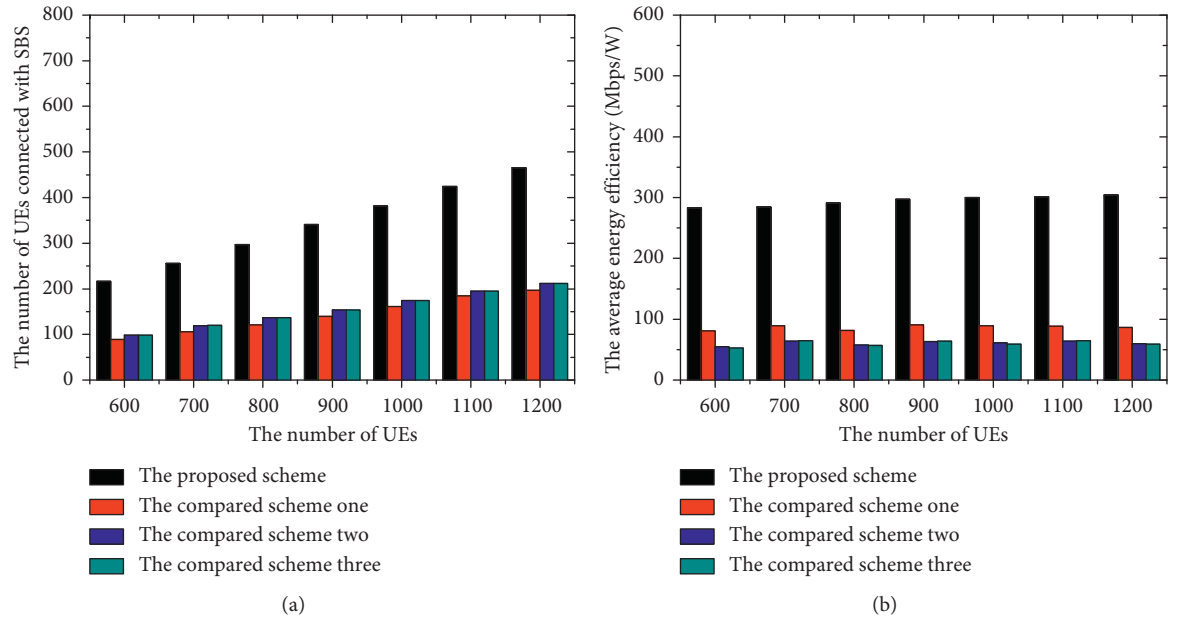

Figure 5: The performance variation trend versus the number of UEs for the above four schemes.

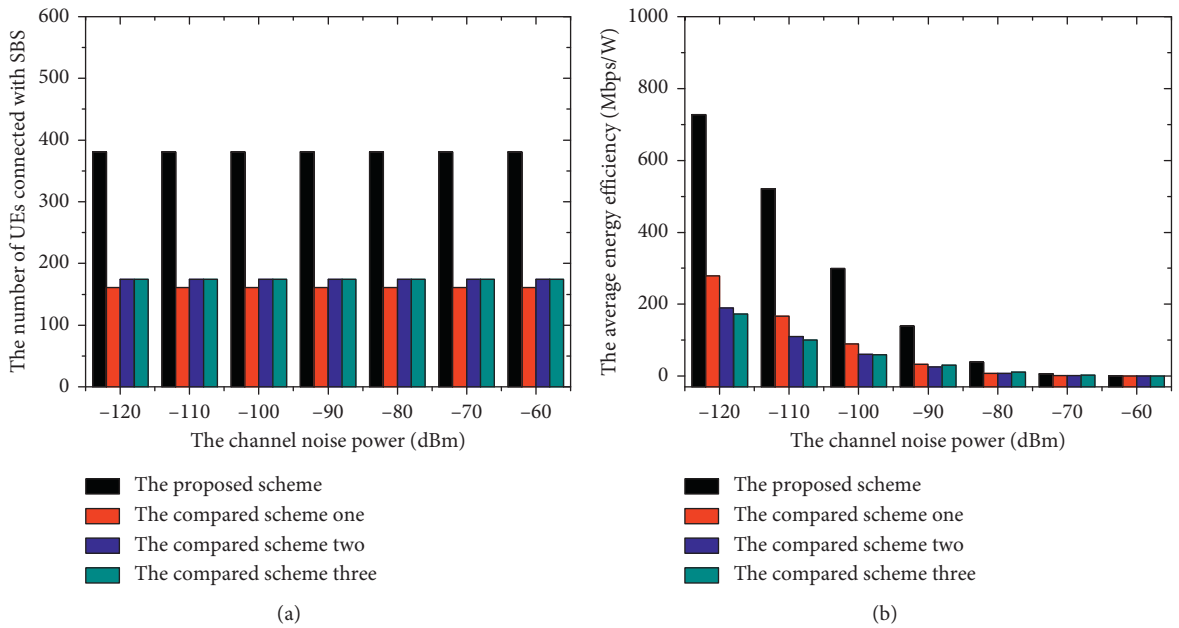

Figure 6: The performance variation trend versus the channel noise power for the above four schemes.

certain range. However, with the increase in transmission power, the increase in data rate becomes smaller and smaller according to Shannon theorem. The simulation results in Figure 6 also demonstrate that the proposed scheme outperforms the other three schemes in the same two performance aspects, where the interpretation for Figure 5 can be applied to Figure 6.

When the channel noise power, the number of UEs, and the threshold r_{th} for energy efficiency are set as −100 dBm, 1000, and 700 Mbps/W respectively, the performance variation trend with the ratio of working UEs is shown in Figure 7. As shown in Figure 7(a), the number of UEs connected with SBS grows monotonically with the ratio of working UEs in the four schemes. The reason is obviously that the number of UEs that can communicate with SBSs in a direct or indirect manner is positively correlated with the number of working UEs.

Figure 7(b) shows that the average energy efficiency of UEs decreases slightly with the ratio of working UEs in the proposed scheme, while it hardly changes in the other three schemes. There are two main reasons. First, as the ratio of working UEs increases, the number of idle UEs decreases,

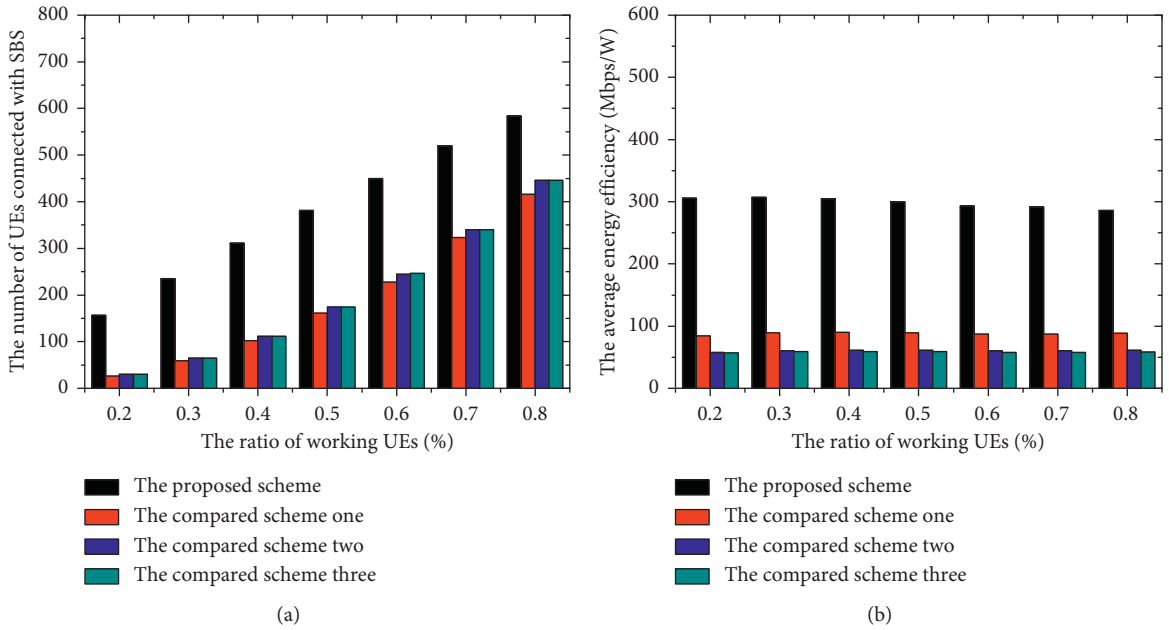

Figure 7: The performance variation trend versus the number of UEs under different parameter combinations.

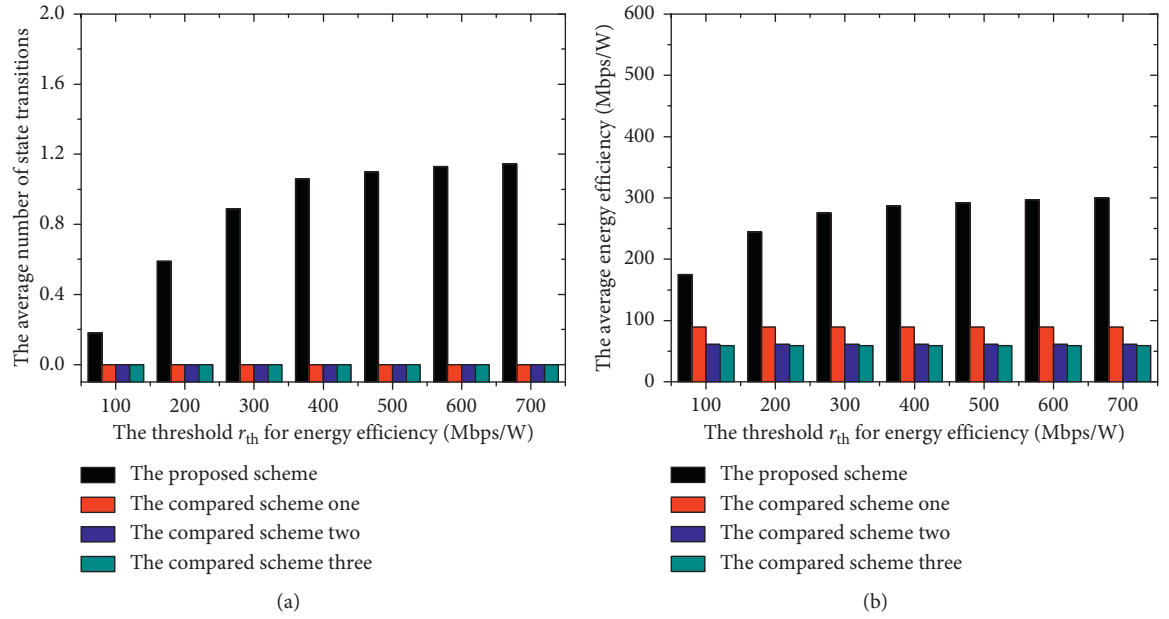

Figure 8: The performance variation trend versus the threshold r_{th} for energy efficiency under different parameter combinations.

and thus there are fewer candidates relaying UEs. Second, more working UEs means more demand for relaying UEs. Therefore, in the case of a higher ratio of working UEs, there is less probability that a working UE requiring a relay can choose a suitable relay, which will affect its energy efficiency. However, the other three schemes do not involve the use of relays, and thus they are hardly affected by the ratio of working UEs.

When the channel noise power is fixed as −100 dBm and the number of UEs is fixed as 1000, the performance variation trend with the threshold r_{th} for energy efficiency is shown in Figure 8, which shows that both the average number of state transitions and the average energy efficiency increase with the threshold r_{th} for energy efficiency in the proposed scheme. The results confirm the analysis in Section 4.5. However, the other three schemes do not involve the state transition, and thus they are hardly affected by the threshold r_{th} for energy efficiency. Figure 8(a) shows the mean of the results over 10 time intervals. When the threshold value is relatively low, there are hardly state transitions in most time intervals. Therefore, the average number of state transitions is less than one.

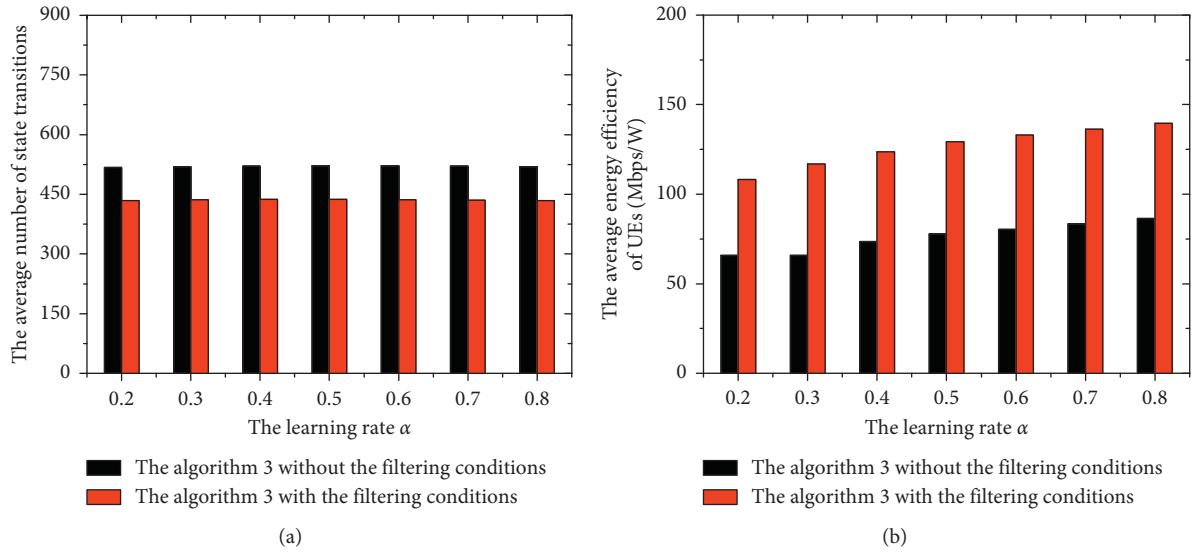

Figure 9: The performance variation trend versus the value of α.

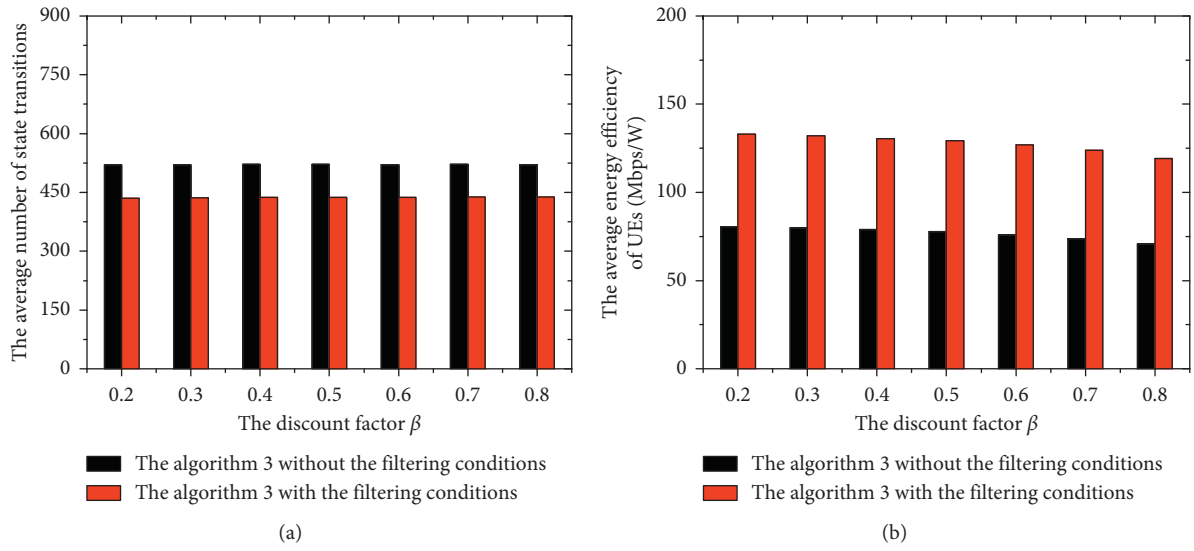

Figure 10: The performance variation trend versus the value of β.

In addition to the above simulations, we also compare the performance of Algorithm 3 before and after adding the filter conditions. When the discount factor β is fixed as 0.5, the performance variation trend with the learning rate α is shown in Figure 9. Also, when the learning rate α is fixed as 0.5, the performance variation trend with the discount factor β is shown in Figure 10. Since the number of episodes is set to a fixed value (e.g., 1000 in our simulations), the average number of state transitions can be used to measure the running cost of the different schemes indirectly. In Figures 9 and 10, the energy efficiency refers to the best energy efficiency value of a training path (corresponding to an episode), while the average energy efficiency is an average value in terms of the best energy efficiency values in all the training paths.

From Figures 9(a) and 10(a), we can observe that the Algorithm 3 with the filtering conditions is clearly superior to that without the filtering conditions in terms of the average number of state transitions. This shows that it does reduce running overhead for an agent to filter some unnecessary initial states during the training process. Also, Figures 9(a) and 10(a) show that the average number of state transitions hardly changes as the learning rate α and the discount factor β change. The main reason is that the number of state transitions depends on the selection of initial states and the number of training rounds but has little relationship with α and β.

From Figures 9(b) and 10(b), we also see that Algorithm 3 with the filtering conditions outperforms that without the filtering conditions in terms of the average energy efficiency. The reason behind this phenomenon is mainly attributed to that the state transitions from some initial states can fail to achieve a desired state that has a good

energy efficiency. Therefore, it is necessary that these initial states will be filtered out during the training process. In practice, UEs rarely encounter these initial states. In addition, Figure 9(b) shows that the average energy efficiency increases as the learning rate α increases, while Figure 10(b) shows that the average energy efficiency decreases as the discount factor β increases.

This is because a larger value of α means that an agent pays more attention to the immediate interests at present instead of the past rewards. Since the past rewards do not necessarily adapt to the present situation, a larger value of α helps speed up the training process and thus gets better training results. Also, a larger value of β means that an agent attaches greater importance to future rewards instead of the current rewards. Due to the uncertainty of future expectations, a larger value of β does not helps speed up the training process and thus hardly improves training results.

5.3. Comparison of Q-Learning-Based Method and Online Learning Solution. As mentioned in the introduction, both Q-learning model and MAB theoretical model are in the category of reinforcement learning. Therefore, in this section, we briefly examine the performance of these two models used to address the concerns of this paper. Through modeling our concern problem as a contextual multiarmed bandit problem, we design a contextual online learning algorithm to compare with our Q-learning method. Based on the expression (5) defined in Subsection 3.2, we define $b_{i,j} \times d_{i,j}$ of the expression (5) as the context space of UE i under the coverage of SBS j, which is divided into six context subspaces.

Also, we define $p_{i,j} \times g_{i,j} \times l_{i,j}$ of the expression (5) as the set of resources from which UE i can request. According to the actual state of UE i, the context subspace to which it belongs can be found from the context space $b_{i,j} \times d_{i,j}$. Under this context subspace, for each of the set of resources $p_{i,j} \times g_{i,j} \times l_{i,j}$, there is a corresponding state that belongs to one state of the state space defined in the expression (5), and thus the performance under this state can be determined by formula (9).

Using the same design idea of Algorithm 1 in [39], we can design a similar contextual online learning algorithm, which learns the expected resource performances under different contexts online over time. This algorithm works on the assumption that for, similar UE contexts, the performance of a particular resource will on average be similar. This contextual online learning algorithm performs the following steps for UE i. It first uniformly partitions the context space into six context subspaces and learns about the performance of different resources independently in each of these context subspaces.

Then, in each period, the algorithm performs either an exploration or an exploitation. A control function is used to determine which phase it enters. In exploration phases, the algorithm randomly selects one from the set of resources for UE i, while in exploitation phases, the algorithm selects the resource that showed the best performance when selected in previous periods. By observing the amount of data transmitted by UE i, the algorithm acquires performance estimate of the selected resource. Therefore, it learns the performance of the different resources under different UE contexts over time. In order to save the paper space, we omit the algorithm description. Please refer to [39] for the algorithm description details. The comparison of the characteristics of the two types of learning algorithms is listed in Table 1.

Although the training phase of Q-learning algorithm requires lots of computing resources, the speed at which decisions are made for UE i according to a well-trained Q table is faster than that of the online learning solution that does not need an additional training phase, which is also illustrated by the results in Figure 11. Based on the above simulation settings, in our Q-learning-based method, the average decision overhead can be approximated by averaging the number of state transitions across all the feasible decision paths on the well-trained Q table.

For the contextual online learning solution, the size of its resource space is same as that of action space of our Q-learning-based method. When UE i knows its context subspace, it still needs to traverse its resource space to arrive at the decision result. Therefore, the contextual online learning solution requires at most n comparisons to obtain the best resource if a simple enumeration mode is adopted, where n represents the size of the resource space of UE i.

If we rank the performance metrics of resources in its resource space, we can directly select the best performing resource that is known at decision time. However, the performance metric needs to be updated after each decision based on the actual performance of the resource, so the newly updated value needs to be reordered in the original ordered table to facilitate the next decision, which cannot be omitted. If we find the new position of the updated value in the original ordered table based on the binary search method, the maximum number of operations is $\log_2 n$, and thus the corresponding average value is about $(\log_2 n/2)$.

Based on the above simulation settings, the average number of state transitions across all the feasible decision paths on the well-trained Q table with 192 states and 32 actions fluctuates between 1.2 and 1.3 when the number of UEs varies between 600 and 1200, while it fluctuates between 1.1 and 1.3 when the channel noise power varies between −110 dBm and −60 dBm.

For the contextual online learning solution, since the size of its resource space is same as that of action space of our Q-learning-based method, n takes 32 and thus $(\log_2 n/2) = 2.5$. Therefore, the overhead of updating the original ordered table after each decision can be approximated as 2.5, which can be used to approximate the decision process overhead of the contextual online learning solution. For intuition, we show the results in Figures 11(a) and 11(b).

When the channel noise power is less than −110 dBm, it indicates that the channel conditions are very good, where no matter what the initial state is, the room to improve performance by changing the state is very limited, so the average number of state transitions is very small in our Q-learning-based method. However, in the contextual online learning solution, the decision process overhead is essentially unchanged. As a result, the ratio of decision

Table 1: The comparison of the characteristics of the two types of learning algorithms.

The type of learning algorithm	Advantages	Disadvantages
Q-learning-based method	When the Q table is fully trained, the decision-making speed based on the Q table is very fast	Full training of the Q tables requires sufficient computing resources and takes a long time
Online learning solution	No training process is required, and thus training resources are saved	Decision process is slower than the well-trained Q table-based decision process

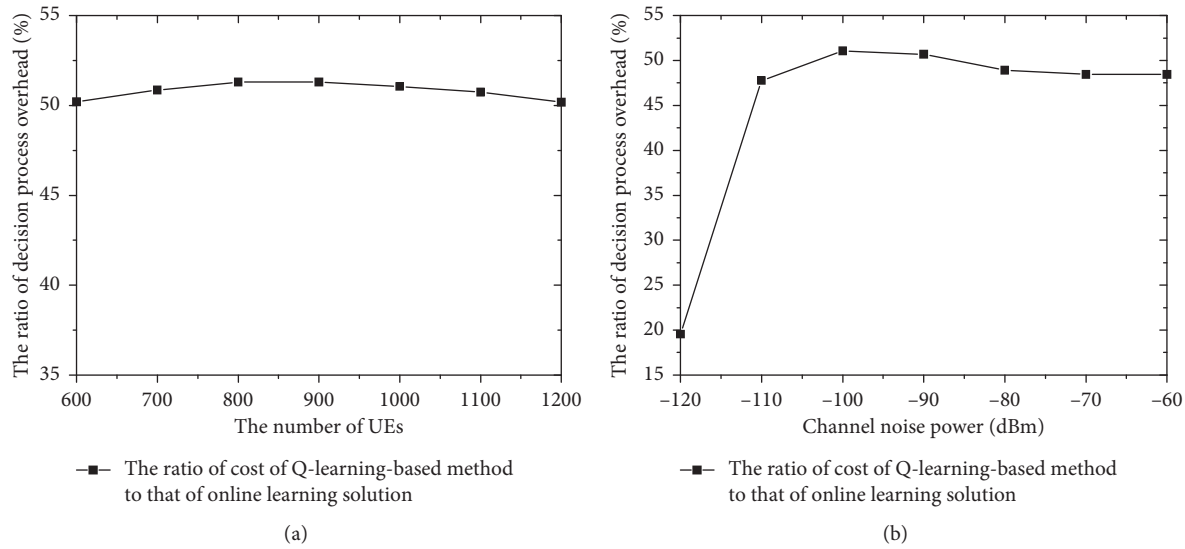

Figure 11: The ratio of the decision process overhead in the Q-learning method to that in online learning solution.

overhead in the Q-learning method to that in online learning solution drops abruptly.

6. Conclusions

In this paper, we investigated the transmission capacity problem in mmWave networks and proposed a Q-learn-based scheme to stabilize transmission capacity in mmWave links from an energy efficiency optimization perspective. The proposed scheme was compared with the other three schemes in terms of the number of UEs connected with SBS, the average number of state transitions, and the average energy efficiency. Also, we discussed how to reduce the cost of the Q table training process. The simulation results show that the proposed scheme keeps the most number of UEs connected with SBS, while it also achieves the best average energy efficiency among the four schemes. At the same time, the simulation results show that the Q table training process can be accelerated by filtering some unnecessary states and actions, and Q table performance can also meet the decision requirements.

Data Availability

The simulation data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant (no. 61873352 and 61803387).

References

[1] J. Tan, W. Liu, T. Wang, M. Zhao, A. Liu, and S. Zhang, "A high-accurate content popularity prediction computational modelling for mobile edge computing by using matrix completion technology," *Transactions on Emerging Telecommunications Technologies*, 2020.

[2] Y. Liu, Z. Zeng, X. Liu, X. Zhu, and M. Bhuiyan, "A novel load balancing and low response delay framework for edge-cloud network based on SDN," *IEEE Internet of Things Journal*, p. 1, 2019.

[3] J. Luo, X. H. Deng, H. G. Zhang, and H. M. Qi, "QoE-driven computation offloading for edge computing," *Journal of Systems Architecture*, vol. 97, pp. 34–39, 2019.

[4] M. Huang, W. Liu, T. Wang, A. Liu, and S. Zhang, "A cloud-MEC collaborative task offloading scheme with service orchestration," *IEEE Internet of Things Journal*, p. 1, 2019.

[5] M. Chen, T. Wang, K. Ota, M. Dong, M. Zhao, and A. Liu, "Intelligent resource allocation management for vehicles network: an A3C learning approach," *Computer Communications*, vol. 151, pp. 485–494, 2020.

[6] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2133–2168, 2018.

[7] M. Peng, W. Liu, T. Wang, and Z. Zeng, "Relay selection joint consecutive packet routing scheme to improve performance for wake-up radio-enabled WSNs," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 7230565, 32 pages, 2020.

[8] Z. Li and J. Gui, "Energy-efficient resource allocation with hybrid TDMA-NOMA for cellular-enabled machine-to-machine communications," *IEEE Access*, vol. 7, no. 1, pp. 105800–105815, 2019.

[9] X. Liu, A. Liu, T. Wang et al., "Adaptive data and verified message disjoint security routing for gathering big data in energy harvesting networks," *Journal of Parallel and Distributed Computing*, vol. 135, pp. 140–155, 2020.

[10] Z. Kuang, G. Liu, G. Li, and X. Deng, "Energy efficient resource allocation algorithm in energy harvesting-based D2D heterogeneous networks," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 557–567, 2019.

[11] X. Deng, J. Luo, L. He, Q. Liu, X. Li, and L. Cai, "Cooperative channel allocation and scheduling in multi-interface wireless mesh networks," *Peer-to-Peer Networking and Applications*, vol. 12, no. 1, pp. 1–12, 2019.

[12] L. Yin, J. S. Gui, and Z. W. Zeng, "Improving energy efficiency of multimedia content dissemination by adaptive clustering and D2D multicast," *Mobile Information Systems*, vol. 2019, Article ID 5298508, 16 pages, 2019.

[13] J. Gui, L. Hui, and X. Zhou, "Improving lifetime of cell-edge smart sensing devices by incentive architecture based on dynamic charging," *IEEE Access*, vol. 7, no. 1, pp. 72703–72715, 2019.

[14] J. Deng, O. Tirkkonen, R. Freij-Hollanti, T. Chen, and N. Nikaein, "Resource allocation and interference management for opportunistic relaying in integrated mmWave/sub-6 GHz 5G networks," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 94–101, 2017.

[15] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5G network capacity: key elements and technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, 2014.

[16] W. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 106–112, 2014.

[17] A. Gupta and R. K. Jha, "A survey of 5G network: architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[18] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter wave massive MIMO communication for future wireless systems: a survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 836–869, 2018.

[19] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029–3056, 2015.

[20] M. R. Akdeniz, Y. Liu, M. K. Samimi et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[21] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2015.

[22] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: a MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, 2015.

[23] Qualcomm Unveils First mmWave 5G Antennas for Smartphones, https://www.theverge.com/2018/7/23/17596746/qualcomm-mmwave-5g-antenna-smartphones-qtm052-networking-speeds-size Qualcomm Unveils First mmWave 5G Antennas for Smartphones.

[24] S. K. Haider, A. Jiang, M. A. Jamshed, H. Pervaiz, and S. Mumtaz, "Performance enhancement in P300 ERP single trial by machine learning adaptive denoising mechanism," *IEEE Networking Letters*, vol. 1, no. 1, pp. 26–29, 2019.

[25] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5G small cells," *IEEE Wireless Communications*, vol. 23, no. 3, pp. 64–73, 2016.

[26] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[27] G. Alnwaimi, S. Vahid, and K. Moessner, "Dynamic heterogeneous learning games for opportunistic access in LTE-based macro/femtocell deployments," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2294–2308, 2015.

[28] O. Onireti, A. Zoha, J. Moysen et al., "A cell outage management framework for dense heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2097–2113, 2016.

[29] Z. B. Gao, B. Wen, L. F. Huang, C. B. Chen, and Z. W. Su, "Q-learning-based power control for LTE enterprise femtocell networks," *IEEE Systems Journal*, vol. 11, no. 4, pp. 2699–2707, 2017.

[30] Z. Wei, Y. Zhang, X. Xu, L. Shi, and L. Feng, "A task scheduling algorithm based on Q-learning and shared value function for WSNs," *Computer Networks*, vol. 126, pp. 141–149, 2017.

[31] H. Bayat-Yeganeh, V. Shah-Mansouri, and H. Kebriaei, "A multi-state Q-learning based CSMA MAC protocol for wireless networks," *Wireless Networks*, vol. 24, no. 4, pp. 1251–1264, 2018.

[32] J. Zhu, Y. Song, D. Jiang, and H. Song, "A new deep-Q-learning-based transmission scheduling mechanism for the cognitive internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2375–2385, 2018.

[33] A. Carie, M. Li, C. Liu, P. Reddy, and W. Jamal, "Hybrid directional CR-MAC based on Q-learning with directional power control," *Future Generation Computer Systems*, vol. 81, pp. 340–347, 2018.

[34] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multi-agent reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4539–4551, 2018.

[35] B. Malila, O. Falowo, and N. Ventura, "Intelligent NLOS backhaul for 5G small cells," *IEEE Communications Letters*, vol. 22, no. 1, pp. 189–192, 2018.

[36] J. Gui, Y. Lu, X. Deng, and A. Liu, "Flexible resource allocation adaptive to communication strategy selection for cellular clients using stackelberg game," *Ad Hoc Networks*, vol. 66, no. 11, pp. 64–84, 2017.

[37] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 2002.

[38] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213–6230, 2017.

[39] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2487–2500, 2018.