Received February 3, 2022, accepted February 15, 2022, date of publication February 17, 2022, date of current version March 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152545

Load-Aware Dynamic Mode Selection for Network-Assisted Full-Duplex Cell-Free Large-Scale Distributed MIMO Systems

YUE ZHU [1], (Graduate Student Member, IEEE), JIAMIN LI [1,2], (Member, IEEE), PENGCHENG ZHU [1], (Member, IEEE), DONGMING WANG [1,2], (Member, IEEE), HENG YE [3], (Member, IEEE), AND XIAOHU YOU [1,2], (Fellow, IEEE)

[1] National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China
[2] Purple Mountain Laboratories, Nanjing 211111, China
[3] Intel China Research Center Ltd., Beijing 100080, China

Corresponding author: Jiamin Li (jiaminli@seu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2900300; and in part by the National Natural Science Foundation of China (NSFC) under Grant 61971127, Grant 61871465, and Grant 61871122.

ABSTRACT The network-assisted full-duplex (NAFD) system realizes flexible duplex in the spatial domain within the same time-frequency resource. With the explosive growth of the number of users and remote antenna units (RAUs) under 6G scenario, the resource utilization of the system is lower. When the resource of users is selected by the RAUs to send or receive, collisions or congestion may occur due to mechanisms such as grant-free. Aiming at making better use of system resources, a load-aware dynamic mode selection scheme with NAFD scheme is proposed to improve the access efficiency and resource utility of the system. This paper first propose a centralized Q-learning algorithm which determines a clever strategy to approach the ultimate goal by itself and excels in environment dynamics. However, the size of the Q-table used in the centralized Q-learning algorithm for storage is huge. Further, a distributed multi-agent Q-learning algorithm is proposed which has a smaller size of Q-table and lower complexity to suit for actual scenarios. The simulation results showed that the proposed load-aware dynamic mode selection scheme can significantly improve resource utility and throughput performance than other traditional schemes.

INDEX TERMS Full duplex, load-aware, dynamic mode selection, Q-learning.

I. INTRODUCTION

Currently, ultra-reliable and low-latency communication (URLLC) related theories and technologies in 6G are in urgent need of breakthrough. In order to reduce the traditional half duplex (HD) system latency, a full duplex (FD), equipped with transmit antennas and receive antennas, has been widely studied in the literature to enable simultaneous transmission and reception in the same frequency band, with theoretically doubled throughput [1]. Self interference (SI) is the main barrier in implementing FD. Active and passive SI suppression techniques have been studied in [2], [3], which makes FD a realistic technology for modern wireless systems. [4] studied a FD cell-free massive multiple-input multiple-output (MIMO) network, where the APs employ a simple conjugate beamforming/matched filtering scheme with the channel state information acquired through the uplink training with orthogonal pilots transmitted from the users. It provided a simple power control method to mitigate residual self-interference. [5] derived closed-form spectral efficiency (SE) lower bounds for FD cell-free MIMO system with maximum-ratio combining/maximum-ratio transmission processing and optimal uniform quantization. Recently, the problem of maximization of SE and energy efficiency (EE) of the FD cell-free MIMO system is considered in [6]. However, to achieve URLLC, SI cancellation processing latency in FD system should be considered [7]. Network-assisted full-duplex (NAFD) under cell-free massive MIMO network was proposed in [8], which does not have SI at the remote antenna unit (RAU) level and can solve the cross-link interference (CLI) problem by using joint processing [9] thus reducing the latency of interference cancellation. It realizes

The associate editor coordinating the review of this manuscript and approving it for publication was Cunhua Pan.

flexible duplex transmission by selecting the uplink and downlink working modes of the RAUs in the spatial domain within the same time-frequency resource block at the network level, reducing the delay of the time division duplex (TDD) system and improving the spectral efficiency (SE) and energy efficiency (EE) of the system. NAFD scheme is quite promising to achieve URLLC due to the reduction of latency in HD and TDD system and SI cancellation latency in traditional full-duplex scheme.

With the explosive growth of the number of mobile terminals, it is necessary to study the reliable multi-access and resource utilization mechanism of NAFD scheme. To improve the utilization of UL/DL resources, the working mode selection of RAUs as uplink reception or downlink transmission based on the traffic loads and quality of service (QoS) of the users is investigated in this paper. Proposed load-aware dynamic mode selection scheme directly assigns DL transmitting or UL receiving for each RAU according to the traffic loads of the whole network. There is no need for the users to establish a handshake mechanism with RAUs which reduces signaling overhead, simplifies the access procedure, reduces the access latency, and improves the access efficiency of the massive access scenario.

On the other hand, most of the work usually fix the working mode of the RAUs to further analyze the system performance with NAFD scheme [10]–[12]. [10] estimated the effective CSI (inner products of beamforming and channel vectors) instead based on beamforming training scheme. [11] investigated the problem of joint transceiver design for NAFD systems under cell-free massive MIMO network with simultaneous wireless information and power transfer (SWIPT) considering the fronthaul capacity as a constraint. [12] focused on the optimization of SE of the systems taking SWIPT ratio design into consideration. [13] only focused on the problem of maximization of SE through mode selection scheme where the traffic loads and quality of service (QoS) of the users has not been considered. To the best our understanding, the working mode selection problem of the RAUs considering the traffic loads and QoS of the users has not yet been explored in the literature. As far as we know, there are no researches focusing on the resource utilization of the FD cell-free MIMO systems. This paper focuses on the resource utilization problem and system performance of NAFD cell-free large-scale distributed MIMO systems. In this paper, a load-aware dynamic mode selection scheme with flexible duplexing based on reinforcement learning is proposed. The load-aware technic has been studied in [14] and [15], where the authors proposed the load-aware system utility function based on their assumed scenarios to reflect the proposed performance improvement according the traffic loads. Specifically, in this paper, the reinforcement learning method that maximizes the expected benefits in a dynamic environment is used to optimize the working mode of the RAUs for uplink reception or downlink transmission. Q-learning is a classic method of reinforcement learning which does not require a deep neural network for function approximation [16]–[19]. The load-aware dynamic mode selection scheme based on Q-learning approaches the ultimate goal by taking clever strategies and excels in environment dynamics. The proposed algorithms can be utilized in practical RAU-mode-selection scenario with limited computation power. We initially proposed a centralized Q-learning algorithm which viewed all RAUs as an agent. However, this algorithm has the problem of explosive growth of the size of Q-table for storage. We further proposed a distributed multi-agent Q-learning method. The distributed algorithm viewed each RAU as an independent agent and thus had a smaller size of storage unit with lower complexity.

The main contributions of this paper are highlighted as follows:

- The dynamic mode selection scheme is hard to be modeled in the actual scene. In order to determine the RAUs' working mode, two binary assignment vectors $\mathbf{x}_{\mathrm{u}}, \mathbf{x}_{\mathrm{d}} \in \{0, 1\}^{M\times 1}$ are used to model the mode selection problem. To improve the resource utility and access efficiency of the system, a load-aware dynamic mode selection scheme is further proposed.
- A utility function is defined to reflect both the proposed performance improvement in resource allocation and the associated overhead costs of any coalition formation. The defined utility function leads the UEs preferentially select RAUs with fewer RB resources under the premise of satisfying its own QoS, thereby improving the utilization of RBs.
- Q-learning is a classic method of reinforcement learning which does not require a deep neural network for function approximation which can be utilized in practical RAU-mode-selection scenario with limited computation power. A load-aware dynamic mode selection scheme based on centralized Q-learning is proposed to solve the resource utilization problem.
- We further proposed a distributed multi-agent Q-learning method to avoid the problem of explosive growth of the size of Q-table for storage in centralized Q-learning. The distributed algorithm viewed each RAU as an independent agent and thus had a smaller size of storage unit with lower complexity. The effectiveness of the proposed scheme was verified through simulations.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

We consider a network-assisted full-duplex cell-free large-scale distributed MIMO system with M_{u} RAUs performing uplink reception and M_{d} RAUs performing downlink transmission at each time slot, where $M_{\mathrm{u}} + M_{\mathrm{d}} = M$. Each RAU is equipped with N half-duplex(HD) antennas, while the terminals are single-antenna and HD capable as illustrated in Fig.1. RAUs serve arbitrarily distributed K_{u} uplink users and K_{d} downlink users, abandoning the traditional cell structure.

Previous works focused on the performance analysis of NAFD scheme in the fixed mode, which equally split the

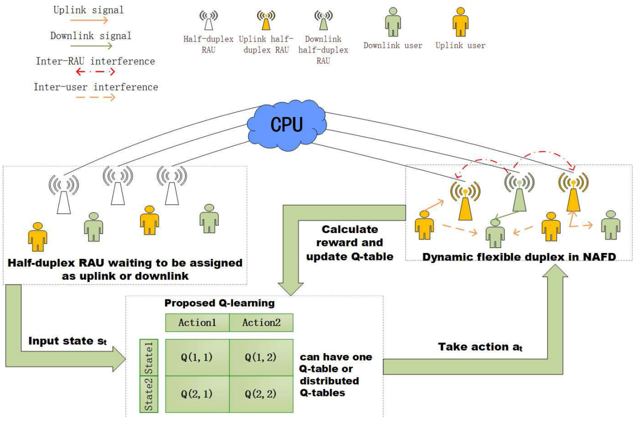

FIGURE 1. System model of the proposed Q-learning framework.

RAUs to uplink and downlink mode. Fixed-mode encounters bottlenecks in the use of UL and DL resources. So dynamic RAU mode selection plays a significant role in improving the system performance. To improve the resource utility and access efficiency of the system, a load-aware dynamic mode selection scheme is further proposed. In order to determine the RAUs' working mode, two binary assignment vectors $\mathbf{x}_{\mathrm{u}}, \mathbf{x}_{\mathrm{d}} \in \{0, 1\}^{M\times 1}$ are used to model the mode selection problem. $x_{\mathrm{u},i}(x_{\mathrm{d},j})$ is equal to 1 if RAU $i(j)$ is used for UL reception (DL transmission), or equal to 0 otherwise. This paper assumes all the antennas of the same RAU working in the same mode and an RAU is either uplink or downlink only, which means $\mathbf{x}_{\mathrm{u}} + \mathbf{x}_{\mathrm{d}} = 1$. We define $\mathbf{X}_{\mathrm{u}} = \mathrm{diag}(\mathbf{x}_{\mathrm{u}})$ and $\mathbf{X}_{\mathrm{d}} = \mathrm{diag}(\mathbf{x}_{\mathrm{d}})$, and derive the effective received signal, transmitted signal and the load-aware mode selection model.

B. DOWNLINK SIGNAL MODEL

For downlink transmission, the baseband signals are compressed by the central processing unit (CPU) and conveyed to each downlink RAUs through the downlink fronthaul links. The downlink RAUs decompress the downlink signals received from the CPU and then forward them to the downlink users.

For downlink transmission, in each scheduled time slot, M_{d} downlink RAUs jointly send signals to K_{d} downlink users. Specifically, the signal received by the DL user j can be expressed as

$$\tilde{r}_{\mathrm{d},j} = \tilde{\boldsymbol{h}}_{\mathrm{d},j}^{\mathrm{H}} \sum_{m=1}^{K_{\mathrm{d}}} \mathbf{w}_{\mathrm{d},m} s_{\mathrm{d},m} + \sum_{i=1}^{K_{\mathrm{u}}} g_{i,j}\sqrt{p_{\mathrm{u},i}}x_i + \eta_{\mathrm{d},j}, \tag{1}$$

where $\tilde{\boldsymbol{h}}_{\mathrm{d},j} = \mathbf{X}_{\mathrm{d}}\mathbf{h}_{\mathrm{d},j}$ denote the effective DL channel vector, $\mathbf{h}_{\mathrm{d},j} \in \mathbb{C}^{M_{\mathrm{d}}N\times 1}$ denotes the channel vector between the receiving DL user j and all the downlink RAUs, which can be modeled as $\mathbf{h}_{\mathrm{d},j} = \sqrt{\boldsymbol{\Lambda}_{\mathrm{d},j}}\mathbf{g}_{\mathrm{d},j}$, where $\boldsymbol{\Lambda}_{\mathrm{d},j} = \mathrm{diag}(\lambda_{\mathrm{d},j,1}, \lambda_{\mathrm{d},j,2} \ldots \lambda_{\mathrm{d},j,M_{\mathrm{d}}}) \otimes \mathbf{I}_N$, $\lambda_{\mathrm{d},j,m} = d_{j,m}^{-\alpha_{\mathrm{d}}}$ represents large-scale fading, $d_{j,m}$ denotes the corresponding distance between RAUs and UEs, α_{d} and $\mathbf{g}_{\mathrm{d},j}$ denote path-loss and irrelevant small-scale fast fading respectively, $\mathbf{w}_{\mathrm{d},m}$ is the downlink precoding vector, $s_{\mathrm{d},m}$ denotes the transmitted data signal with $\mathbb{E}[s_{\mathrm{d},m}s_{\mathrm{d},m}^{\mathrm{H}}] = 1$ in the DL, $p_{\mathrm{u},i}$ denotes the uplink transmit power of the user i, x_i denotes the transmitted data signal with $\mathbb{E}[x_i x_i^{\mathrm{H}}] = 1$ by UL user i, and $g_{i,j}$ denotes the interfering channel vector between the UL transmitter user equipment (UE) i and the DL receiver UE j, $\eta_{\mathrm{d},j} \sim \mathcal{CN}(0, \sigma_{\mathrm{d}}^2)$ is additive white Gaussian noise at DL user j. Then, the signal-to-interference-plus-noise ratios (SINR) of downlink user j can be expressed as

$$\gamma_{\mathrm{d},j} = \frac{|\tilde{\boldsymbol{h}}_{\mathrm{d},j}^{\mathrm{H}}\mathbf{w}_{\mathrm{d},j}|^2}{\sum\limits_{j'\neq j, j'\in K_{\mathrm{d}}} |\tilde{\boldsymbol{h}}_{\mathrm{d},j}^{\mathrm{H}}\mathbf{w}_{\mathrm{d},j'}|^2 + \sum\limits_{i\in K_{\mathrm{u}}} p_{\mathrm{u},i}|g_{i,j}|^2 + \sigma_{\mathrm{d}}^2}, \tag{2}$$

where $R_{\mathrm{D},j}$ represents the downlink rate which can be denoted as $R_{\mathrm{D},j} = \log_2(1 + \gamma_{\mathrm{d},j})$.

C. UPLINK SIGNAL MODEL

For uplink transmission, each uplink RAU compresses the received signals that are transmitted by uplink users and sends them to the CPU. The CPU receives the compressed signals that are transmitted by all uplink RAUs and then performs joint decoding of all uplink users based on the received compressed signals.

For uplink transmission, all uplink RAUs jointly receive signals from UL users. The received signal can be expressed as

$$\tilde{\boldsymbol{r}}_{\mathrm{u}} = \sum_{i=1}^{K_{\mathrm{u}}} \tilde{\boldsymbol{h}}_{\mathrm{u},i}\sqrt{p_{\mathrm{u},i}}x_i + \sum_{j=1}^{K_{\mathrm{d}}} \tilde{\boldsymbol{G}}_{\mathrm{I}}\mathbf{w}_j s_{\mathrm{u},j} + \tilde{\boldsymbol{\eta}}_{\mathrm{u}}, \tag{3}$$

where $\tilde{\boldsymbol{h}}_{\mathrm{u},i} = \mathbf{X}_{\mathrm{u}}\mathbf{h}_{\mathrm{u},i}$, $\tilde{\boldsymbol{G}}_{\mathrm{I}} = \mathbf{X}_{\mathrm{u}}\mathbf{G}_{\mathrm{I}}\mathbf{X}_{\mathrm{d}}$ denote the effective UL channel vector and the channel vector between uplink RAUs and downlink RAUs, $\tilde{\boldsymbol{\eta}}_{\mathrm{u}} = \mathbf{X}_{\mathrm{u}}\boldsymbol{\eta}_{\mathrm{u}}$ denotes the effective noise with distributions $\tilde{\boldsymbol{\eta}}_{\mathrm{u}} \sim \mathcal{CN}(0_{M_{\mathrm{u}}}, \sigma^2\mathbf{X}_{\mathrm{u}})$, $\tilde{\boldsymbol{h}}_{\mathrm{u},i} \in C^{M_{\mathrm{u}}N\times 1}$ denotes the channel vector between the transmitting UL user i and all the uplink RAUs, $\mathbf{G}_{\mathrm{I}} \in C^{M_{\mathrm{u}}N\times M_{\mathrm{d}}N}$ is the real interference channel matrix between downlink RAUs and uplink RAUs. In practice, we assume the channel state information (CSI) between uplink-RAUs and downlink-RAUs is imperfect due to the channel estimation errors. Specifically, we model the inter-RAU channel as:$\mathbf{G}_{\mathbf{I}} = \mathbf{G}_{\mathbf{IRI}}{}' + \mathbf{G}_{\mathbf{IRI}}{}''$, where $\mathbf{G}_{\mathbf{IRI}}{}'$ denotes the estimated channel and $\mathbf{G}_{\mathbf{IRI}}{}''$ denotes the channel estimation error. Then, the SINR for each uplink user i can be expressed as

$$\gamma_{\mathrm{u},i} = \frac{p_{\mathrm{u},i}|\mathbf{v}_{\mathrm{u},i}^{\mathrm{H}}\tilde{\boldsymbol{h}}_{\mathrm{u},i}|^2}{\sum\limits_{i'\neq i, i'\in K_{\mathrm{u}}} p_{\mathrm{u},i'}|\mathbf{v}_{\mathrm{u},i}^{\mathrm{H}}\tilde{\boldsymbol{h}}_{\mathrm{u},i}|^2 + \sum\limits_{m\in M} \sigma_m^2\left\|\mathbf{v}_{\mathrm{u},i,m}\right\|^2 + \mu_i}, \tag{4}$$

where μ_i denotes the interference between DL users and UL users, $\mu_i = \sum_{j\in K_{\mathrm{d}}} \psi_{\mathrm{d},j}\left\|\mathbf{v}_{\mathrm{u},i}\right\|^2$, $\psi_{\mathrm{d},j}$ and $\mathbf{v}_{\mathrm{u},i}$ are the variance of the total interference plus noise in the DL and the corresponding receiver vector respectively. $R_{\mathrm{U},i}$ represents the uplink and rate which can be denoted as $R_{\mathrm{U},i}$=$\log_2$(1+$\gamma_{\mathrm{u},i}$).

D. MODEL FOR THE PROPOSED LOAD-AWARE DYNAMIC MODE SELECTION SCHEME

The load of the system is an important indicator as it greatly affects the performance of multi-connectivity. In multi-connectivity, the user equipments must avoid accessing to RAUs with higher load to the extent possible to achieve system load balancing. In order to improve the resource utility of the system, a load-aware dynamic mode selection scheme is proposed. We assume that each UE has a requirement of resource blocks (RBs) to meet the need of QoS. RAU m has k RBs to allocate, which denotes as $\lambda_m = \{\lambda_{m,1}, \lambda_{m,2}, \ldots, \lambda_{m,k}\}$. The total transmit power is P_m. Evenly allocated transmission power is assumed, then the power of each RB is $p_m = \frac{P_m}{k}$. The number of RBs allocated by RAU m to UE i is $n_{m,i}$, then the power allocated by RAU m to UE i is $p_{m,i} = n_{m,i}p_m$. Considering the traffic load, the achievable rate of UE i associated with RAU m can be expressed as

$$T_i = n_{m,i} b \log_2(1+\gamma_i), \tag{5}$$

where b represents the bandwidth of each RB [20]- [21]. It can be seen that the user rate is related to the number of RBs that can be used. Each user has different service requests and different QoS requirements. RAU guarantees users' QoS by allocating a certain number of RBs to different users. In order to meet the QoS needs of different users, let t_i^{req} denote the bandwidth request of user i

$$n_{m,i} b \log_2(1+\gamma_i) \geq t_i^{\mathrm{req}}, \tag{6}$$

It can be obtained that the number of RBs provided by RAU m for user i is

$$n_{m,i} = \left\lceil \frac{t_i^{\mathrm{req}}}{b\log_2(1+\gamma_i)} \right\rceil, \tag{7}$$

[14] and [15] considering load-aware methods both proposed the load-aware utility function suitable for their research scenarios. Utility function should reflect both the proposed performance improvement in resource allocation and the associated overhead costs of any coalition formation. We propose the utility function for the UE i as U_i

$$U_i = \ln\left(1 + \frac{k - \sum_{a\in K} n_{m,a}}{n_{m,i}}\right), \tag{8}$$

The utility of the UE at each time slot depends on the load of the associated RAUs and the number of RBs allocated to it. If the RAU cannot guarantee the QoS of UE i, the RAU does not provide RBs for UE i. In this case, the value of the utility function for UE i $U_i = 0$. In Fig. 2, we present the trend of the utility function provided that the number of RBs of an RAU is $k = 20$, $k = 25$, $k = 30$, $k = 40$. We assume that the number of RBs allocated to the users associated with this designated RAU is 10. According to the nature of the logarithmic function, when the load of the entire network is too large, UEs will preferentially select RAUs with fewer RB resources under the premise of satisfying its own QoS, thereby improving the utilization of RBs. UEs' utility value will decrease as the overall network load increases. Also, with the increase of the the number of RBs of an RAU acquired, the value of the utility function goes up. Combined with uplink and downlink model concerned with NAFD scheme, the utility function for downlink user j and uplink user i can be expressed as follows respectively

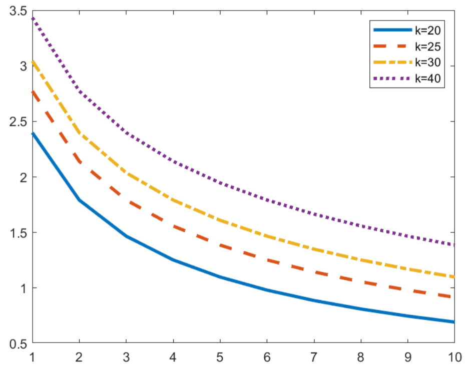

FIGURE 2. Load-aware utility function.

$$U_{\mathrm{D},j} = \sum_{m\in M_{\mathrm{d}}} \ln\left(1 + \frac{\mathbf{k}_{\mathrm{D},m} - \sum_{a\in K_{\mathrm{d}}} n_{a,m}}{n_{m,j}}\right), \tag{9}$$

$$U_{\mathrm{U},i} = \sum_{m\in M_{\mathrm{u}}} \ln\left(1 + \frac{\mathbf{k}_{\mathrm{U},m} - \sum_{a\in K_{\mathrm{u}}} n_{a,m}}{n_{m,i}}\right), \tag{10}$$

Let $\boldsymbol{\beta}_{\mathrm{D},j} = \frac{\mathbf{k}_{\mathrm{D},m} - \sum_{a\in K_{\mathrm{d}}} n_{a,m}}{n_{m,j}}$, then considering the mode selection scheme, $\tilde{\boldsymbol{\beta}}_{\mathrm{D},j} = \boldsymbol{\beta}_{\mathrm{D},j}\mathbf{X}_{\mathrm{d}}$, the effective utility function of downlink is expressed as $U_{\mathrm{D},j} = \ln(1+\tilde{\boldsymbol{\beta}}_{\mathrm{D},j})$. Similarly, the effective utility function of uplink is expressed as $U_{\mathrm{U},i} = \ln(1+\tilde{\boldsymbol{\beta}}_{\mathrm{U},i})$, where $\boldsymbol{\beta}_{\mathrm{U},i} = \frac{\mathbf{k}_{\mathrm{U},m} - \sum_{a\in K_{\mathrm{u}}} n_{a,m}}{n_{m,i}}$, and $\tilde{\boldsymbol{\beta}}_{\mathrm{U},i} = \boldsymbol{\beta}_{\mathrm{U},i}\mathbf{X}_{\mathrm{u}}$.

E. PROBLEM FORMULATION

We aim at maximizing the users' utility function considering the traffic load as follows.

$$\max_{\mathbf{x}_{\mathrm{u}},\mathbf{x}_{\mathrm{d}}} \sum_{i\in K_{\mathrm{u}}} U_{\mathrm{U},i} + \sum_{j\in K_{\mathrm{d}}} U_{\mathrm{D},j}, \tag{11}$$

$$\text{s.t.} \sum_{j\in K_{\mathrm{d}}} ||\mathbf{X}_{\mathrm{d},j}\mathbf{w}_{\mathrm{d},j}||^2 \leq P_{\mathrm{D},j}, \tag{12}$$

$$p_{\mathrm{u},i} \leq P_{\mathrm{U},i}, \tag{13}$$

$$\mathbf{x}_{\mathrm{u}} + \mathbf{x}_{\mathrm{d}} = 1, \tag{14}$$

$$n_{m,i} b \log_2(1+\gamma_i) \geq t_i^{\mathrm{req}}, \tag{15}$$

where $P_{\mathrm{D},j}$ and $P_{\mathrm{U},i}$ are the power consumption budget for downlink RAU j and uplink user i. The binary assignment variables should either be 0 or 1, that is for a certain RAU, it should work at either downlink or uplink working mode. (15) is the QoS needs of different users.

III. PROPOSED LOAD-AWARE DYNAMIC MODE SELECTION ALGORITHM BASED ON REINFORCEMENT LEARNING

Two load-aware dynamic mode selection algorithms based on reinforcement learning for NAFD Cell-Free Large-scale Distributed MIMO Systems is proposed in this part. One is based on centralized Q-learning, the other is based on distributed multi-agent Q-learning.

Reinforcement learning (RL) usually contains five elements, including environment, agent, state, action and reward. The agent has the ability to learn by interacting with the environment constantly and will act on the basis of the observed values combined with its own experience, which is also called a strategy. The state of the environment will be affected by specific action taken by the agent. Two pieces of information from the changing environment will be obtained by the agent: observations and the reward. So the agent can perform new actions based on new observations. The usefulness of performing an action is represented by a numerical value known as the Q-value. The expectation of the long-term return generated by certain strategic actions under the premise of knowing the current state s_t and action a_t is called the state-action value function $Q_\pi(s,a) = \mathbb{E}\left[R_{t+1}+\gamma Q(S_{t+1},A_{t+1})\,|S_t=s,A_t=a\right]$, where γ denotes the discount factor. Then we map our scenario to key elements mentioned above.

A. PROPOSED CENTRALIZED Q-LEARNING ALGORITHM FOR LOAD-AWARE DYNAMIC MODE SELECTION

Q-learning is a classic method of reinforcement learning. The purpose of Q-Learning is to establish a Q-Table with "state" as the row and "action" as the column, and to continuously update the Q value in the Q-Table through the rewards brought by each action, so as to obtain the Q value under a specific action and a specific state. The strategy for taking each action in Q-Learning is ε-greedy strategy, that is, to maintain a delicate balance of exploration and utilization. While the evaluation strategy used when learning to update the Q-Table is the greedy strategy, that is, the best action is always recorded in the Q-Table. Q-Learning is off-policy because its action strategy and evaluation strategy are not the same. In this paper, the proposed centralized Q-learning algorithm treated all RAUs as a whole to assign the working modes of each of them in the system.

1) AGENT

In the proposed centralized Q-learning algorithm, we take the total RAUs as an agent of the proposed reinforcement learning framework. Agent will make intelligent decisions by observations of the environment, including the adaptive selection of working mode of the RAUs. Particularly, the agent can obtain experience and adjust its action strategy.

2) STATE

Define the infinite set for the state space as S, a $1 \times M$ one-dimensional array is used to demonstrate the state of the environment $s_t = [x_1, x_2, x_3 \ldots x_M]$. The value of x_M can be either 0 or 1, where 0 means RAU M is working as uplink reception and 1 means RAU M is working as downlink transmission. At each time slot, an RAU is working as UL reception or DL transmission.

3) ACTION

Since each RAU only has two working mode, we can simply take action to change the original UL RAU to DL RAU or to change the original DL RAU to UL RAU by performing the XOR operation between bits. The agent selects one of the following actions in the current state s_t: "RAU 1 changes its original working mode"… "RAU M changes its original working mode". Hence, M actions is used to model the mode selection scheme. In this way, each RAU's working mode can be changed between uplink and downlink mode according to the strategies taken by the agent.

To obtain the optimal reward, an appropriate adjustment of exploitation and exploration is required because the agent does not possess sufficient information regarding the environment in general. So with the probability of ε(e), the agent selects a random action. While with the probability of $1-\varepsilon$(e), the agent chooses the action with the highest Q-value. If the greedy rate is too high, it is easy to enter the local optimal solution. When we first train the Q function, we must have a large epsilon. As the agent becomes more confident about the estimated Q value, we will gradually decrease epsilon. Hence, the decayed ε-greedy policy is used as follows

$$\varepsilon(e) = \varepsilon_{first}(1-\varepsilon_{first})^{\frac{e}{\varphi\times|\text{action}|}}, \tag{16}$$

where e denotes the the current episode index, ε_{first} is the initial value of ε, φ denotes an exploration parameter that controls the attenuation rate of ε, and |action| is the size of the action set.

4) REWARD

Our goal is to maximize the the users' utility function considering the traffic load, so the reward can be assigned as the sum value of the users' utility function as expressed in (11). The learning process is driven by the reward function in the RL framework, and the system performance can be improved when the design of the reward function for each step is related to the desired objective. The Q-learning algorithm selects the action that can achieve the maximum reward based on the state-action value $Q(s_t, a_t)$. Q-learning uses the current return and the estimated value of the next moment obtained by taking the action that maximizes the value to estimate the value of the current moment. The Q value is updated by the following formula

$$Q(s_t,a_t) = (1-\alpha)\,Q(s_t,a_t) + \alpha\left[R_{t+1}+\gamma Q(s_{t+1},a_{t+1})\right], \tag{17}$$

where α denotes the learning rate. The specific procedures are summarized in Algorithm 1.

An episode is one complete play of the agent interacting with the environment in the general RL setting.

B. PROPOSED DISTRIBUTED MULTI-AGENT Q-LEARNING ALGORITHM FOR LOAD-AWARE DYNAMIC MODE SELECTION

The concept of the distributed algorithm for multi-agent optimization first appeared in [22] where the authors studied the scenario that agents cooperatively minimize a common additive cost with no constraint. Multi-agent reinforcement learning has been used in mobile D2D networks [23] and UAV networks [24]. A distributed Q-learning algorithm for dynamic resource allocation problem with unknown cost functions and unknown resource transition functions was studied in [25]. As a matter of fact, the algorithms based on Q-learning excels a lot in dynamic environments where the users can move randomly in the coverage area and the channel conditions between the RAUs and UEs vary.

In centralized Q-learning, all possible cases explorable by the agent should be considered. Hence, the size of the Q-table of the centralized Q-learning increases exponentially based on the sizes of the state and action sets which can be calculated as $Q_{\text{table-size}} = |\text{state}|_{\text{size}} \times |\text{action}|_{\text{size}}$. In our scenario, the state of the environment is denoted as $s_t = [x_1, x_2, x_3 \dots x_M]$ where the value of x_M can be either 0 or 1. Therefore, the size of our state set comes to 2^M. The size of the Q-table for centralized Q-learning algorithm is $2^M \times M$. By contrast, in distributed multi-agent Q-learning, each agent generates its Q-table considering only its own state set and action set. Therefore, the size of the Q-table of the distributed multi-agent Q-learning can be obtained as ${Q^{dis}}_{\text{table-size}} = N_{\text{agent}} \times |\text{state}|_{\text{own-size}} \times |\text{action}|_{\text{own-size}}$. N_{agent} is the number of agents.

In the proposed distributed multi-agent Q-learning algorithm for load-aware dynamic mode selection, each RAU is regarded as an agent. M RAUs corresponds to M agents. The state set of each RAU is denoted as $s_t = \{s_1, s_2\}$, where s_1 denotes the RAU working as uplink reception, while s_2 denotes the RAU working as downlink transmission. The action taken by each RAU is either "RAU M changes its original working mode from uplink to downlink or from downlink to uplink" or "RAU M stays in its original working mode". The decayed ε-greedy policy in (16) is used in action selection. Subsequently, the size of the Q-table of the proposed distributed multi-agent Q-learning algorithm is $M \times 2 \times 2$. Because distributed multi-agent Q-learning does not require a deep neural network for function approximation, the proposed algorithm can be utilized in practical scenarios for massive terminal access with limited computation power. Furthermore, distributed multi-agent Q-learning method also alleviates the problem of Q-table's size explosion when a large number of terminals access large-scale RAUs. It has high scalability in massive access scenario.

When the state of the RAU indicates its working mode as uplink reception, then the reward of the proposed distributed multi-agent Q-learning framework is defined as

$$R_{\mathrm{m,UL}} = \sum_{i=1}^{K_{\mathrm{u}}} \ln\left(1 + \frac{k - \sum\limits_{a \in K_{\mathrm{u}}} n_{m,a}}{n_{m,i}}\right), \tag{18}$$

When the state of the RAU indicates its working mode as downlink transmision, then the reward is defined as

$$R_{\mathrm{m,DL}} = \sum_{i=1}^{K_{\mathrm{d}}} \ln\left(1 + \frac{k - \sum\limits_{a \in K_{\mathrm{d}}} n_{m,a}}{n_{m,i}}\right), \tag{19}$$

The Q value in the Q-Table is updated according to (17). The specific procedures are summarized in Algorithm 2.

TABLE 1. Simulation parameters.

RAU cell radius/UL and RL user cell radius	600 m/1000 m
Number of RAUs/Number DL users/UL users	10/40/40
Path loss	128.1+37.6log10(d)
Noise power	-90 dBm
UL/DL RAUs and users transmit power	[30 23 dBm]
Optimization parameters [γ, α, φ,ε_{first}]	[0.5,0.5,2,0.99]
RBs to allocate for uplink/downlink RAUs	[40000,80000]

C. COMPLEXITY ANALYSIS

Reinforcement learning provides a robust way to treat environment dynamics and perform sequential decision making by constantly interacting with the uncertainty of the environment, reducing the computational complexity. The time complexity of the proposed centralized Q-learning and the distributed multi-agent Q-learning is $\mathrm{O}(E_{\max}KM)$ and $\mathrm{O}(E_{\max}K)$ respectively. $E_{\max}$ is the number of episodes that make the algorithm converge. K and M here refer to the number of users and RAUs respectively. They are far better than the exhaustion approach which owns the complexity of $\mathrm{O}(2^M)$. The training phase of reinforcement learning is offline so the time it takes for training is out of consideration when implementing it in practice. We can simply get the optimal assignment of RAUs through Q-table that has been trained already. As for the storage unit, centralized Q-learning algorithm requires a $2^M \times M$ Q-table while distributed multi-agent Q-learning only requires $M \times 2 \times 2$ Q-tables. As demonstrated above, the size of Q-table is greatly reduced in distributed multi-agent Q-learning without the implementation of a complicated trained deep neural network which works better under the scenarios for massive terminal access in practice.

IV. NUMERICAL RESULTS AND DISCUSSION

In this section, NAFD cell-free large-scale distributed MIMO system in a circular area is considered. This paper assumes the M RAUs are distributed in a circular area. The system contains K randomly distributed users, including K_{u} uplink users and K_{d} downlink users. Each RAU is equipped with N half-duplex antennas. The detailed simulation parameters are listed in Table 1.

Algorithm 1 Proposed Load-Aware Dynamic Mode Selection Algorithm Based on Centralized Q-Learning

```
Input:
  • initialization: Generate the state set $s_t(\bullet)$ as a $1 \times M$ one-dimensional zero array, create a Q-table scaling $2^M \times M$ and initialize $Q(\bullet)$, $\varepsilon$, $\mathbf{h}_{\mathrm{u},i}$, $\mathbf{h}_{\mathrm{d},j}$, $\mathbf{w}_{\mathrm{d},m}$, $g_{i,j}$, $\mathbf{G}_{\mathrm{I}}$, $\gamma$, $\alpha$, and randomly place $K_{\mathrm{u}}$ uplink users and $K_{\mathrm{d}}$ downlink users.
Repeat:
for Everyepisode do
    $\varepsilon(e) = \varepsilon_{first}(1-\varepsilon_{first})^{\frac{e}{\varphi\times|\text{action}|}}$.
    Determine an action:
    $a_i = \begin{cases} \text{random action, with probability}\varepsilon(\text{e}). \\ \arg\max\limits_{a_i}(Q(s_i,a_i)),\ \text{with probability1-}\varepsilon(\text{e}). \end{cases}$
    The state changes to the next state by taking the action.
    Calculate the reward according to (11), $\max\limits_{\mathbf{x}_{\mathrm{u}},\mathbf{x}_{\mathrm{d}}} \sum\limits_{i\in K_{\mathrm{u}}} U_{\mathrm{U},i} + \sum\limits_{j\in K_{\mathrm{d}}} U_{\mathrm{D},j}$.
    Update the Q-table according to (17).
end for
Return the optimal solutions of state and the optimal reward which correspond to the best assignment of UL/DL RAUs and the biggest value of the utility function respectively.
```

In this study, we considered four conventional schemes as benchmarks to compare the performances of the proposed algorithms in terms of performance metrics.

- **Average RAU scheme:** This scheme is based on randomly equal splitting of the RAUs as half uplink RAUs and half downlink RAUs.
- **TDD scheme:** The TDD scheme is the time division duplex mode.
- **Random scheme:** This scheme randomly chooses an assignment of RAUs in each scenario.
- **Exhaustion scheme:** The exhaustive search provides an optimal solution of the assignment of RAUs with very high computational complexity. The convergence of the proposed reinforcement learning algorithm to the optimal solution can be proved through the comparison between the exhaustive search and the proposed reinforcement learning algorithm.
- **Centralized Q-learning scheme and Distributed multi-agent scheme:** The two schemes refer to the proposed reinforcement learning algorithms mentioned above.

Most of the previous works with respect to FD systems usually fix the assignment of antennas or RAUs and then perform system performance analysis on this basis. The mode selection of RAUs can significantly improve the utilization of UL/DL resources and the load-aware dynamic mode selection scheme improves the access efficiency of massive terminal communications and enhance the transmission reliability of dynamic access links taking the traffic loads in consideration.

Algorithm 2 Proposed Load-Aware Dynamic Mode Selection Algorithm Based on Distributed Multi-Agent Q-Learning

```
Input:
  • initialization: Generate the state set $s_t(\bullet)$ for each RAU, create $M$ Q-tables each scaling $2 \times 2$ and initialize $Q(\bullet)$, $\varepsilon$, $\mathbf{h}_{\mathrm{u},i}$, $\mathbf{h}_{\mathrm{d},j}$, $\mathbf{w}_{\mathrm{d},m}$, $g_{i,j}$, $\mathbf{G}_{\mathrm{I}}$, $\gamma$, $\alpha$, and randomly place $K_{\mathrm{u}}$ uplink users and $K_{\mathrm{d}}$ downlink users.
Repeat:
for Everyepisode do
    $\varepsilon(e) = \varepsilon_{first}(1-\varepsilon_{first})^{\frac{e}{\varphi\times|\text{action}|}}$.
    if State detected as uplink reception then
        for $i = 1 : K_{\mathrm{u}}$ do
            Determine an action for each RAU:
            $a_i = \begin{cases} \text{random action, with probability}\varepsilon(\text{e}). \\ \arg\max\limits_{a_i}(Q(s_i,a_i)),\ \text{with probability1-}\varepsilon(\text{e}). \end{cases}$
            The state changes to the next state by taking the action.
        end for
        Calculate the reward according to (18).
    end if
    if State detected as downlink transmission then
        for $i = 1 : K_{\mathrm{d}}$ do
            Determine an action for each RAU according to the decayed $\varepsilon$-greedy policy.
            The state changes to the next state by taking the action.
        end for
        Calculate the reward according to (19).
    end if
    Calculate the sum reward: $r_{sum} = \sum\limits_{i=1}^{M} R$.
    Update the Q-table according to (17)
end for
Return the optimal solutions of state and the optimal reward which correspond to the best assignment of UL/DL RAUs and the biggest value of the utility function respectively.
```

Fig. 3 and Fig. 4 illustrates the accumulated average reward based on the progress of the episode. The results of optimal rewards and were obtained through an exhaustive search in the entire search space. Fig. 3 was performed under the ZF precoding while Fig. 4 was performed under the MRT precoding. The optimal values is plotted as a constant to present the rate of convergence and optimality of the convergent solutions. As shown in the figure, the proposed Q-learning algorithm can converge quickly and has similar good convergence performance for different precoding which proves the robustness of the algorithm.

Fig. 5 is the CDF of the utility function of different schemes under ZF precoding. Here, we generated 1000 scenarios of randomly distributed users with their required resources requirements. The results indicate that

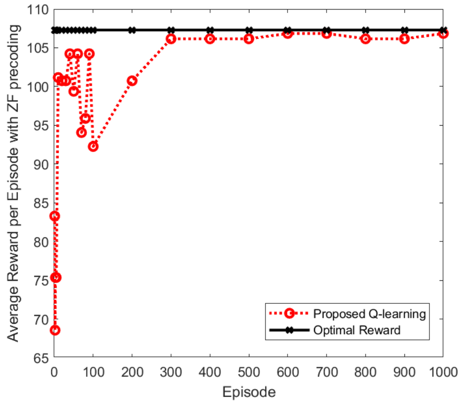

FIGURE 3. Average reward vs. episode under ZF precoding.

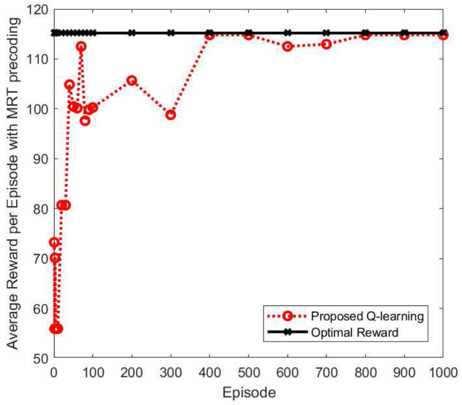

FIGURE 4. Average reward vs. episode under MRT precoding.

the load-aware dynamic mode selection scheme using the proposed centralized Q-learning and distributed multi-agent Q-learning have similar performance. The gain provided by load-aware dynamic mode selection scheme instead of the simple and equally split of RAUs and random assignment scheme is approximately 15% at the probability of 0.5, and it is only 7% poorer than the optimal solution. It also provides nearly 56% gain compared with the TDD scheme.

Fig. 6 is the CDF of the throughput of different schemes. The throughput of the system is calculated under the premise that the values of the above utility function have reached their maximum values of each scheme respectively. The red line which indicates the max throughput is to maximize SE through exhaustive research under the condition of satisfying users' QoS. As shown in Fig. 6, the proposed load-aware mode selection scheme based on Q-learning achieves nearly the same throughput performance as the exhaustive research method, which gains 16% compared with the random scheme, 22% compared with the average RAU scheme and 45% compared with the TDD scheme at the probability of 0.9. This proves the mode selection of RAUs can significantly improve the throughput performance of the system. According to the Nash Equilibrium in Game Theory, a Nash Equilibrium in a game is a list of strategies, one for each player, such that no player can get a better payoff by switching to some other strategy that is available to him while all other players adhere to the strategies specified for them in the list. In our scenario, each RAU's strategy is the best response to other RAU's strategies. Hence, the throughput by maximizing the values of utility function is no larger than the maximizing-throughput in global perspective. Instead, each RAU has made a "no regrets" decision. Maximizing the utility function guides users to prefer RAUs with fewer RB resources on the premise of meeting their own QoS, thus improving the utilization of RBs. With the traffic load and the QoS of users considered, the load-aware dynamic mode selection scheme helps to make better use of system resources, thus improving the efficiency of user access.

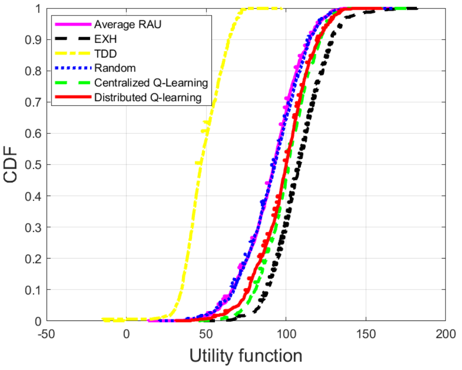

FIGURE 5. CDF of the utility function of different schemes.

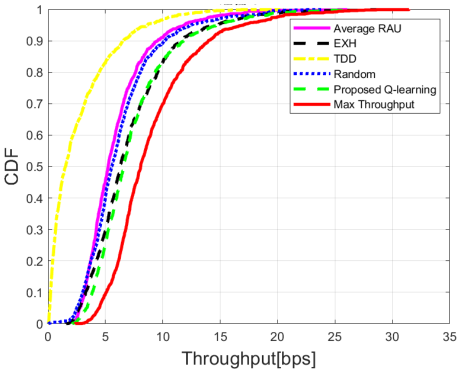

FIGURE 6. CDF of the throughput of different schemes.

Fig. 7 indicates the resource utility between load-aware schemes and non load-aware scheme. Our proposed load-aware dynamic mode selection scheme based on Q-learning gains 13% of the resource utility compared with non load-aware mode selection scheme at the probability of 0.5. It has been proved that the load-aware schemes can significantly improve the resource utility of the system.

Fig. 8 shows the value of the utility function versus the number of RAUs. We assume the number of UL and DL users is 10 respectively. The fixed mode refers to the equally split the RAUs as uplink and downlink working mode. It has been indicated that the proposed load-aware dynamic mode selection scheme based on centralized Q-learning and distributed Q-learning achieves nearly the same performance regardless of the advantages of distributed Q-learning in computing and storage. The exhaustive method can get the theoretically optimal solution with high complexity. As the number of RAUs increases, the values of the utility function go higher under

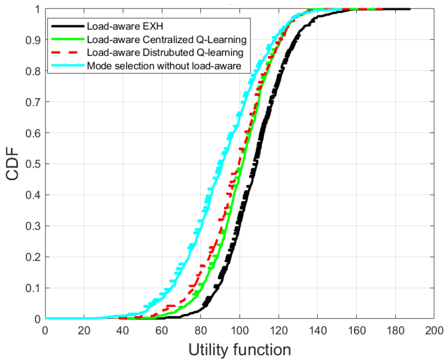

FIGURE 7. CDF of the utility function of load-aware schemes and non load-aware scheme.

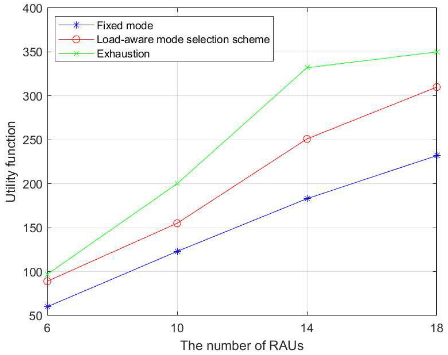

FIGURE 8. The utility function versus the number of RAUs.

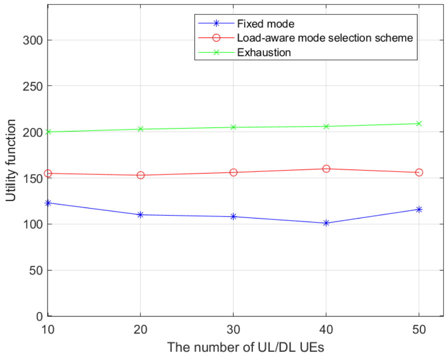

FIGURE 9. The utility function versus the number of UL/DL UEs.

these three schemes. The proposed load-aware dynamic mode selection scheme achieves better performance than the fixed mode scheme and gets closer to the exhaustive method with the growing number of RAUs.

Fig. 9 shows the value of the utility function versus the number of UL/DL UEs. The numbers of uplink and downlink users are equal and correspond to the numbers on the horizontal axis. The values of the utility function don't fluctuate much as the numbers of UL/DL UEs vary which indicates our proposed load-aware dynamic mode selection scheme is robust even when the number of UEs increases.

With the access of a large number of terminals, it is essential to find an flexible mode selection scheme that can maximize the use of system resources. The load-aware dynamic mode selection scheme based on centralized Q-learning determines a clever strategy to approach the ultimate goal by itself and excels in environment dynamics while the distributed multi-agent Q-learning method reduces the computational complexity using a distributed approach. The distributed method has a higher scalability which is more suitable in the actual scene. The proposed Q-learning algorithm can be a reasonable approach for obtaining an optimal solution rapidly and with low complexity.

V. CONCLUSION

In this paper, a load-aware dynamic mode selection scheme of RAUs was studied under NAFD cell-free large-scale distributed MIMO systems. A utility function was proposed to measure the utilization of the system traffic loads. The centralized Q-learning and distributed multi-agent Q-learning algorithms with different complexity were investigated. Through intensive simulations, we demonstrated that the proposed algorithms outperformed conventional schemes, i.e., equally splitting of the RAUs, random scheme, and TDD scheme with far lower complexity compared with exhaustive research method. The load-aware dynamic mode selection can better exploit the system resources and enhance the performance. The distributed multi-agent Q-learning algorithm is proved to more suitable in the actual scene with smaller storage unit and lower complexity. For the sake of real time, CSI overhead and limited fronthaul capacity may be caused by NAFD scheme in the scenario with a large number of RAUs and users. Therefore, in future research, some clustering algorithms will be considered on Aps and UEs to further investigate the scalable mechanism of our systems.

REFERENCES

[1] M. M. Razlighi and N. Zlatanov, "Buffer-aided relaying for the two-hop full-duplex relay channel with self-interference," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 477–491, Jan. 2018.

[2] B. Deballie, D. Broek, C. Lavin, B. V. Liempd, E. Klumperink, C. Palacios, J. Craninckx, and B. Nauta, "Rf self-interference reduction techniques for compact full duplex radios," in *Proc. Veh. Technol. Conf.*, May 2015, pp. 1–6.

[3] J. Zhou, T.-H. Chuang, T. Dinc, and H. Krishnaswamy, "Integrated wideband self-interference cancellation in the RF domain for FDD and full-duplex wireless," *IEEE J. Solid-State Circuits*, vol. 50, no. 12, pp. 3015–3031, Dec. 2015.

[4] T. T. Vu, D. T. Ngo, H. Q. Ngo, and T. Le-Ngoc, "Full-duplex cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[5] S. Datta, D. N. Amudala, E. Sharma, R. Budhiraja, and S. S. Panwar, "Full-duplex cell-free massive MIMO systems: Analysis and decentralized optimization," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 31–50, 2022.

[6] V. H. Nguyen, D. Nguyen, O. A Dobre, S. K. Sharma, S. Chatzinotas, B. Ottersten, and O.-S. Shin, "On the spectral and energy efficiencies of full-duplex cell-free massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1698–1718, Aug. 2020.

[7] Y. Jiang, H. Duan, X. Zhu, Z. Wei, T. Wang, F.-C. Zheng, and S. Sun, "Toward URLLC: A full duplex relay system with self-interference utilization or cancellation," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 74–81, Feb. 2021.

[8] D. Wang, M. Wang, P. Zhu, J. Li, J. Wang, and X. You, "Performance of network-assisted full-duplex for cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1464–1478, Mar. 2020.

[9] D. Wang, Z. Zhao, Y. Huang, H. Wei, X. Wang, and X. You, "Large-scale multi-user distributed antenna system for 5G wireless communications," in *Proc. IEEE 81st Veh. Technol. Conf.*, May 2015, pp. 1–5.

[10] J. Li, Q. Lv, P. Zhu, D. Wang, J. Wang, and X. You, "Network-assisted full-duplex distributed massive MIMO systems with beamforming training based CSI estimation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2190–2204, Apr. 2021.

[11] H. Yang, X. Xia, J. Li, P. Zhu, and X. You, "Joint transceiver design for network-assisted full-duplex systems with SWIPT," *IEEE Syst. J.*, early access, Apr. 7, 2021, doi: 10.1109/JSYST.2021.3062455.

[12] X. Xia, P. Zhu, J. Li, H. Wu, D. Wang, Y. Xin, and X. You, "Joint user selection and transceiver design for cell-free with network-assisted full duplexing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7856–7870, Dec. 2021.

[13] Y. Zhu, J. Li, P. Zhu, H. Wu, D. Wang, and X. You, "Optimization of duplex mode selection for network-assisted full-duplex cell-free massive MIMO systems," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3649–3653, Nov. 2021.

[14] S. Bassoy, M. A. Imran, S. Yang, and R. Tafazolli, "A load-aware clustering model for coordinated transmission in future wireless networks," *IEEE Access*, vol. 7, pp. 92693–92708, 2019.

[15] L. Liu, Y. Zhou, V. Garcia, L. Tian, and J. L. Shi, "Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2741–2755, Mar. 2018.

[16] D. Pandey and P. Pandey, "Approximate Q-learning: An introduction," in *Proc. 2nd Int. Conf. Mach. Learn. Comput.*, 2010, pp. 317–320.

[17] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653–133667, 2019.

[18] L. Shoufeng, L. Ximin, and D. Shiqiang, "Q-learning for adaptive traffic signal control based on delay minimization strategy," in *Proc. IEEE Int. Conf. Netw., Sens. Control*, Apr. 2008, pp. 687–691.

[19] M. D. Awheda and H. M. Schwartz, "Exponential moving average Q-learning algorithm," in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn. (ADPRL)*, Apr. 2013, pp. 31–38.

[20] F. Zhukov, O. Galinina, E. Sopin, S. Andreev, and K. Samouylov, "On load-aware cell association schemes for group user mobility in mmWave networks," in *Proc. 11th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops (ICUMT)*, Oct. 2019, pp. 1–6.

[21] X. Ba and Y. Wang, "Load-aware cell select scheme for multi-connectivity in intra-frequency 5G ultra dense network," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 354–357, 2019.

[22] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[23] W. Jiang, G. Feng, S. Qin, T. S. P. Yum, and G. Cao, "Multi-agent reinforcement learning for efficient content caching in mobile D2D networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1610–1622, Mar. 2019.

[24] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[25] P. Dai, W. Yu, and D. Chen, "Distributed q-learning algorithm for dynamic resource allocation with unknown objective functions and application to microgrid," *IEEE Trans. Cybern.*, pp. 1–11, 2021.

YUE ZHU (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from Shanghai University, Shanghai, China, in 2020. She is currently pursuing the M.S. degree with the College of Information Science and Engineering, Southeast University. Her research interests include flexible full duplex and massive MIMO.

JIAMIN LI (Member, IEEE) received the B.S. and M.S. degrees in communication and information systems from Hohai University, Nanjing, China, in 2006 and 2009, respectively, and the Ph.D. degree in information and communication engineering from Southeast University, Nanjing, in 2014. He joined the National Mobile Communications Research Laboratory, Southeast University, in 2014, where he has been an Associate Professor, since 2019. His research interests include cell-free distributed massive MIMO, massive ultra-reliable low-latency communications (mURLLC), and artificial intelligence and its applications in future mobile communications.

PENGCHENG ZHU (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree in communication and information science from Southeast University, Nanjing, China, in 2009. He has been a Lecturer with the National Mobile Communications Research Laboratory, Southeast University, since 2009. His research interests include communication and signal processing and include limited feedback techniques and distributed antenna systems.

DONGMING WANG (Member, IEEE) received the B.S. degree from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 1999, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2002, and the Ph.D. degree from Southeast University, Nanjing, in 2006. He joined the National Mobile Communications Research Laboratory, Southeast University, in 2006, where he has been an Associate Professor, since 2010. His research interests include turbo detection, channel estimation, distributed antenna systems, and large-scale MIMO systems.

HENG YE (Member, IEEE) received the bachelor's degree in automation engineering from Tsinghua University and the master's degree in telecommunications and information system from the China Academy of Telecommunication and Technology. He is currently a System Software Architect and a Technical Lead of the Wireless Access Network Division, Network Platform Group, Intel Corporation. He has over 20 years of research and development experience in the wireless industry. He has contributed extensively to 3G/4G/5G radio interface and system architecture design and development.

XIAOHU YOU (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Nanjing Institute of Technology, Nanjing, China, in 1982, 1985, and 1989, respectively. From 1987 to 1989, he was a Lecturer with the Nanjing Institute of Technology. Since 1990, he has been with Southeast University, first as an Associate Professor and later as a Professor. He is the Chief of the Technical Group of China's 3G/B3G Mobile Communication Research and Development Project. His research interests include mobile communications, adaptive signal processing, and artificial neural networks with applications to communications and biomedical engineering. He received the Excellent Paper Prize from the China Institute of Communications in 1987 and the Elite Outstanding Young Teacher Awards from Southeast University in 1990, 1991, and 1993. He was also a recipient of the 1989 Young Teacher Award of the Fok Ying Tung Education Foundation, State Education Commission of China.