User Access Control in Open Radio Access Networks: A Federated Deep Reinforcement Learning Approach

Yang Cao, *Graduate Student Member, IEEE*, Shao-Yu Lien, *Member, IEEE*, Ying-Chang Liang, *Fellow, IEEE*, Kwang-Cheng Chen, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*— Targeting at implementing the next generation radio access networks (RANs) with virtualized network components, the open RAN (O-RAN) has been regarded as a novel paradigm towards fully open, virtualized and interoperable RANs. Through particularly introducing RAN intelligent controllers (RICs), machine learning (ML) can be unprecedentedly installed, adapting to various vertical applications and deployment environments without sophisticated planning efforts. However, the O-RAN also suffers two critical challenges of load balancing and frequent handovers in the massive base station (BS) deployment. In this paper, an intelligent user access control scheme with deep reinforcement learning (DRL) is proposed. To optimize the performance of distributed deep Q-networks (DQNs) trained by user equipments (UEs), a federated DRL-based scheme is proposed with a global model server installed in the RIC to update the DQN parameters. To further predictively train a global DQN with acceptable signaling overheads, the upper confidence bound (UCB) algorithm to select the optimal UE set and a dueling structure to decompose the DQN parameters are developed. With the proposed scheme, each UE effectively maximizes the long-term throughput and avoids frequent handovers. The simulation results well justify the outstanding performance of the proposed scheme over the-state-of-the-arts, to serve as references for the O-RAN standardization.**

***Index Terms*— Open radio access networks (O-RANs), user access control, deep reinforcement learning (DRL), federated learning (FL), RAN intelligent controller (RIC), deep Q-networks (DQNs).**

Manuscript received December 30, 2020; revised May 28, 2021 and September 23, 2021; accepted October 12, 2021. Date of publication November 3, 2021; date of current version June 10, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1801105, in part by the National Natural Science Foundation of China under Grant 61631005 and Grant U1801261, in part by the Key Areas of Research and Development Program of Guangdong Province, China, under Grant 2018B010114001, in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2019Z022, in part by the Program of Introducing Talents of Discipline to Universities under Grant B20064, and in part by the Ministry of Science and Technology (MOST) under Contract 108-2221-E-194-020-MY3 and Contract 110-2224-E-305 -001-. An earlier version of this paper was presented in part at the IEEE International Conference on Communications (ICC) 2021 [DOI: 10.1109/ICC42927.2021.9500603]. The associate editor coordinating the review of this article and approving it for publication was L. Duan. *(Corresponding author: Ying-Chang Liang.)*

Yang Cao is with the National Key Laboratory on Communications and the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou 313001, China (e-mail: cyang9502@gmail.com).

Shao-Yu Lien is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 62102, Taiwan (e-mail: sylien@ccu.edu.tw).

Ying-Chang Liang is with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou 313001, China (e-mail: liangyc@ieee.org).

Kwang-Cheng Chen is with the Department of Electrical Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: kwangcheng@usf.edu).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3123500.

Digital Object Identifier 10.1109/TWC.2021.3123500

I. Introduction

UNLIKE the International Mobile Telecommunications-advanced (IMT-advanced) systems solely supporting a single class of wireless services, such as undiscriminating latency, reliability, and data rate, the key feature of the IMT-2020 systems lies in enabling support multi-class wireless services with different requirements in terms of latency, reliability, connection density, and data rates. This emerging feature however greatly increases the design and deployment complexity of the radio access network (RAN), and drives the global industry to reach the consensus on two core principles to evolve the current RAN to the next generation RAN [1]: 1) **openness** to attract more participants and collaborations of venders/researchers in the designs and implementations of RANs; and 2) **intelligence** to enable machine learning (ML) based optimization in RANs for autonomous deployment. To this end, an innovative solution, known as the open RAN (O-RAN), has emerged since 2018 to create a new paradigm of wireless infrastructure.

To achieve openness, the functions of the RAN in the O-RAN are virtualized and divided into four key units [2], [3]. (i) The Radio Unit (RU) embraces the radio frequency (RF) components and lower-physical layer. (ii) The Distributed Unit (DU) sustains the higher-physical layer, medium access control (MAC) layer and radio link control (RLC) layer. (iii) The Central Unit (CU) hosts the radio resource control (RRC) protocol, packet data convergence protocol (PDCP)

1536-1276 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

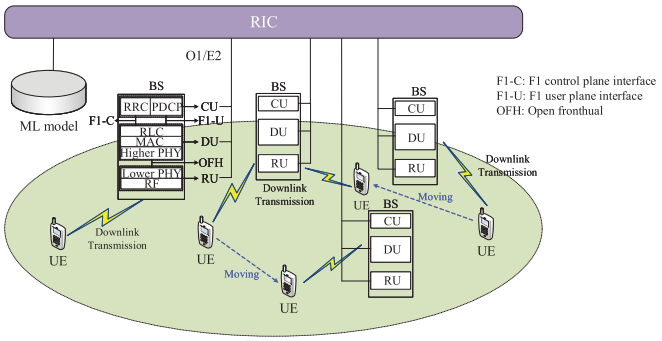


Fig. 1. The general deployment of the O-RAN considered in this paper, in which there are N BSs and M UEs. The global model server is installed at the RIC to train the global model parameters for UEs.

and service data adaptation protocol (SDAP). (iv) Most importantly, the RAN Intelligent Controller (RIC) is unprecedentedly introduced. To achieve intelligence, various ML algorithms can be installed to the RIC to perform predictive resource/communication configuration for CUs/DUs, so as to optimize the performance for any vertical applications over any deployment scenarios. With the aid of this architecture, a base station (BS) can be constructed by connecting an RU, DU and CU through standardized interference [3], and the RIC is responsible for the optimal resource/communication configuration for one or multiple BSs, as illustrated in Fig. 1. In this case, each technology contributor therefore can only focus on the design/implementation of a single function of these four units.

A. Challenges and Related Works

Although the O-RAN provides a revolutionary architecture for an intelligent and agile deployment of the IMT-2020 infrastructure, this innovative technology suffers from the engineering challenges in the multi-BS deployment. Based on the architecture of the O-RAN, the functions of BSs are decomposed/virtualized to CUs/DUs/RUs, which inherently leads to a massive deployment of CUs/DUs/RUs. In such a massive deployment, it is expected to assign each user equipment (UE) to access a proper BS so as to maximize the overall throughput (i.e., the sum of all individual throughputs of UEs), known as user access control or user association [4]. In conventional user access control schemes, the user access decisions are made based on certain metrics such as *received signal strength* (RSS) or capacity [5] measured by UEs, by which UEs always access the BS with the strongest RSS. As a result, an RSS-based scheme may incur two critical concerns: frequent handovers and load balancing.

Aiming at optimizing the throughput performance, in existing handover schemes, the handover procedure is triggered when certain conditions (i.e., the RSS difference with a particular hysteresis margin is larger than the pre-determined threshold) occur. To reduce the number of unnecessary handovers, the handover parameters (i.e. the hysteresis margin, comparison threshold and time-to-trigger) should be designed based on the average dynamics instead of the instantaneous gains. For this goal, a fuzzy logic based algorithm has been proposed to optimize the handover parameters in [6]. Additionally, a ping-pong timer has been introduced to distinguish the types of handovers incurred by movements in [7], and a handover skipping technique has been proposed to enable high-speed UEs to avoid inefficient-in-throughput handovers along their trajectories in [8]. On the other hand, to balance the loads among BSs with different coverage sizes, a range expansion mechanism has been developed, in which BSs can adjust their range expansion biases based on RSSs and throughputs of UEs for attracting UEs to access [9], and a load-aware user association scheme has been proposed in [10] to maximize the downlink throughput and balance the loads by exploiting traffic load information and link qualities of different BSs. However, in all these schemes, handover parameters are optimized based on the fixed tracking areas of BSs. While in the O-RAN, since the functions of BSs are decomposed/virtualized to CUs/DUs/RUs, the massive deployment of CUs/DUs/RUs inherently leads to a "cell-less" architecture [11]. In such a massive deployment, performing the optimization of user access control solely at the RAN side may be practically intractable due to unaffordable signaling overheads and complexity. In this case, each UE autonomously selecting proper BSs (or CUs/DUs/RUs) to access has manifested the effectiveness in significantly reducing the control complexity and signaling overheads. As a result, to successfully deploy the O-RAN, an effective user access control performed at the UE side to maximize the overall throughput and avoid frequent handovers is urgently desired.

In [12], the user access control has been formulated as a convex optimization with proper constraint relation and decomposition. In [13] and [14], game-theoretical approaches such as matching game have been applied to user association and resource allocation in RANs. However, these approaches rely on the prior knowledge and beliefs of each UE's decisions, channel conditions between each UE and BSs, and mobility of each UE, which may not be generally available for all the deployment scenarios. To obtain the optimal UE association policy without the above prior knowledge and beliefs, reinforcement learning (RL) has been shown as a promising remedy to address the frequent handover problem and maximize the long-term performance of UEs [15], [16]. Nevertheless, such tabular RL algorithms require a large memory space to store the experiences of the interactions with the environment, and an outstanding performance relies on sufficient updates of these experiences. Unfortunately, in the practical deployment, it is very likely that only limited experiences are available.

Recently, deep reinforcement learning (DRL) has been shown as an effective method to fundamentally enhance the performance of RL [17]–[19]. By performing DRL in each UE, the variation pattern hidden in the environment and the optimal access policies can be learned by UEs only using local observations [20], [21]. In the meantime, unlike in the RSS-based schemes, each UE in DRL targets at maximizing the long-term throughput instead of the instantaneous throughput. Unnecessary handovers therefore can be avoided. However, if DRL in each UE operates independently without any information exchange, it may take an unacceptably long time to achieve convergence. To guarantee the performance

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

convergence, multiple DRL agents need to exchange local observations with each other and take actions made by other agents into considerations [22], [23] to infer global network information [24]. However, such observation-sharing mechanisms may lead to an unaffordable amount of communication overheads. Therefore, to implement effectively distributed DRL algorithms, UEs are only allowed to exchange a limited amount of parameters with other UEs.

Despite that only a limited amount of parameter exchanges are allowed among UEs, if these parameters can be adequately trained, the performance can be significantly enhanced. To this end, the federated learning (FL) method [25] can be adopted to train a global DRL model using the exchanged parameters from UEs [26], [27]. The spirit of FL is to permit each UE to train its own (local) model with local observations [28]. Subsequently, instead of directly exchanging model parameters among UEs, each UE can send its trained model parameters to a global model server, in which a global model is updated using the model parameters provided by UEs [29]. After training the global model, the parameters of the global model are disseminated to UEs to further improve the local models. With the aid of the O-RAN architecture, the RIC connecting to multiple CUs/DUs/RUs is therefore perfectly suitable to sustain and train the global model and perform FL computation for UEs.

B. Main Contributions

Although the O-RAN can potentially sustain different ML algorithms to tackle complex optimization models for manifold scenarios, the very first ML scheme should effectively address user access control, so as to implement the O-RAN. Even though the above technical merits render FL a recent innovation for user access control, significant efforts and particular designs are still required. With the facilitation of the RIC to accommodate the global model server for FL, in this paper, we consequently propose a federated DRL-based scheme to address user access control in the O-RAN. To the best of our knowledge, our work is the first design to adopt FL and DRL to establish intelligent user-centric access control mechanism to optimize the overall throughput and avoid frequent handovers in the O-RAN, in which UEs can access proper BSs and RBs autonomously based on their local information. In such case, the user access control problem can be formulated as a long-term utility optimization, and the following contributions are provided to efficiently solve the optimization in the O-RAN.

- In the developed federated DRL-based scheme for user access control, each UE trains two deep Q-networks (DQNs) using its own observations, and makes access decisions (i.e., selecting proper BS and RB to access) based on its DQNs. The trained DQN parameters are subsequently forwarded to a global model server installed in the RIC to train a global DQN, and the parameters of the global DQN are then disseminated to each UE to further improve its access decision.
- To substantially reduce the signaling overheads between UEs and the global model server, the global model server may select a part of UEs and only these selected UEs should send their DQN parameters to the global model server. Since the global model server may not be fully aware of the link conditions in all the UE-BS pairs, there is a tradeoff between exploration and exploitation in the UE selection. To tackle this tradeoff, an upper confidence bound (UCB)-based UE selection scheme is developed to choose the optimal UE set.
- To optimally estimate/learn statistics of the environment, the global model server should obtain independent samples from the environment (i.e., from the parameters of selected UEs). However, even though a particular set of UEs are selected to provide the DQN parameters, it is still possible that different UEs may have a common knowledge of the interacting environment. In this circumstance, if each selected UE provides a complete set of the DQN parameters, these parameters may not be independent and the convergence of learning results may not be warranted in general. With such a disfavored situation, in the proposed scheme, a dueling structure is introduced to decompose the DQN parameters of each UE into three different parts, and only two parts of the parameters are sent to the global model server for the global DQN training.
- Through comprehensively evaluating the performance of the proposed scheme, simulation results show that the proposed scheme outperforms the distributed training scheme, centralized training, RSS-based scheme, heuristic algorithm and conventional reinforcement learning (RL) methods in terms of the long-term throughput and the number of handovers. The proposed scheme is also able to achieve any levels of tradeoff between the throughput and number of handovers through properly arranging the handover punishment factor.

C. Organizations

The rest of this paper is organized as follows. After describing the system model and problem formulation in Section II. Subsequently, the proposed federated DRL-based user access control in the O-RAN and corresponding motivations are presented in Section III. In Section IV, sufficient simulation studies are provided, and this paper is concluded in Section V.

II. System Model and Problem Formulation

A. System Model

In this paper, the O-RAN deployment as shown in Fig. 1 is considered, in which N BSs (indexed by $j = 0, 1, \cdots, N-1$) connecting to an RIC through E2 or O1 interface are deployed to serve M UEs (indexed by $i = 0, \cdots, M-1$), where $M > N$. Let $\mathcal{B}$ and $\mathcal{U}$ denote the sets of BSs and UEs, respectively. Since 3GPP New Radio (NR) [30] adopting *orthogonal frequency division multiple access* (OFDMA) is the only system supported by the O-RAN, in this paper, OFDMA is also considered and all the BSs share a common resource pool $\mathcal{K}$ composed of K RBs (indexed by $k = 0, \cdots, K-1$). In the time domain, aligning with the frame structure of 3GPP NR, the resources are arranged in equal-duration time slots.

Suppose that each UE is permitted to select only one BS from $\mathcal{B}$ to access and is allocated with a fixed number of

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

TABLE I
MAJOR NOTATIONS ADOPTED IN THIS PAPER

Notations	Definitions	Notations	Definitions	Notations	Definitions
N, $\mathcal{B}$	Number and set of BSs	M, $\mathcal{U}$	Number and set of UEs	K, $\mathcal{K}$	Number and set of channels
$u_{i,j}(t)$	UE access indicator	$f_{i,k}(t)$	RB allocation indicator	$g_{i,j}^k(t)$	Channel gain between UE i and BS j in RB k at time slot t
$h_{i,j}^k(t)$	Small-scale fading between UE i and BS j in RB k at time slot t	$l_{i,j}(t)$	Large-scale fading between UE i and BS j at time slot t	ρ	Coherent coefficient of small-scale fading
P_j	Transmit power of BS j	B	Bandwidth of each RB	$\gamma_{i,j}^k(t)$	SINR of UE i from BS j in RB k at time slot t
$c_{i,j}^k(t)$	Achievable rate of UE i from BS j in RB k at time slot t	$R_{i,j}(t)$	Receiving rate of UE i from BS j at time slot t	C	Equivalent handover cost
β	Discounting factor in DRL	$\boldsymbol{s}_i(t), \boldsymbol{a}_i(t), r_i(\boldsymbol{s}_i(t), \boldsymbol{a}_i(t))$	State, action and reward of UE i at time slot t	$\omega_i(t)$	Receiving rate of UE i at time sot t
η	Punishment factor in reward function	$\boldsymbol{\theta}_i^c(t), \boldsymbol{\theta}_i^v(t), \boldsymbol{\theta}_i^a(t)$	Common-network, value-function and advantage-function parameters of UE i	$\boldsymbol{\theta}_G^c(t), \boldsymbol{\theta}_G^v(t)$	Global common-network and value-function parameters

RBs for downlink transmissions at each time slot (since this fixed number of allocated RBs does not impact the operation of the proposed scheme, we can simply take this number as one as an example). To indicate whether UE i accesses BS j at time slot t, a binary indicator $u_{i,j}(t)$ is adopted, where $u_{i,j}(t) = 1$ if UE i accesses BS j, and $u_{i,j}(t) = 0$ otherwise. A RB allocation indicator $f_{i,k}(t)$ is further adopted to indicate whether UE i is allocated with RB k at time slot t, where $f_{i,k}(t) = 1$ if RB k is allocated to UE i, and $f_{i,k}(t) = 0$ otherwise. In Table I, major notations adopted in this paper are summarized.

1) Channel Model: The channel gain $g_{i,j}^k(t)$ between UE i and BS j in RB k at time slot t is composed of two components, i.e., the large-scale fading component $l_{i,j}(t)$ and the small-scale fading component $h_{i,j}^k(t)$. The large-scale fading component $l_{i,j}(t)$ is determined by the distance $d_{i,j}(t) = \sqrt{(x_i(t) - x_j(t))^2 + (y_i(t) - y_j(t))^2}$ between UE i located at $(x_i(t), y_i(t))$ and BS j located at $(x_j(t), y_j(t))$. With the facilitation of the mature channel equalization methods, the small-scale fading component $h_{i,j}^k(t)$ can be regarded as unchanged within a time slot. To capture the variation of small-scale fading between two contiguous time slots, the Jake's model [31] can be applied, i.e.,

$$h_{i,j}^k(t+1) = \rho h_{i,j}^k(t) + \delta_{i,j}^k(t), \tag{1}$$

where ρ is the coherent coefficient between two contiguous time slots. $h_{i,j}^k(0)$ is a Gaussian random variable with zero mean and unit variance $h_{i,j}^k(0) \sim \mathcal{CN}(0,1)$. $\delta_{i,j}^k(t)$ also can be regarded as a Gaussian random variable with zero mean and a variance of $1-\rho^2$, $\delta_{i,j}^k(t) \sim \mathcal{CN}(0, 1-\rho^2)$. Hence, the channel gain $g_{i,j}^k(t)$ from BS j to UE i in RB k at time slot t can be expressed by

$$g_{i,j}^k(t) = l_{i,j}(t) \left| h_{i,j}^k(t) \right|^2. \tag{2}$$

2) Downlink Transmissions: If multiple UEs are allocated with the same RB, the co-channel interference induced by other BSs should be considered. In this case, the signal-to-noise-plus-interference-ratio (SINR) at UE i in RB k from BS j at time slot t is given by

$$\gamma_{i,j}^k(t) = \frac{P_j f_{i,k}(t) g_{i,j}^k(t)}{\sum_{m \in \mathcal{B}_k \setminus j} P_m f_{i,k}(t) g_{i,m}^k(t) + \sigma^2}, \tag{3}$$

where $\mathcal{B}_k$ is the set of BSs allocating RB k to other UEs, P_j and P_m are the transmission power levels of BS j and m, respectively, and σ^2 denotes noise power. Therefore, to indicate downlink transmission rate in each RB, the Shannon capacity equation is adopted, and the expression of the achievable rate of UE i from BS j in RB k at time slot t can be given by

$$c_{i,j}^k(t) = B \log_2(1 + \gamma_{i,j}^k(t)), \tag{4}$$

where B is the bandwidth of each RB. Since all the RBs have the same slot length, based on (4), the receiving rate of UE i from BS j at time slot t is the sum of the achievable rates in all the allocated RBs, which is given by

$$R_{i,j}(t) = \sum_{k \in \mathcal{K}} f_{i,k}(t) c_{i,j}^k(t). \tag{5}$$

B. Problem Formulation

The objective of BS selection in each UE is to maximize the downlink throughput. Due to mobility of UEs, the distance between a UE and a particular BS may change rapidly, leading

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

to a highly varying channel gain between a UE and a BS. As a result, the throughput from the BS (mainly determined by the channel gain) may also change drastically. In this case, handover may take place frequently if the RSS-based BS access scheme is adopted, which is harmful in the practical O-RAN deployment. On the other hand, if multiple UEs access the same BS, this BS could be heavily loaded, which leads to a poor throughput performance. To jointly optimize the throughput and avoid unnecessary handovers, an optimization maximizing a utility function taking both throughput and handover into consideration should be designed.

Definition 1: The utility function of a UE jointly considering the handover cost and throughput is defined by

$$\Gamma_{i,j}(t) \triangleq \begin{cases} R_{i,j}(t), & u_{i,j}(t) = u_{i,j}(t-1), \\ R_{i,j}(t) - C, & u_{i,j}(t) \neq u_{i,j}(t-1), \end{cases} \tag{6}$$

where C is the equivalent cost of the handover, which is incurred by the signalling overheads and power consumptions. Particularly, C can be regarded as a system parameter indicating preference of the operator between the overall throughput and the number of handovers in the practical deployment. A large C implies that less handovers and a lower throughput are allowed, while a small C implies that more handovers and a higher throughput are preferred.

To avoid frequent handovers, an effective scheme may lie in adopting the long-term performance as the objective of the optimization, instead of optimizing the immediate/instantaneous performance. For this purpose, in this paper, we target at maximizing the long-term utilities of all the UEs in the O-RAN, and such optimization can be formulated in the following.

Optimization 1: The optimization of the long-term utilities of all the UEs from time slot $t = 0$ to time slot $t = T - 1$ is given by

$$\max_{\boldsymbol{U},\boldsymbol{F}} \sum_{t=0}^{T-1} \sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{B}} u_{i,j}(t)\Gamma_{i,j}(t)$$

$$\text{s.t.} \sum_{j\in\mathcal{B}} u_{i,j}(t) = 1, \quad \forall i \in \mathcal{U}, \tag{7}$$

$$\sum_{k\in\mathcal{K}} f_{i,k}(t) = 1, \quad \forall i \in \mathcal{U}, \tag{8}$$

$$u_{i,j}(t) \in \{0,1\}, \quad \forall (i,j) \in \mathcal{U}\times\mathcal{B}, \tag{9}$$

$$f_{i,k}(t) \in \{0,1\}, \quad \forall (i,k) \in \mathcal{U}\times\mathcal{K}, \tag{10}$$

where $\boldsymbol{U} = [u_{i,j}(t), \forall i,j,t]$ and $\boldsymbol{F} = [f_{i,k}(t), \forall i,k,t]$ are vectors of $u_{i,j}(t)$ and $f_{i,k}(t)$, respectively. The constraints (7) *and* (8) *indicate that each UE is permitted to access one BS only and is allocated with one RB for downlink transmissions at each time slot.*

Optimization is an integer programming problem, which has been regarded as an intractable problem using the conventional methods such as the convex optimization. In addition to the mathematical programming, solving this optimization in the practical deployment also creates critical issues to be reckoned with. To perform this optimization with a single/centralized agent, global network information such as positions and channel state information of UEs should be available to this agent, which induces unaffordable signaling overheads. On the contrary, toward performing this optimization in a distributed manner, the objective of each UE in the proposed scheme is to maximize its individual long-term utility, which also can be regarded as the general-sum game. Since the numbers of UEs, BSs and RBs are finite, there exists the Nash equilibria for UEs, which can only be derived if the completed network information is common knowledge for all the UEs.

However, due to the resource limitations and co-channel interferences, the long-term utility performance of each UE may be impacted by other UEs' access decisions, and therefore convergence to the optimum performance is not guaranteed. On the other hand, since each UE cannot obtain the accurate network information, each UE needs to infer beliefs about other UEs' access strategies to perform decision-making tasks. To this end, there is an urgent need to develop an effective mechanism at the UE side for inferring the mutual decision impacts among UEs. These engineering concerns thus motivate the design of an intelligent access control scheme based on DRL, which plays a crucial role in the family of ML methods to tackle the sequential-decision-making tasks. In DRL, the objective of an agent is to find the optimal policy π^* to maximize the long-term accumulated rewards from continuous interactions with the environment, i.e.,

$$Q(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}\left[\sum_{t=0}^{T-1} \beta^t r(\boldsymbol{s}(t),\boldsymbol{a}(t)) | \boldsymbol{s}(0) = \boldsymbol{s}, \boldsymbol{a}(0) = \boldsymbol{a}\right], \tag{11}$$

where $\mathbb{E}[\cdot]$ is the expectation operation, $\boldsymbol{s}$ and $\boldsymbol{a}$ are the initial state and action, respectively, and $r(\boldsymbol{s}(t),\boldsymbol{a}(t))$ is the instantaneous reward obtained by the agent visiting state $\boldsymbol{s}(t)$ and selecting action $\boldsymbol{a}(t)$ at time slot t, and $0 \leq \beta \leq 1$ is the discounting factor. Therefore, DRL can be adopted to enable each UE to obtain the optimal access policy for maximizing its long-term utility by continuously interacting with the environment. To further enhance the performance of DRL with acceptably low signaling overheads, a federated DRL-based scheme particularly feasible for the O-RAN is consequently proposed in this paper.

III. Federated DRL-Based Scheme for User Access Control in O-RAN

In the proposed federated DRL-based scheme, **Optimization 1** can be transformed to a *Markov Decision Process* (MDP) $\mathcal{M} = \{\mathcal{S},\mathcal{A},\mathcal{R},\mathcal{P}\}$, which is composed of the state space $\mathcal{S}$, action space $\mathcal{A}$, reward function $\mathcal{R}$, and transition probability space $\mathcal{P}$, and the optimum access decision can be derived using the dynamic programming (DP) methods if the state transition probabilities of the MDP are available. However, in practical O-RAN operations, the state transition probabilities of the MDP are generally unavailable, and this knowledge should be predictively estimated through interacting with the environment. In our proposed scheme, DRL is introduced to enable each UE to learn the hidden variation patterns, and their DQN parameters can be regarded as the average knowledge of the environment. Additionally, since the objective is to optimize the overall performance,

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

UEs should make access decisions cooperatively according to the estimated global knowledge in addition to their local observations, and the DQN parameters of the UE set with higher accumulated rewards should be selected to update the global knowledge of the environment to optimize the overall performance. With the aid of FL architecture, the global DQN parameters can be acquired by aggregating DQN parameters from multiple UEs selected by global model server installed in the RIC. For this purpose, the proposed scheme is composed of three stages. In the first stage, each UE trains two local DQNs based on its local observations. In the second stage, the global model server installed in the RIC selects certain UEs to acquire their parameters of DQNs, and trains a global DQN using the obtained parameters from the selected UEs. In the third stage, the parameters of the global DQN are disseminated to all the UEs to further refine their local DQNs. Each UE then makes access decisions based on the local DQNs.

A. Proposed DRL for Each UE

Since the MDP model may significantly impact the performance of the corresponding DRL, in the section, the state space, action space and reward function for the MDP should be designed.

Definition 2 (State Space): The state of UE i at time slot t denoted as $\boldsymbol{s}_i(t)$ is composed of five components. The first component is a vector of the BS access indicators at time slot $t-1$. The second component is a vector of the RB allocation indicators at time slot $t-1$. The third component is a vector of the RSSs in different RBs from BS $\hat{j} = \arg\max_{j\in\mathcal{B}} \boldsymbol{U}_i(t-1)$ at time slot $t-1$, where $\boldsymbol{U}_i(t-1) = \{u_{i,j}(t-1)\}, j \in \mathcal{B}$. The fourth component is a vector of the RSSs from different BSs in RB $\hat{k} = \arg\max_{k\in\mathcal{K}} \boldsymbol{F}_i(t-1)$ at time slot $t-1$, where $\boldsymbol{F}_i(t-1) = \{f_{i,k}(t-1)\}, k \in \mathcal{K}$. The final component $\omega_i(t-1) = \sum_{j\in\mathcal{B}} R_{i,j}(t-1)$ is the receiving rate of UE i at time slot $t-1$. $\boldsymbol{s}_i(t)$ is therefore given by

$$\begin{aligned}\boldsymbol{s}_i(t) \triangleq \{&u_{i,0}(t-1),\cdots,u_{i,N-1}(t-1);\\ &f_{i,0}(t-1),\ldots,f_{i,K-1}(t-1);\\ &P_{\hat{j}}g^0_{i,\hat{j}}(t-1),\ldots,P_{\hat{j}}g^{K-1}_{i,\hat{j}}(t-1);\\ &P_0 g^{\hat{k}}_{i,0}(t-1),\ldots,P_{N-1}g^{\hat{k}}_{i,N-1}(t-1);\\ &\omega_i(t-1)\}.\end{aligned} \tag{12}$$

Definition 3 (Action Space): The action of UE i at time slot t denoted as $\boldsymbol{a}_i(t)$ is a BS-and-RB selection, and UE i sends a request to the selected BS to acquire an RB for downlink transmissions, i.e.,

$$\boldsymbol{a}_i(t) \triangleq \{u_{i,0}(t),\cdots u_{i,N-1}(t), f_{i,0}(t),\cdots,f_{i,K-1}(t)\}. \tag{13}$$

Definition 4 (Reward Function): Since the objective of each UE is to maximize the throughput and avoid frequent handovers, the reward function of UE i after taking action $\boldsymbol{a}_i(t)$ at state $\boldsymbol{s}_i(t)$ integrates the receiving rate and the handover costs, i.e.,

$$r_i(\boldsymbol{s}_i(t),\boldsymbol{a}_i(t)) \triangleq \begin{cases}\omega_i(t), & \boldsymbol{U}_i(t)=\boldsymbol{U}_i(t-1),\\ \omega_i(t)-\eta C, & \boldsymbol{U}_i(t)\neq\boldsymbol{U}_i(t-1),\end{cases} \tag{14}$$

where η is a punishment factor to balance the throughput and the handover cost.

In RL, to balance the exploration and exploitation, the ϵ-greedy policy has been shown as an optimum policy for action selection, in which non-greedy actions being selected stochastically for explorations. However, in the case that a greedy action is not selected, all the non-greedy actions are selected with an equal probability may limit the overall performance. To further enhance the performance, explorations (non-greedy action selection) can be performed based on the accumulated scores of BSs and RBs, which are updated by the corresponding historical RSS information stored in states. For this purpose, in the following an exploration mechanism known as the score-based exploration is proposed.

Proposition 1 (Score-Based Exploration): At each time slot, UE i selects the action based on the following policy, i.e.,

$$\boldsymbol{a}_i(t) = \begin{cases}\arg\max\limits_{\boldsymbol{a}_i(t)\in\mathcal{A}} Q(\boldsymbol{s}_i(t),\boldsymbol{a}_i(t)), & \text{with } 1-\epsilon,\\ \arg\max\limits_{\boldsymbol{a}_i(t)\in\mathcal{A}} \upsilon_i(\boldsymbol{a}_i(t)), & \text{with } \epsilon\end{cases} \tag{15}$$

where

$$\begin{aligned}\upsilon_i(\boldsymbol{a_i}(t)) &= \upsilon_i(\boldsymbol{U}_i(t),\boldsymbol{F}_i(t))\\ &= \frac{\sum\limits_{\tau=0}^{t-1}\mathcal{F}^i_{bs}(\boldsymbol{U}_i(t),\tau)+\sum\limits_{\tau=0}^{t-1}\mathcal{F}^i_{rb}(\boldsymbol{F}_i(t),\tau)}{2(t+1)}\end{aligned} \tag{16}$$

where $\mathcal{F}^i_{bs}(\boldsymbol{U}_i(t),\tau)$ and $\mathcal{F}^i_{rb}(\boldsymbol{F}_i(t),\tau)$ are the scores of the BS selection $\boldsymbol{U}_i(t)$ and RB selection $\boldsymbol{F}_i(t)$ of UE i at time slot τ, respectively. Specifically, the RSSs from different BSs (RBs) in state $\boldsymbol{s}_i(\tau)$ are sorted in a descending order, and the order of BS selection $\boldsymbol{U}_i(t)$ (RB selection $\boldsymbol{F}_i(t)$) is set as its corresponding score $\mathcal{F}^i_{bs}(\boldsymbol{U}_i(t),\tau)$ ($\mathcal{F}^i_{rb}(\boldsymbol{F}_i(t),\tau)$) at time slot t.

In the proposed DRL, the input of the DQN is the state defined in (12), and the outputs are the action values corresponding to actions. Moreover, each UE needs to construct two DQNs, i.e., the trained DQN $Q_i(\boldsymbol{s}_i,\boldsymbol{a}_i;\boldsymbol{\theta}_i)$ and the target DQN $Q_i(\boldsymbol{s}_i,\boldsymbol{a}_i;\boldsymbol{\theta}_i^-)$. However, since the purpose of the target DQN is only to obtain the target of approximation in (17), we can only focus on the trained DQN in all the following sections. Additionally, each agent needs to establish a replay memory $\mathcal{D}_i$ to store its experiences with the environment. At each time slot, UE i inputs its present state $\boldsymbol{s}_i(t)$ into its (trained) DQN, and chooses an action $\boldsymbol{a}_i(t)$ based on the approximated action value $Q_i(\boldsymbol{s}_i(t),\boldsymbol{a}_i;\boldsymbol{\theta}_i)$ according to the action selection policy in (15). After interacting with the environment using the selected action, the UE receives a reward $r_i(\boldsymbol{s}_i(t),\boldsymbol{a}_i(t))$ and a new state $\boldsymbol{s}_i(t+1)$, and this experience $\{\boldsymbol{s}_i(t),\boldsymbol{a}_i(t),r_i(\boldsymbol{s}_i(t),\boldsymbol{a}_i(t)),\boldsymbol{s}_i(t+1)\}$ is stored in the replay memory of UE i. Subsequently, the (trained) DQN is trained by the stochastic gradient descent (SGD) algorithm using a mini-batch size of experience samples from

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

the memory $\mathcal{D}_i$ to update the current parameters for action value approximation $\boldsymbol{\theta}_i$. Specifically, when updating the DQN parameters, the objective of UE i is to minimize the mean square error of the temporal difference (TD) over each training batch of samples, which can be formulated as a loss function.

Definition 5: The loss function of UE i is the mean square error of the TD over each training batch

$$\mathbb{L}(\boldsymbol{\theta}_i) \triangleq \mathbb{E}\left[\left(y_i^{DQN} - Q_i(\boldsymbol{s}_i(t), \boldsymbol{a}_i(t); \boldsymbol{\theta}_i)\right)^2\right], \tag{17}$$

where $y_i^{DQN} = r(\boldsymbol{s}_i(t), \boldsymbol{a}_i(t)) + \beta \max_{\boldsymbol{a}_i(t+1)\in\mathcal{A}} Q_i(\boldsymbol{s}_i(t+1), \boldsymbol{a}_i(t+1); \boldsymbol{\theta}_i^-)$ *is the target of approximation.*

In practice, UE i updates the parameters of the target DQN with the parameters of the trained DQN once per Z time slots [20].

B. Proposed Design for UE Selection

After presenting the proposed DRL for each UE, we next focus on the FL-enabled training procedure for the global DQN. As aforementioned, if each UE trains its DQN individually, convergence among the parameters of DQNs trained by different UEs may not be guaranteed. On the other hand, although convergence could be achieved using a centralized training method through collecting all the experiences from UEs, the signaling overheads may not be affordable. To address the above issues in the proposed design, FL [25] is introduced to develop global parameters in a distributed manner. In FL, the global parameters are disseminated to all the UEs, and UEs adopt the trained global model parameters to further refine their local model parameters. By iteratively aggregating the trained model parameters form different UEs, the global parameters are updated until achieving a desirable accuracy. Generally, the architecture of FL offers two key technical merits: 1) The global model server does not require UEs' complete data (or model) but only needs their local model parameters, to reduce the signaling overheads for UEs. 2) The nature of the distributed data storage/process in FL significantly reduces the complexity of the centralized data storage/process.

Aligning with the spirit of FL, the proposed FL-enabled training procedure is composed of two phases: UE selection and global model aggregation, while the proposed UE selection is presented in this section. To develop FL-enabled training procedure for maintaining global DQN parameters, a global model server installed in the RIC should select a certain set of UEs to perform local training (i.e., UE updates the parameters of its DQN based on its observations). The updated parameters are then sent to the global model server for further aggregation (this part will be detailed in the next section). Similar to the formulation of MDP for DRL, the UE selection is of crucial importance and may largely impact the performance of FL. Since the objective of FL is to minimize the long-term global loss function after a period of training, the optimization of the UE selection in FL can be formulated as a long-term loss function minimization problem, i.e.

$$\min_{\mathcal{U}'(t)} \mathbb{L}\left(\boldsymbol{\theta}_G(t)\right) = \min_{\mathcal{U}'(t)} \frac{1}{|\mathcal{U}'(t)|} \sum_{z=t}^{T_{train}} \sum_{i\in\mathcal{U}'(t)} \mathbb{L}\left(\boldsymbol{\theta}_i(z)\right), \tag{18}$$

where $|\cdot|$ is the 1-norm operation, $\mathcal{U}'(t)$ is the set of the selected UEs at time slot t, $\boldsymbol{\theta}_i(t)$ and $\boldsymbol{\theta}_G(t)$ are the local model parameters of UE i and global model parameters at time slot t, respectively, and T_{train} is the length of the training period.

Remark 1: In the Q-learning based DRL, the loss function is the mean square of the TD errors (i.e., the differences between the target of approximation and present approximation) during the interactions with the environment, as shown in (17)*, and the target of approximation is actually the expected accumulated reward. Hence, minimizing TD errors of all UEs is equivalent to maximizing the long-term accumulated rewards of UEs. Based on the loss function of DRL, above loss function minimization problem can be rewritten as*

$$\max_{\boldsymbol{b}_s(t)} \sum_{t=0}^{T} \sum_{i=0}^{M-1} b_i(t) \cdot r_i(\boldsymbol{s}_i(t), \boldsymbol{a}_i(t)), \tag{19}$$

where $\boldsymbol{b}_s(t)$ *is a vector of indicators*

$$\boldsymbol{b}_s(t) = \{b_0(t), \cdots, b_{M-1}(t)\} \tag{20}$$

and $b_i(t)$ *for all* i *is a binary indicator to indicate whether UE* i *is selected at time slot* t*. UE* i *is selected at time slot* t *if* $b_i(t) = 1$*, and* $b_i(t) = 0$ *otherwise. Therefore, the UE selection problem can be transformed to a long-term reward maximization problem.*

In **Remark 1**, let $\mathcal{A}_s$ denote the space of $\boldsymbol{b}_s(t)$. Since the total number of UEs is M, the number of possible solutions for the UE selection is $|\mathcal{A}_s| = 2^M$. Due to mobility of UEs, there exits uncertainty on the set of UEs having the best access decisions. Different from the greedy selection, in which the set of UEs with the current best performance are always chosen, the UCB algorithm [32] may provide a better exploration ability to predictively search for the best actions so as to achieve long-term maximization. With this technical merit, a UCB algorithm is adopted in our scheme to select UEs to perform local training, and maximize the long-term global model training performance in FL. UCB is a sort of RL algorithms, which therefore aims at maximizing the mean of accumulated rewards through choosing a proper action (i.e., UE selections). For this purpose, the reward at each time step of UCB should be adequately designed. Since the objective of UE selections is also to maximize the overall throughput and avoid frequent handovers, which aligns to the objective of DRL in each UE, the reward of UCB can be constructed based on the reward of DRL in each UE.

Proposition 2 (Reward of UCB): Note that the objective of the UE selection is to maximize the long-term accumulated overall rewards of UEs, which can be calculated by the global model server. Suppose that the UE selection is performed every n_f $(n_f \geq 0)$ *time slots, the reward of the designed UCB algorithm* $R_s(t)$ *is the sum of the accumulated rewards from all the UEs in* n_f *continuous time slots, i.e.,*

$$R_s(t) = \sum_{\tau=t}^{t+n_f-1} \sum_{i=0}^{M-1} r_i(\boldsymbol{s}_i(\tau), \boldsymbol{a}_i(\tau)). \tag{21}$$

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

With the provided reward, the global model server performs UE selection based on the following criterion

$$\boldsymbol{b}_s(t) = \arg\max_{\iota \in \mathcal{A}_s} \left(\hat{\mu}_\iota + \sqrt{\frac{2\ln t}{T_\iota (t-1)}} \right), \tag{22}$$

where T_ι is the number of time slots that action $\iota \in \mathcal{A}_s$ has been selected prior to time slot t, and $\hat{\mu}_\iota$ is an estimated value with the following updating rule

$$\hat{\mu}_\iota \leftarrow \hat{\mu}_\iota + \alpha_s \left(R_s(t) - \hat{\mu}_\iota \right). \tag{23}$$

where α_s is the learning rate of UCB. From (22), when action ι has been continuously selected for a long time, $\hat{\mu}_\iota + \sqrt{\frac{2\ln t}{T_\iota(t-1)}}$ becomes smaller. As a result, other $\boldsymbol{b}_s(t)$ may be selected. This design therefore facilitates exploration and avoids the result being trapped in the local optimum.

C. Design for Aggregation Stage

In this section, we continue the discussion on the proposed design for the model aggregation in FL. In the proposed scheme, all the UEs have a common objective to maximize the overall throughput and minimize the number of handovers.

Proposition 3: In FL, the FedAvg *algorithm [33] is a widely adopted method to train the global model parameters using the following rule*

$$\boldsymbol{\theta}_G(t) = \frac{1}{\sum\limits_{i \in \mathcal{U}'(t)} D_i} \sum_{i \in \mathcal{U}'(t)} D_i \boldsymbol{\theta}_i(t), \tag{24}$$

where D_i is the size of the dataset of UE i. From (24)*, the ratio of each selected UE's datasize to the sum of datasizes in each model aggregation phase can be regarded as the weight for corresponding UE's DQN parameters. In the proposed scheme, for all the UEs with a common objective, UEs can be regarded as homogeneous with the same datasize in the training phase, and the DQN parameters from each selected UE can be given the same weight. Therefore, the updating rule of the model aggregation operation can be the average of parameters from the selected UEs,*

$$\boldsymbol{\theta}_G(t) = \frac{1}{|\mathcal{U}'(t)|} \sum_{i \in \mathcal{U}'(t)} \boldsymbol{\theta}_i(t). \tag{25}$$

Since the reward of each UE is influenced by other UEs' actions, UEs may suffer from non-stationary issues in the training phase. To alleviate this issue, in our proposed scheme, each UE regards the other UEs as part of the environment, and the global DQN parameters as the average common knowledge for the environment are shared among UEs. In the following theorem, the convergence of the global DQN parameters using the proposed model aggregation is justified.

Theorem 1: When UEs are selected uniformly and the UE selection is performed for a large number of runs, the final parameters of the global DQN trained by the global model server using FL converge to

$$\boldsymbol{\theta}_G^* = \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{\theta}_i^*. \tag{26}$$

where $\boldsymbol{\theta}_i{}^*$ are the converged DQN parameters of UE i.

Proof: According to (25), the updated parameters of the global DQN in the model aggregation can be rewritten as

$$\begin{aligned} \lim_{t\to\infty} \boldsymbol{\theta}_G(t+1) &= \lim_{t\to\infty} \frac{1}{|\mathcal{U}'(t)|} \sum_{i \in \mathcal{U}'(t)} \left(\boldsymbol{\theta}_G(t) - g_{\boldsymbol{\theta}_i}(t) \right) \\ &= \lim_{t\to\infty} \boldsymbol{\theta}_G(t) - \lim_{t\to\infty} \frac{1}{|\mathcal{U}'(t)|} \sum_{i \in \mathcal{U}'(t)} g_{\boldsymbol{\theta}_i}(t), \end{aligned} \tag{27}$$

where $g_{\boldsymbol{\theta}_i}(t) = (\boldsymbol{\theta}_i(t+1) - \boldsymbol{\theta}_i(t))/\alpha$ is the gradient of UE i at time slot t. Noted that the state space and action space defined in Section III-A are finite and the reward function is bounded. When the learning rate α satisfies $\lim\limits_{T\to\infty} \sum\limits_{t=0}^{T-1} \alpha = \infty$ and $\lim\limits_{T\to\infty} \sum\limits_{t=0}^{T-1} \alpha^2 < \infty$, the convergence of tabular Q-function can be achieved after a large number of updates [34]. According to [35], the convergence of the DQN (with particular activation functions) can be proved theoretically, which indicates that the convergence of the parameters $\boldsymbol{\theta}_i$ can be achieved after a large number of updates. Therefore, when $t \to \infty$, $g_{\boldsymbol{\theta}_i}(t)$ approaches zero, and the convergence of $\boldsymbol{\theta}_G$ can be achieved in (27).

Furthermore, since each UE is selected uniformly, the probability of UE i being selected is $\frac{1}{M}$. After a large number of runs, the sample average of the parameters of selected UEs approaches to the mean of the parameters of selected UEs

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{|\mathcal{U}'(t)|} \sum_{i \in \mathcal{U}'(t)} \boldsymbol{\theta}_i(t) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=0}^{M-1} \frac{1}{M} \boldsymbol{\theta}_i(t) \\ &= \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{\theta}_i^*. \end{aligned} \tag{28}$$

By substituting (28) to (25), the parameters of the global DQN converge to (26). □

Theorem 1 shows that the converged global DQN parameters are actually the mean value of the converged DQN parameters trained by all UEs separately. Thus, the converged DQN parameters can be regarded as a consensus of the best decisions obtained by multiple UEs. In this case, before reaching convergence, there must exist a bias $\chi_i(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i)$ in the estimated action value of UE i along with the consensus result. The action value can therefore be generally formulated as the sum of a consensus function $y_c(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i)$ and a bias function $\chi_i(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i)$, i.e.,

$$Q_i(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i) = y_c(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i) + \chi_i(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i). \tag{29}$$

If all the DQN parameters of UEs are aggregated directly to generate the global DQN parameters, the bias of each UE i cannot be adequately presented and effectively eliminated. As a consequence, the global model parameters may not converge to the optimum result [36]. To avoid this undesirable case, only a part of DQN parameters of UEs should be aggregated as the consensus of all the UEs, while the bias function is trained by each UE locally.

As illustrated in Fig. 2, different from the traditional DQN structure, $Q_i(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_i)$ in the adopted dueling DQN [36],

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

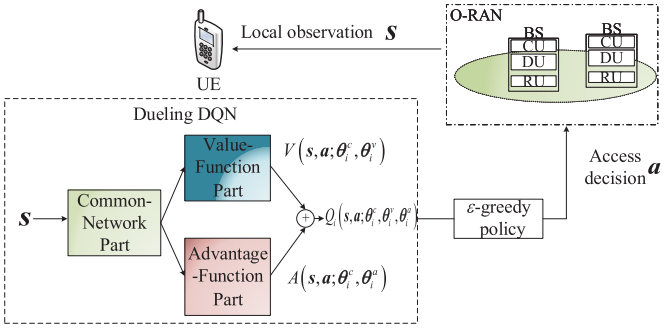


Fig. 2. The structure of dueling DQN.

rewritten as $Q_i(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v,\boldsymbol{\theta}_i^a)$, is decomposed as two parts: the value function $V(\boldsymbol{s};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v)$ and the advantage function $A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a)$, where $\boldsymbol{\theta}_i^c$, $\boldsymbol{\theta}_i^v$ and $\boldsymbol{\theta}_i^a$ are the common-network parameters, value-function parameters, and advantage-function parameters, respectively. The common-network part, which can be a group of fully-connected layers in the DNN, is deployed as the front part in the dueling DQN to extract features from the inputted state. The copies of the output of the common-network part as two independent streams are inputted into the value-function part and advantage-function part, respectively. The value function is only used to estimate the value of state, and the advantage function represents the importance of each action in the given state. The expression of $Q_i(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i)$ in the dueling structure therefore can be decomposed to

$$Q_i(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v,\boldsymbol{\theta}_i^a) = V(\boldsymbol{s};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v) + A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a). \quad (30)$$

Since the consensus of all UEs is the aggregation of decision policies, the value-function parameters $\boldsymbol{\theta}_i^v$ and the common-network parameters $\boldsymbol{\theta}_i^c$ are used to represent the consensus function $y_c(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_G)$. While since the bias function can be regarded as the preference of each action for UE i, the advantage-function parameters $\boldsymbol{\theta}_i^a$ are used to represent the bias function $\chi_i(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i)$. Since the results of the value function and advantage function cannot be recovered given a specific action value, (30) is unidentifiable and the performance may largely degrade if (30) is applied directly. To improve the identifiability of $V(\boldsymbol{s};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v)$ and $A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a)$, a constant can be added to $V(\boldsymbol{s};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v)$ and the same constant can be subtracted from $A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a)$. The average of $A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a)$ is therefore introduced and the expression used in practical implementations is given by

$$Q_i(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v,\boldsymbol{\theta}_i^a) = V(\boldsymbol{s};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v) + A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a) - \frac{1}{|\mathcal{A}|}\sum_{\boldsymbol{a}\in\mathcal{A}} A(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^a). \quad (31)$$

In each training step of FL, the selected UEs only present their common-network and value-function parameters $(\boldsymbol{\theta}_i^c,\boldsymbol{\theta}_i^v)$ to the global model server in the uplink reports. In the following proposition, the model aggregation operations of $\boldsymbol{\theta}_i^c$ and $\boldsymbol{\theta}_i^v$ at the global model server side are provided.

Proposition 4: In the global model aggregation phase, the global common-network and value-function parameters are obtained by averaging the corresponding parameters from the selected UEs, i.e.,

$$\boldsymbol{\theta}_G^c(t) = \frac{1}{|\mathcal{U}'(t)|}\sum_{i\in\mathcal{U}'(t)} \boldsymbol{\theta}_i^c(t), \quad (32)$$

$$\boldsymbol{\theta}_G^v(t) = \frac{1}{|\mathcal{U}'(t)|}\sum_{i\in\mathcal{U}'(t)} \boldsymbol{\theta}_i^v(t). \quad (33)$$

After aggregation, the obtained parameters $\boldsymbol{\theta}_G^c(t)$ and $\boldsymbol{\theta}_G^v(t)$ are sent back to each UE. UE i then combines the newly obtained parameters and the locally trained advantage-function parameters $\boldsymbol{\theta}_i^v(t)$ as new parameters of its local (trained) DQN, i.e., $\boldsymbol{\theta}_i(t+1) = \{\boldsymbol{\theta}_i^a(t),\boldsymbol{\theta}_G^c(t),\boldsymbol{\theta}_G^v(t)\}$. In the subsequent time slot, each UE makes decisions based on the new local (trained) DQN.

D. The Proposed Federated DRL-Based Scheme

After providing the UE selection scheme and the dueling structure, we are now ready to propose the federated DRL-based scheme for user access control, as illustrated in Fig. 3. To implement access control at the UE side, each UE needs to learn the hidden long-term variation pattern by using proposed DRL framework to interact with the environment. Since the objective is to maximize the overall performance, UEs should not only consider their local information but also concentrate on the common knowledge of the environment, which is presented by the global DQN parameters. Moreover, since the decisions of UEs are influenced by each other, if the complete local DQN parameters of UEs are considered, these complete local DQN parameters may not be independent. To optimally train global DQN parameters, the common knowledge of the environment should be represented without considering UEs' local preferences in actions. In this case, only partial DQN parameters of each UE are adopted to present the consensus of UEs, and partial parameters are trained locally to preserve local preferences to access actions. Specifically, in the proposed scheme, each UE makes access decisions based on its trained DQN with the parameters $\{\boldsymbol{\theta}_i^a(t),\boldsymbol{\theta}_G^c(t),\boldsymbol{\theta}_G^v(t)\}$, which is trained by an FL-enabled training algorithm. Additionally, to achieve long-term performance optimizations, a global model server installed in the RIC is responsible for selecting UEs to perform local training by using the provided UCB-based selection scheme, and aggregating the parameters collected from the selected UEs to update the global DQN parameters. The temporal structure of the proposed scheme consists of two phases, i.e., the initialization phase and the training phase.

- *Initialization Phase:* At the beginning of the initialization phase, each UE needs to establish two DQNs, i.e. the trained DQN $Q_i(\boldsymbol{s}_i,\boldsymbol{a}_i;\boldsymbol{\theta}_i)$ and the target DQN $Q_i(\boldsymbol{s}_i,\boldsymbol{a}_i;\boldsymbol{\theta}_i^-)$ with the arbitrarily initialized parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i^-$, and the replay memory $\mathcal{D}_i$ to store its interaction experience $\{\boldsymbol{s}_i(t),\boldsymbol{a}_i(t),r_i(\boldsymbol{s}_i(t),\boldsymbol{a}_i(t)),\boldsymbol{s}_i(t+1)\}$ at each time slot. In the first N_b slots, UE i makes access decisions using the RSS-based method, and subsequently receives N_b corresponding rewards. Then, these N_b interaction experiences are stored into its replay memory as initialization training data.

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

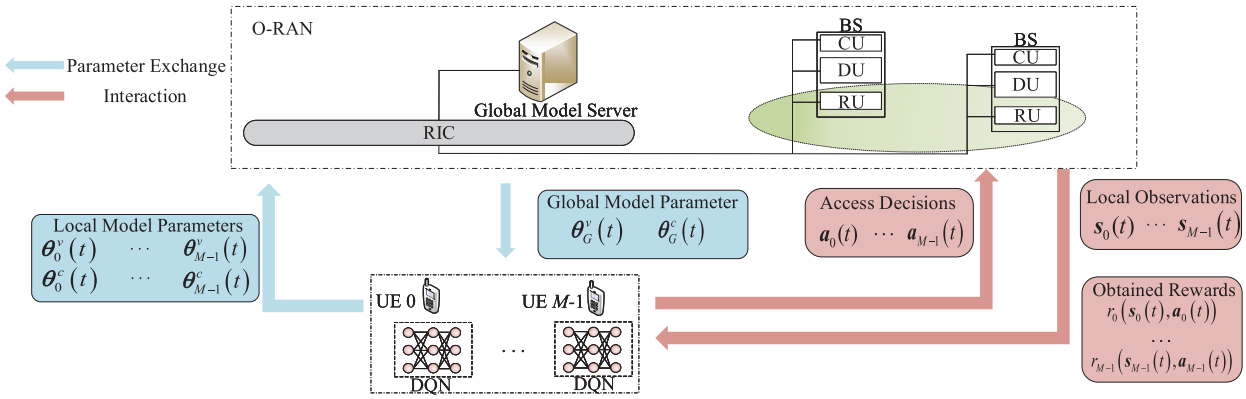


Fig. 3. The proposed user access control scheme.

- *Training Phase:* In each time slot t $(t > N_b)$ of the training phase, UE i visits a state, and makes its access decision based on the trained DQN. When multiple UEs send access requests to the same BS for the same RB, the BS would allocate the RB to the UE with the strongest request in terms of signal strength, and allocate its available RBs to other UEs. After executing the selected action, each UE receives an immediate reward and visits a new state. The interaction experiences are stored into its replay memory. Let t_{start} denote a local record of time slot for the last training round. In every n_f time slots, (i.e., $t - t_{start} = n_f$), the global model server would select a number of UEs for performing the local DQN training according to the UCB-based scheme proposed in Section III-B. Then, each selected UE samples a mini-batch size of experiences randomly from the replay memory as training data. The parameters $\boldsymbol{\theta}_i(t)$ are updated using the SGD algorithm to minimize the loss function defined in (17). Subsequently, each selected UE sends its common-network and value-function parameters $(\boldsymbol{\theta}_i^c(t), \boldsymbol{\theta}_i^v(t))$ to the global model server. Next, the global model server performs model aggregation to obtain new global common-network and value-function parameters $(\boldsymbol{\theta}_G^c(t), \boldsymbol{\theta}_G^v(t))$, and sends these newly obtained parameters to all the UEs. Finally, when UEs receive these global DQN parameters, the received parameters are combined with the locally trained advantage-function parameters $\boldsymbol{\theta}_i^a(t)$ as the new local DQN parameters, which are used for access decision in the next time slot.

In Algorithm 1, above procedure of the proposed scheme is summarized. If a new UE arrives at the O-RAN, it only needs to download parameters of the global DQN from the global model server, and then makes access decisions based on its DQN with the downloaded parameters.

E. Complexity and Signaling Overheads

In this section, we analyze the complexity and signaling overheads of the proposed scheme. For performance comparison, two conventional schemes are considered, i.e., the distributed training scheme and centralized training scheme.

- *Distributed Training:* In the distributed training scheme, each UE trains its DQN with local interaction experiences independently. In particular, there is no information exchange between each UE, and all the UEs make decisions according to their own trained DQNs.
- *Centralized Training:* In this scheme, there exists a centralized trainer responsible for training a global DQN for UEs. In the training phase, UEs present their interaction experiences to the centralized trainer as training data. The trainer updates the parameters of the DQN based on the collected experiences, which are sent back to UEs for access decisions.

In this paper, the fully-connected network is adopted to construct the DQN for each UE. Let Z^c, Z^v, Z^a and n_z^c, n_z^v, n_z^a denote the number of hidden layers of the common-network part, value-function network part, advantage-function network part, and the number of neurons in the zth hidden layers of the common-network part, value-function network part, advantage-function network part, respectively, the time complexity of the adopted DQN can be $\zeta_{DQN} = J((2(N+K)+1) \times n_1^c + n_{Z^v}^v + n_{Z^a}^a NK + n_{Z^c}^c(n_1^v + n_1^a) + \sum_{z=1}^{Z^c-1} n_z^c \times n_{z+1}^c + \sum_{z=1}^{Z^v-1} n_z^v \times n_{z+1}^v + \sum_{z=1}^{Z^a-1} n_z^a \times n_{z+1}^a)$, where $J(\cdot)$ is the time complexity for updating the parameters of the fully-connected layers. Therefore, for each UE, the complexity of training a DQN in a single time slot is $\mathcal{O}(\zeta_{DQN} \times N_b)$. Hence, the time complexity of the proposed scheme is $\mathcal{O}(\sum_{z=1}^{\bar{T}} \zeta_{DQN} \times N_b \times |\mathcal{U}'(z \times n_f)|)$, where $\bar{T} = \left\lfloor \frac{T}{n_f} \right\rfloor$. For the distributed training scheme, the time complexity is $\mathcal{O}(\bar{T} \times M \times \zeta_{DQN} \times N_b)$ since all the UEs train their DQNs independently. For the centralized scheme, only the centralized trainer node trains the global DQN, and therefore the time complexity is $\mathcal{O}(T \times \zeta_{DQN} \times N_b)$.

After analyzing the complexity, we further analyze the signaling overheads of these three schemes. In the proposed scheme, the selected UEs transmit partial DQN parameters with size of D_F to the global model server in the uplink reports, and the global model server sends the aggregated DQN parameters back to all the UEs in the downlink transmissions, where $D_F = (2(N+K)+1) \times n_1^c + n_{Z^v}^v + n_{Z^c}^c(n_1^v) + \sum_{z=1}^{Z^c-1} n_z^c \times n_{z+1}^c + \sum_{z=1}^{Z^v-1} n_z^v \times n_{z+1}^v$ is the datasize of transmitted partial parameters. While in the distributed training

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

scheme, each UE trains its DQN independently without information exchange. Therefore, there is no signaling overhead in this distributed training scheme. In the conventional centralized training scheme, each UE sends its interaction experience to the centralized trainer at each time slot, and the total datasize of transmitted experience is $4M(N+K+1)$. Then the centralized trainer announces the parameters of the trained DQN with size of D_C in the downlink transmissions at each time slot, where $D_C = D_F + n^a_{Z^a} NK + n^c_{Z^c}(n^a_1) + \sum_{z=1}^{Z^a-1} n^a_z \times n^a_{z+1}$ is the datasize of the complete DQN parameters.

Algorithm 1 The Proposed Federated DRL-Based Scheme

1: **Initialize Stage**:
2: Each UE constructs two DQNs, i.e., $Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i; \boldsymbol{\theta}_i)$ and $Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i; \boldsymbol{\theta}_i^-)$ with randomly initialized parameters, and a replay memory $\mathcal{D}_i$. The global model server installed in the RIC is initialized.
3: In the first N_b time slots, each UE accesses the BS with the strongest RSS and stores corresponding interaction experiences into its replay memory.
4: **Training Stage**:
5: $t_{start} \leftarrow t$
6: **repeat**
7: Each UE visits a state $\boldsymbol{s}_i(t)$ at each time slot t, and makes access decision $\boldsymbol{a}_i(t)$ based on its trained DQN.
8: Each UE obtains a reward $r_i(\boldsymbol{s}_i(t), \boldsymbol{a}_i(t))$ and visits the state $\boldsymbol{s}_i(t+1)$ at time slot $t+1$.
9: Store experience $\{\boldsymbol{s}_i(t), \boldsymbol{a}_i(t), r_i(\boldsymbol{s}_i(t), \boldsymbol{a}_i(t)), \boldsymbol{s}_i(t+1)\}$ into its replay memory $\mathcal{D}_i$.
10: $t \leftarrow t+1$
11: **if** $t - t_{start} == n_f$ **then**
12: The global model server selects UEs from the UE set $\mathcal{U}$ according to the provided UCB-based scheme in (22).
13: Each selected UE randomly samples a mini-batch size of experiences from its replay memory $\mathcal{D}_i$ and updates its local DQN parameters by minimizing the loss function defined in (17) using the SGD algorithm.
14: Each selected UE reports its trained common-network parameters $\boldsymbol{\theta}_i^c(t)$ and value-function parameters $\boldsymbol{\theta}_i^v(t)$ to the global model server.
15: The global model server aggregates global parameters from the selected UEs based on (32) and (33).
16: The server sends the newly obtained parameters back to all the UEs.
17: Each UE combines its local advantage-function parameters $\boldsymbol{\theta}_i^a(t)$ and newly obtained parameters as new local DQN parameters to make access decision.
18: The server updates the UCB-based UE selection scheme based on (23)
19: **end if**
20: **until** The global loss decreases a desirable value (e.g., 0.05).

IV. Performance Evaluation

A. Simulation Settings

To evaluate the performance of the proposed scheme, we consider an O-RAN deployment in which $N = 6$ BSs with different transmission power levels share $K = 10$ RBs to serve $M = 20$ UEs located randomly in a 1000 m $\times$ 1000 m square area. Each UE moves with a fixed velocity along a certain direction. The Gauss-Markov mobility model [37] is adopted to generate the mobility patterns of UEs. For the construction of DQNs, the provided dueling structure is adopted, and the hyperparameters of the DQN are determined using the cross-validation algorithm [38]. Specifically, there are 2 fully-connected layers with a sigmoid function, 3 fully-connected layers with a sigmoid function and 3 fully-connected layers with a sigmoid function in a DQN to represent the common-network part, value-function part and advantage-function part, respectively. Other main simulation parameters are summarized in Table II.

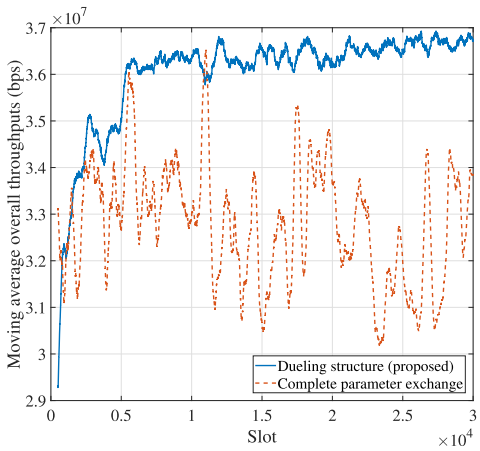


Fig. 4. Moving average overall throughputs with different DQN parameter exchange methods with $n_f = 10$ (moving average of the results in previous 500 slots).

For performance comparison benchmarkings, the RSS-based scheme [5], Q-learning [39], particle swarm optimization (PSO) algorithm, distributed training scheme and centralized training scheme are considered. In the RSS-based scheme, each UE accesses the BS and RB with the strongest RSS in each time slot. In the Q-learning, each UE makes access decisions based on its Q-table according to a softmax selection policy. The PSO algorithm is an evolutionary searching algorithm to find access decisions for UEs [40], which is a commonly method to solve a binary integer programming problem. On the other hand, the operations of the distributed training scheme and the centralized training scheme are aforementioned in Section III-E.

B. Results and Analysis

1) Performance With Different DQN Parameter Exchange Methods: As aforementioned in (29), if all the DQN parameters of UEs are averaged directly to generate the global DQN parameters, the global model parameters may not converge to the optimum result. To address this issue, a dueling structure has been provided in our scheme. To justify the performance of this design, we investigate the throughputs of the proposed scheme with different DQN parameter exchange methods in Fig. 4. From Fig. 4, we can observe that the overall throughput grows significantly over the time if the proposed dueling structure is adopted, instead of the complete parameter exchange method. Particularly, since the decision-makings are performed based on historical network information, the performance

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

TABLE II
PARAMETERS FOR SIMULATION

Parameters	Value	Parameters	Value
BS transmission power	40 dBm / 30 dBm	Common-network part	128×64
Bandwidth of each RB	180 kHz	Value-function part	$32 \times 16 \times 1$
Noise power density	-174 dBm/Hz	Advantage-function part	$64 \times 64 \times NK$
BS large-scale fading	$34 + 40\log(d)/37 + 30\log(d)$ [24]	SGD Optimizer	Adam [21]
Carrier frequency	2 GHz	Learning rate (α)	0.001 [23]
Discounting factor	0.9 [24]	Mini-batch size (N_b)	32 [24]
Replay memory size	2000 [23]	Target DQN updating period (Z)	200 [18]

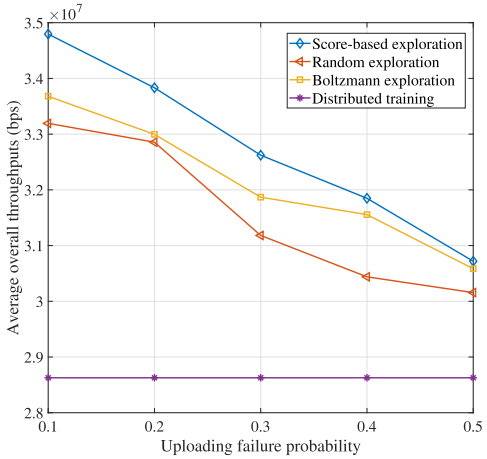


Fig. 5. Average overall throughputs with different uploading failure probabilities.

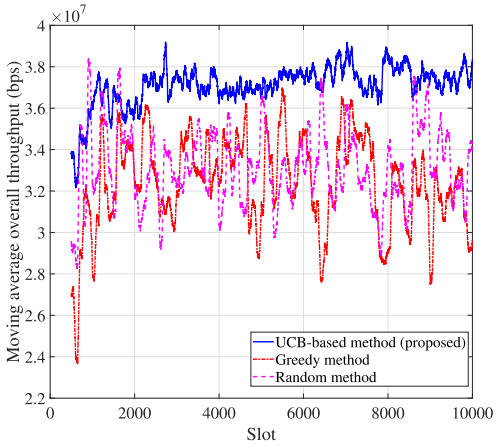


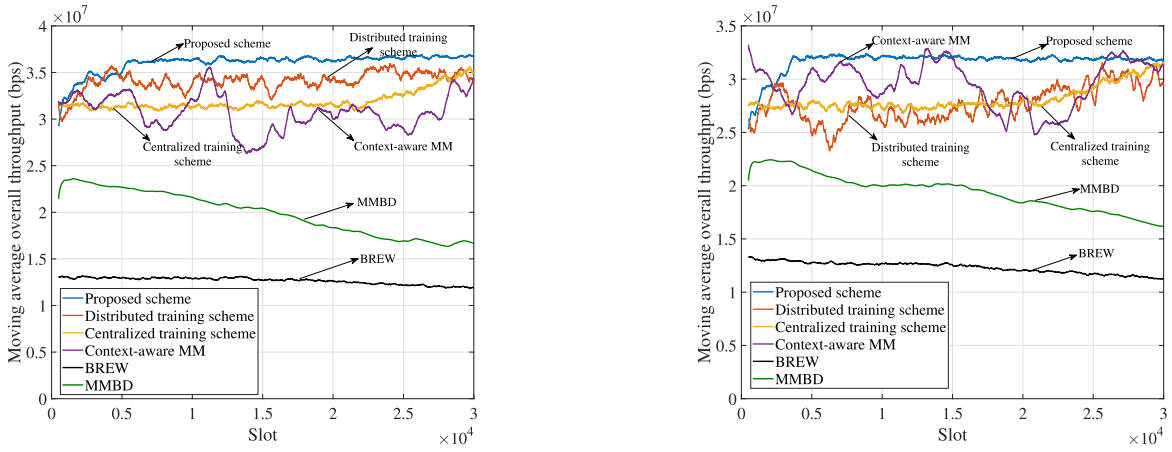
Fig. 6. Moving average overall throughputs with different UE selection methods.

achieved by the proposed scheme is sub-optimal. The result shows that UEs are able to learn proper access policies from interactions with the environment. Nevertheless, the overall throughput of the adopted dueling structure is higher than that of the complete parameter exchange method, which justifies that the adopted dueling structure offers a greater generation ability in the FL-enabled training algorithm.

2) Performance With Different Uploading Failure Probabilities: Next, due to the imperfect channel conditions, UEs uploading their local model parameters may suffer from different levels of transmission latency. A transmission failure may consequently occur if the transmission latency is too severe. To analyze the impact of transmission latency (and thus transmission failure), we should evaluate the overall throughputs of the proposed scheme with different uploading failure probabilities in Fig. 5. In the meantime, we also evaluate the performances of the distributed training scheme (i.e., only distributed DRL in each UE without FL), the proposed scheme with a random exploration algorithm (i.e., in the ϵ-greedy action selection, each non-greedy action is selected with an equal probability), the proposed scheme with a Boltzmann exploration algorithm (i.e., each non-greedy action is sampled based on a Boltzmann distribution calculated by current action values), and the proposed scheme with a score-based exploration algorithm (i.e., in the ϵ-greedy policy selection, if the greedy action is not selected, the non-greedy action with the highest score is selected, which is calculated according to (16)). From Fig. 5, we can observe that the overall throughput achieved by the proposed scheme with the score-based exploration is higher than those of the random and Boltzmann exploration since historical information is exploited in the score-based exploration, which guides the agent to select actions with the highest long-term channel quality instead of blind exploration or current estimated action distribution. Additionally, with the increase of the uploading failure probabilities, the performance of the proposed scheme decreases and is gradually close to that of the distributed training scheme. Nevertheless, the proposed scheme outperforms the distributed DRL training scheme in the range of practical uploading failure probabilities. These results therefore suggest the robustness of the proposed scheme against potential transmission delay.

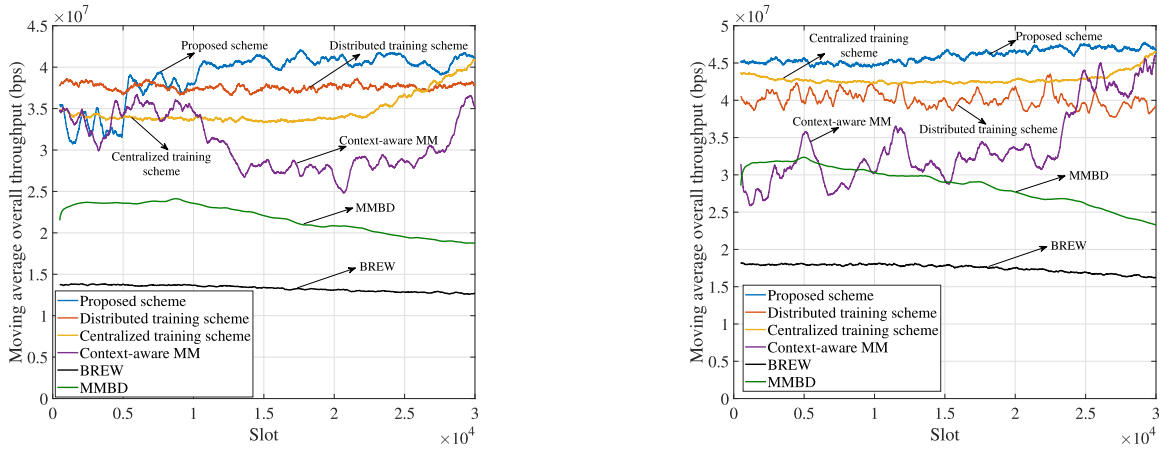
3) Performance With Different UE Selection Methods: Next, to justify the practicability of the provided UCB-based UE selection method, the throughputs of the proposed federated DRL-based user access control scheme with different UE selection methods are evaluated in Fig. 6, including the proposed UCB-based selection method, greedy selection method and random selection method. In the greedy method, UEs with the present maximum throughput are selected for training local DQNs in the training phase. In other words, the greedy method only selects UEs with the best immediate performance but ignores the long-term performance, which may lead to a poor generalization ability. On the other hand, in the random selection method, UEs are randomly selected in the training phase. Since the objective of the UCB-based method is to maximize the long-term overall throughput, it can effectively strike the balance between exploration and exploitation. As a

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.



(a) $\rho = 0.9$ (b) $\rho = 0.2$

Fig. 7. Moving average overall throughputs with different coherent coefficients with $M = 20$.



(a) $M = 30$ and $N = 6$ (b) $M = 40$ and $N = 8$

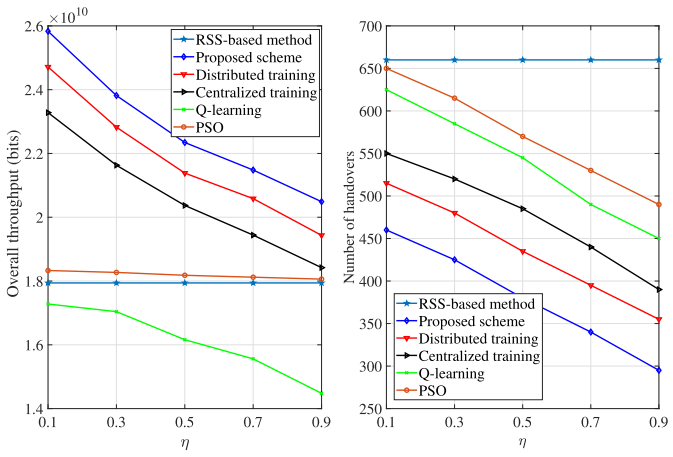
Fig. 8. Moving average overall throughputs with different numbers of UEs and BSs for $\rho = 0.9$.

result, the provided UCB-based UE selection method achieves the highest overall throughput over the other two methods.

4) Performance With Different Coherent Coefficients: To evaluate the performance of the proposed scheme under different channel conditions, in Fig. 7, the simulation results in terms of the overall throughputs of the proposed scheme, distributed training scheme and centralized training scheme under different coherent coefficients are provided. Moreover, we also evaluate the performances of context-aware mobility management (MM) algorithm [9], batched randomization with exponential weighting (BREW) algorithm [15] and mobility management with batched D-UCB (MMBD) algorithm [16], which are based on multi-arm bandit model in RL. Particularly, in context-aware MM algorithm, user access control is first performed based on the adjusted cell biases of BSs then RBs are allocated to connected UEs according to their historical throughputs and velocities, and the algorithm can be regarded as one kind of cell-based scheme. Additionally, since there is no RB allocation design involved in BREW and MMBD, UEs are allocated by RBs with the best channel qualities after accessing a specific BS in our simulations. For $\rho = 0.9$, the small-scale fading is not severe, and the channel gain varies slowly. On the contrary, the channel gain varies fast for $\rho = 0.2$. From Fig. 7, we can observe that the throughputs of all the three DRL-based schemes (i.e., proposed scheme, distributed and centralized training schemes) when $\rho = 0.9$ are higher than those when $\rho = 0.2$, since the variation pattern of small-scale fading can be effectively learned when the channel condition varies slowly. Under both cases of a slowly fading channel ($\rho = 0.9$) and a fast fading channel ($\rho = 0.2$), the proposed scheme significantly outperforms the other training schemes with the aid of model aggregation in the global model server. The reason is that the influences between UEs are not considered in BREW and MMBD algorithms. For the context-aware MM algorithm, it mainly concentrates on the historical throughput for fairness among UEs, and channel gain variations between BS-UE pairs are not considered. These results justify the effectiveness and robustness of the proposed scheme under any channel conditions.

5) Performance With Different Numbers of UEs and BSs: Subsequently, we evaluate the performance of the proposed scheme with different numbers of UEs and BSs in Fig. 8.

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.



(a) Overall throughputs (b) Number of handovers

Fig. 9. Overall throughput and numbers of handover with different handover costs and punishment factors in the reward function with $M = 20$ and $\rho = 0.9$. (All the results are calculated over 500 time slots in which the performances of all the schemes have converged.

For performance comparison, we also evaluate performances of the benchmarkings adopted in Fig. 7. When there are large numbers of UEs and BSs ($M = 40$ and $N = 8$), the overall throughput achieved by the proposed scheme is higher than that when there are smaller numbers of UEs and BSs ($M = 30$ and $N = 6$). This indicates that the overall throughput increases gradually with the growth of numbers of UEs and BSs. Besides, in different network scales, the proposed scheme outperforms the other three bandit-based algorithms (i.e., context-aware MM, BREW and MMBD) and the distributed training scheme since UEs are only able to estimate the environment variations from their own observations in these algorithms. The performances of the proposed scheme and the centralized scheme are around the same level, while the convergence time of the proposed scheme is far shorter than that of the centralized scheme. This indicates that information sharing between UEs is necessary as the network scale increases. Additionally, the convergence of the proposed scheme is not influenced by the numbers of UEs and BSs in the O-RAN.

6) Performance With Different Handover Punishment Factors: Finally, we evaluate the impact of the punishment factor η in the reward function to the overall throughput and the number of handovers. For performance comparison, we also evaluate the overall throughputs of the RSS-based method, PSO algorithm, Q-learning, distributed training method and centralized method. From Fig. 9(a) and 9(b), we can observe that UEs with the proposed scheme can achieve the highest overall throughput and the least number of handovers as compared with the other five schemes. In the RSS-based scheme, multiple UEs would access the same BS, and the access decisions are heavily influenced by mobility of UEs, leading to frequent handovers. For Q-learning, since there exists a high-dimensional decision space to explore, the Q-table cannot be fully updated, resulting in a poor throughput performance. Additionally, the overall throughput and the number of handovers decrease as the value of η increases. Hence, the proposed scheme with a larger η leads to a lower overall throughput and less handovers. Different levels of tradeoff between the overall throughput and the number of handovers can therefore be achieved by adopting different η.

V. Conclusion

To address the foundational issues in the O-RAN to maximize the overall throughput and avoid frequent handovers, UEs need to access proper BSs at each time slot through maximizing the long-term utility. To tackle this optimization problem in practical O-RAN deployment, we propose a federated DRL-based scheme for user access control, in which each UE independently makes access decisions with its DQN, and a global model server installed in the RIC updates the global DQN parameters by aggregating DQN parameters obtained from the selected UEs. To select proper UEs to provide their DQN parameters, we develop a UCB-based UE selection method to find the optimal UE set in the training phase of the proposed FL-enabled training algorithm. To further facilitate convergence of the proposed scheme, a dueling structure is provided to decompose the DQN parameters at the UE side, and the selected UEs only exchange the common-network and value-function parameters with the global model server. Through comparing the performance with different training schemes and the state-of-the-art schemes, the simulation results have shown that a significant throughput enhancement and less number of handovers can be achieved by the proposed scheme over these benchmarkings under different channel conditions, numbers of UEs and BSs and handover punishment factors. Our proposed scheme therefore offers an effective design of intelligent user access control for the O-RAN. It can be interesting to extend the proposed scheme to construct RL-based decision scheme for multi-layer networks with different control cycles and establish more general resource allocation mechanisms (e.g. power control and dynamic beamforming) in the O-RAN.

References

[1] O-RAN Alliance, "O-RAN: Towards an open and smart RAN," O-RAN White Paper, Oct. 2018.
[2] O-RAN Alliance, "O-RAN architecture description," O-RAN-WG1-O-RAN Architecture Description-v01.00.00, Tech. Rep., Oct. 2020.
[3] O-RAN Alliance, "AI/ML workflow description and requirements," O-RAN-WG2-AIML-v01.01, Tech. Rep., Mar. 2020.
[4] H. G. Moussa and W. Zhuang, "Access point association in uplink two-hop cellular IoT networks with data aggregators," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5386–5400, Jun. 2020.
[5] H. A. Mahmoud, I. Guvenc, and F. Watanabe, "Performance of open access femtocell networks with different cell-selection methods," in *Proc. IEEE 71st Veh. Technol. Conf.*, May 2010, pp. 1–5.
[6] K. Da Costa Silva, Z. Becvar, and C. R. L. Frances, "Adaptive hysteresis margin based on fuzzy logic for handover in mobile networks with dense small cells," *IEEE Access*, vol. 6, pp. 17178–17189, 2018.
[7] K. Ghanem, H. Alradwan, A. Motermawy, and A. Ahmad, "Reducing ping-pong handover effects in intra EUTRA networks," in *Proc. 8th Int. Symp. Commun. Syst., Netw. Digit. Signal Process. (CSNDSP)*, Jul. 2012, pp. 1–5.
[8] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in 5G and beyond: A topology aware skipping approach," *IEEE Access*, vol. 4, pp. 9073–9081, 2016.
[9] M. Simsek, M. Bennis, and I. Guvenc, "Mobility management in HetNets: A learning-based perspective," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 26, Feb. 2015.
[10] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Channel-access-aware user association with interference coordination in two-tier downlink cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5579–5594, Jul. 2016.

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

[11] C.-H. Liu, D.-C. Liang, K.-C. Chen, and R.-H. Gau, "Ultra-reliable and low-latency communications using proactive multi-cell association," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3879–3897, Jun. 2021, doi: 10.1109/TCOMM.2021.3065979.
[12] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[13] S. Bayat, R. H. Y. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027–3043, Aug. 2014.
[14] C. Xu, G. Zheng, and L. Tang, "Energy-aware user association for NOMA-based mobile edge computing using matching-coalition game," *IEEE Access*, vol. 8, pp. 61943–61955, 2020.
[15] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3854–3868, Dec. 2016.
[16] Y. Zhou, C. Shen, and M. van der Schaar, "A non-stationary online learning approach to mobility management," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1434–1446, Feb. 2019.
[17] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.
[18] K. Qu, W. Zhuang, X. Shen, X. Li, and J. Rao, "Dynamic resource scaling for VNF over nonstationary traffic: A learning approach," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 2, pp. 648–662, Jun. 2021, doi: 10.1109/TCCN.2020.3018157.
[19] J. Tian, Y. Pei, Y.-D. Huang, and Y.-C. Liang, "Modulation-constrained clustering approach to blind modulation classification for MIMO systems," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 4, pp. 894–907, Dec. 2018.
[20] Y.-C. Liang, *Dynamic Spectrum Management: From Cognitive Radio to Blockchain and Artificial Intelligence*. Singapore: Springer, 2020, doi: 10.1007/978-981-15-0776-2.
[21] Q. Zhang, Y.-C. Liang, and H. V. Poor, "Intelligent user association for symbiotic radio networks using deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4535–4548, Jul. 2020.
[22] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.
[23] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.
[24] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.
[25] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.
[26] Y. Cao, S.-Y. Lien, Y.-C. Liang, and K.-C. Chen, "Federated deep reinforcement learning for user access control in open radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6, doi: 10.1109/ICC42927.2021.9500603.
[27] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.
[28] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.
[29] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, Nov. 2019.
[30] S.-Y. Lien, S.-L. Shieh, Y. Huang, B. Su, Y.-L. Hsu, and H.-Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.
[31] L. Liang, J. Kim, S. C. Jha, K. Sivanesan, and G. Y. Li, "Spectrum and power allocation for vehicular communications with delayed CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 458–461, Aug. 2017.
[32] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.
[33] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016, *arXiv:1602.05629*.
[34] F. S. Melo, "Convergence of Q-learning: A simple proof," Inst. Syst. Robot., Lisboa, Portugal, Tech. Rep., 2001, pp. 1–4.
[35] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," 2019, *arXiv:1901.00137*.
[36] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1–6.
[37] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Commun. Mobile Comput.*, vol. 2, no. 5, pp. 483–502, Sep. 2002.
[38] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," 2009, *arXiv:0907.4728*.
[39] C. Dhahri and T. Ohtsuki, "Q-learning cell selection for femtocell networks: Single- and multi-user case," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 4975–4980.
[40] Z. Li, S. Guo, W. Li, S. Lu, D. Chen, and V. C. M. Leung, "A particle swarm optimization algorithm for resource allocation in femtocell networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Paris, France, Apr. 2012, pp. 1212–1217.

Yang Cao (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China (UESTC), China, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include machine learning techniques and radio access networks.

Shao-Yu Lien (Member, IEEE) received the Ph.D. degree from National Taiwan University in 2011. After the military service, he joined National Taiwan University and the Massachusetts Institute of Technology (MIT) as a Post-Doctoral Researcher. He was with National Formosa University as an Assistant Professor and an Associate Professor from 2013 to 2017. He is currently with National Chung Cheng University as an Associate Professor. He has been the Technical Director of the Smart System Institute, Institute for Information Industry (III), Taiwan, since 2020. Particularly, he has been a 3GPP Standardization Delegate, since 2009, for LTE, LTE-A, LTE Pro, and 5G NR, and in this role he has contributed more than 70 technical documents and patents in conjunction with HTC Corporation, III, the Industrial Technology Research Institute (ITRI), and Huawei. His research interests include configurable networks, cyber-physical systems, radio access networks, and robotic networks. He received a number of prestigious research recognitions, including the IEEE Tainan Section Best Young Professional Member Award 2019, the IEEE Communications Society Asia–Pacific Outstanding Paper Award 2014, the Scopus Young Researcher Award (issued by Elsevier) 2014, the URSI AP-RASC 2013 Young Scientist Award, and the IEEE ICC 2010 Best Paper Award. His research contributions on radio access congestion control has been officially included in EU FP7 Project "EXALTED," and three of his IEEE journal papers have been listed in ESI Highly Cited Paper. He was a Guest Editor of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING in 2019 and *Wireless Communications and Mobile Computing* (WCMC) in 2017. In the meantime, he also served as the Leading Organizer for a number of technical workshops in IEEE VTC-Spring 2015, IEEE GLOBECOM 2015, Qshine 2015 and 2016, IEEE PIMRC 2017, IEEE GLOBECOM 2019, IEEE ICC 2020, and GLOBECOM 2020.

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.

Ying-Chang Liang (Fellow, IEEE) was a Professor with The University of Sydney, Australia; a Principal Scientist and a Technical Advisor with the Institute for Infocomm Research, Singapore; and a Visiting Scholar with Stanford University, USA. He is currently a Professor with the University of Electronic Science and Technology of China, China, where he leads the Center for Intelligent Networking and Communications (CINC). His research interests include wireless networking and communications, cognitive radio, symbiotic communications, dynamic spectrum access, the Internet-of-Things, artificial intelligence, and machine learning techniques.

Dr. Liang is a Foreign Member of the Academia Europaea. He has been recognized by Thomson Reuters (now Clarivate Analytics) as a Highly Cited Researcher since 2014. He received the prestigious Engineering Achievement Award from The Institution of Engineers, Singapore, in 2007; the Outstanding Contribution Appreciation Award from the IEEE Standards Association in 2011; and the Recognition Award from the IEEE Communications Society Technical Committee on Cognitive Networks in 2018. He was a recipient of numerous paper awards, including the IEEE Jack Neubauer Memorial Award in 2014 and the IEEE Communications Society APB Outstanding Paper Award in 2012. He was the Chair of the IEEE Communications Society Technical Committee on Cognitive Networks and served as the TPC Chair and the Executive Co-Chair for the IEEE GLOBECOM 2017. He is also the Founding Editor-in-Chief of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS: Cognitive Radio Series, and the Key Founder and currently the Editor-in-Chief of the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is also serving as an Associate Editor-in-Chief for *China Communications*. He was a Guest Editor/an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the *IEEE Signal Processing Magazine*, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORK. He was also an Associate Editor-in-Chief of the *Random Matrices: Theory and Applications* (World Scientific). He was a Distinguished Lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society.

Kwang-Cheng Chen (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, USA, in 1987 and 1989, respectively. From 1987 to 1998, he worked with SSE, COMSAT, IBM Thomas J. Watson Research Center, and National Tsing Hua University. From 1998 to 2016, he was a Distinguished Professor with National Taiwan University and also served as the Director of the Graduate Institute of Communication Engineering and the Communication Research Center, and the Associate Dean for academic affairs with the College of Electrical Engineering and Computer Science from 2009 to 2015. Since 2016, he has been a Professor of electrical engineering with the University of South Florida, Tampa, FL, USA. His recent research interests include wireless networks, multi-robot systems and machine learning, quantum information systems and cybersecurity, and social networks. He has received a number of awards, including the 2011 IEEE COMSOC WTC Recognition Award, the 2014 IEEE Jack Neubauer Memorial Award, and the 2014 IEEE COMSOC AP Outstanding Paper Award. He has been actively involving in the organization of various IEEE conferences as the general/TPC chair/co-chair, and has served in editorships with a few IEEE journals. He also actively participates in and has contributed essential technology to various IEEE 802, Bluetooth, LTE and LTE-A, 5G-NR, and ITU-T FG ML5G wireless standards.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is currently an University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, the Internet of Things, 5G and beyond, and vehicular networks. He is a registered Professional Engineer of ON, Canada; an Engineering Institute of Canada Fellow; a Canadian Academy of Engineering Fellow; a Royal Society of Canada Fellow; and a Chinese Academy of Engineering Foreign Member. He received the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R. A. Fessenden Award in 2019 from IEEE, Canada, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, the Education Award in 2017 from the IEEE Communications Society (ComSoc), and the Joseph LoCicero Award in 2015, and the Technical Recognition Award from the Wireless Communications Technical Committee in 2019 and the AHSN Technical Committee in 2013. He has also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He served as the Technical Program Committee Chair/Co-Chair for IEEE GLOBECOM 2016, IEEE INFOCOM 2014, IEEE VTC 2010 Fall, and IEEE GLOBECOM 2007, and the Chair for the IEEE ComSoc Technical Committee on Wireless Communications. He is the President Elect of the IEEE ComSoc. He was the Vice President for Technical and Educational Activities, the Vice President for Publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of IEEE Fellow Selection Committee of the ComSoc. He served as the Editor-in-Chief for the IEEE INTERNET OF THINGS JOURNAL, *IEEE Network*, and *IET Communications*. He is also a Distinguished Lecturer of the IEEE Vehicular Technology Society and the Communications Society.

Authorized licensed use limited to: University of Waterloo. Downloaded on July 14,2022 at 02:05:31 UTC from IEEE Xplore. Restrictions apply.