Smart Multi-RAT Access Based on Multiagent Reinforcement Learning

Mu Yan, Gang Feng, *Senior Member, IEEE*, Jianhong Zhou, *Member, IEEE*, and Shuang Qin, *Member, IEEE*

***Abstract*—The ongoing increasing traffic in the era of big data yields unprecedented demands in user experience and network capacity expansion. The users of next generation mobile networks (5 G) should be able to use 3GPP, IEEE, and other technologies simultaneously. The integration of multiple radio access technologies (RATs) of licensed or unlicensed bands has been widely deemed as a cost-efficient way to greatly increase the network capacity. In this paper, we propose a smart aggregated RAT access (SARA) strategy with aim of maximizing the long-term network throughput while meeting diverse traffic quality of service (QoS) requirements. We consider the scenario that users with different QoS requirements access to a heterogeneous network with coexisting cellular-WiFi. In order to maximize system throughput while meeting diverse traffic QoS requirements in such a complex and dynamic environment, we exploit multiagent reinforcement learning to perform RAT selection in conjunction with resource allocation for individual user access requests, through sensing dynamic channel states and traffic QoS requirements. In SARA, we first use Nash Q-learning to provide a set of feasible RAT selection strategies while decreasing the strategy space in learning process, and then employ Monte Carlo tree search (MCTS) based Q-learning to perform resource allocation. Numerical results reveal that the network throughput can be maximized while meeting various traffic QoS requirements with limited number of searches by using our proposed SARA algorithm. For bulk arrival access requests, a suboptimal solution can be obtained as high computational complexity is incurred for achieving global optimality. Another attractive feature of SARA is that a tradeoff between the solution optimality and learning time can be readily made by terminating the search of MCTS according to the time constraint. Compared with traditional WiFi offloading schemes, SARA can significantly improve network throughput while guaranteeing traffic QoS requirements.**

***Index Terms*—Licensed band, unlicensed band, access control, reinforcement learning, Monte-Carlo tree search.**

I. Introduction

ACCORDING to the Cisco Visual Networking Index [1], mobile traffic has occupied a large portion of the big datasets. In order to meet the ever-increasing demands of mobile traffic for next generation wireless networks (5G), the network capacity as well as user access efficiency should be greatly improved.

To a certain extent, the capacity of cellular networks has been significantly increased by shrinking the cell size [2], which is known as network densification. Specifically, densely deployed small cell base stations (SBS) relieved the burden on overloaded macrocells, and lots of researches and developments have been made to efficiently offload traffic from macro cells to small cells [3], [4]. However, the large-scale deployment of small cell base stations is blocked because of severe co-channel interferences between small cells and adjacent base stations. With the rapid growth of network load and spectrum demand, the cellular networks will not be able to satisfy all the associated performance requirements [5]. As a complement and coexisting Radio Access Technology (RAT) of cellular network, the deployment of WiFi hotspot which operates in unlicensed band is now nearly ubiquitous. As WiFi shares some responsibilities of user data bearings for cellular network, it has been considered as an important candidate to provide extra spectrum resources for cellular networks [6]. All spectrum sharing by using multi-RATs becomes an inevitable solution for 5G and beyond [7].

Access control and resource scheduling in multi-RAT access are challenging due to the complexity in cross-domain management and diverse QoS requirements of user applications [8]. Considering the difficulty in optimal decision making for multi-RAT access, recently emerging Artificial Intelligence (AI) technology, such as machine learning, provides an effective tool to address the issue in a dynamic and complex environment. Machine learning gives computers the ability to learn without being explicitly programmed [9]. Among conventional machine learning mechanisms, Reinforcement Learning (RL) inspires agent to learn the variations and find potential solutions. By interacting with the environment, RL becomes a powerful tool for sequential decision making under uncertainty [10]. Recently, the AlphaGo developed by Google DeepMind performs excellently in Go competition by using MCTS based RL and deep learning algorithms [11], demonstrating the extreme power of AI in

Manuscript received July 11, 2017; revised November 4, 2017 and December 13, 2017; accepted January 9, 2018. Date of publication January 15, 2018; date of current version May 14, 2018. This work was supported in part by the National Science Foundation of China under Grant 61631005 and Grant 61471089, and in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2015Z005. The review of this paper was coordinated by Dr. K. Adachi. *(Corresponding author: Gang Feng.)*

M. Yan and S. Qin are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: 826103068@qq.com; blueqs@uestc.edu.cn).

G. Feng is with the National Key Laboratory of Science and Technology and the Center for Cyber Security, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: fenggang@uestc.edu.cn).

J. Zhou is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the School of Computer and Software Engineering, Xihua University, Chengdu 610039, China (e-mail: zhoujh@uestc.edu.cn)

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

Digital Object Identifier 10.1109/TVT.2018.2793186

solving sequential decision-making. In the underlying multi-RAT access scenario, the factors for decision making are dynamic, as the number of UEs requesting access, the QoS requirements of each User Equipment (UE) may vary over time, and especially the wireless channels are error-prone and time-varying. RL allows agent to timely sense the variations in network traffic, user demand etc, and thus smartly coordinate strategic behaviors of UEs and BSs.

In this paper, we propose a Smart Aggregated RAT Access (SARA) strategy based on multi-agent reinforcement learning to address the multi-RAT access problem, with aim to maximize network throughput while guaranteeing diverse UE traffic QoS requirements in HetNets. Specifically, we construct a Semi-Markov Decision Process (SMDP) based Hierarchical Decision Framework (HDF), which consists of RAT/Channel Selection Process (RSP) and Resource Allocation Process (RAP). Considering that there are multiple agents making decisions and the prior transition probability between network states is usually unavailable, we use the model-free multi-agent RL [12] to solve this SMDP problem. In RSP, since the decision-making process for the sub-flows of a specific UE constitutes a simultaneous game, we use Nash Q-learning to solve this problem [13], and obtain a set of feasible access strategies which can avoid access collisions and out-of-order of sub-flows. In RAP, we employ Monte Carlo Tree Search (MCTS) based Q-learning algorithm to search the optimal strategy from the set of candidate strategies to maximize system throughput while meeting the QoS requirements. The strategies searched by MCTS keep improving with learning process [14], and the optimal solution can be eventually accomplished in finite search steps. Moreover, the learning search can be terminated at any time to obtain a "up-to-now" best strategy, so as to meet the timeliness of a specific application.

The remainder of the paper is organized as follows. We present related work in Section II. The system model of LTE-WiFi radio level aggregation is presented in Section III. In Section IV, we formulate the multi-RAT aggregation problem as optimization problems from network and UE perspective respectively, and analyze the computational complexity. Then we construct a Semi-Markov Decision Process (SMDP) model in Section V, and use multi-agent reinforcement learning algorithm to solve the SMDP problem in Section VI. In Section VII we present the numerical results as well as discussions, and finally conclude the paper in Section VIII.

II. Related Work

In recent years, many researchers have focused on efficiently utilizing unlicensed band in HetNets to increase cellular network capacity. In this section, we first brief introduce the basic mechanisms proposed by 3GPP on the combination of licensed and unlicensed bands, and then review related work on RAT selection and associated channel allocation problems, especially those based on machine learning and MDP.

1) *Combination mechanisms of licensed and unlicensed bands* include the mechanisms developed within standardization bodies and industry: LTE in Unlicensed spectrum (LTE-U), Licensed-Assisted Access (LAA), LTE-WLAN radio level Aggregation (LWA), etc. [15]. LAA and LTE-U are proposed for providing carrier-grade wireless service in the 5 Ghz unlicensed band, where the unlicensed carrier is used as a secondary component carrier in the LTE carrier aggregation framework. The advantages of improved link performance, medium access control, mobility management and excellent coverage have made LTE a better RAT when compared with 802.11 WiFi in unlicensed band. However, LAA performs poorly if WiFi always occupies the unlicensed band as the Listen-Before-Talk (LBT) scheme is used. Moreover, in LTE-U, LBT is not employed, resulting in unfairness in its co-existence with WiFi. Thus we can see that in LAA and LTE-U, the interference between LTE and WiFi is inevitable and non-ignorable. LWA is a data aggregation scheme at Radio Access Network (RAN), in which packets are scheduled to be served on both LTE and WiFi radio links. The scheduling decisions for each link can be made at packet time level based on real-time channel conditions. The advantage of LWA is that it can provide better control and resource utilization on both links. Moreover, the Multi-RAT access control proposed in this work is based on the LWA mechanism.
2) *Traffic offloading through RAT of WiFi* is usually modeled as an optimization problem with aim to efficiently offload traffic from cellular, and optimally allocate channel resources to UEs from the two bands. The authors of [16] consider several utilization mechanisms of unlicensed bands and propose an optimization model with the object of maximizing the throughput of small cell users while keeping the interference from small cells to macrocells below a predefined threshold. Besides, the authors of [17] propose a dynamic switching and aggregation scheme of licensed and unlicensed bands in small cells, which is also formulated as an optimization problem. However, the algorithms proposed in [16] and [17] do not seem to have a good adaptability in a complex and time-varying environment, as the authors ignore various types of UE traffic which could request for services at any time. It is necessary to design a more flexible and fair resource allocation algorithm for traffic flows with different QoS requirements to improve long-term system throughput.
3) *Reinforcement Learning and Semi-Markov Decision Process* have been applied to solving RAT selection and resource allocation problems recently. The key driver for networks is to maximize the network throughput while satisfying user preferences. The authors of [18] proposed a network-assisted RAT access approach, where the BSs signal global network information to the mobiles users, so as to assist users to make decisions depending not only on their own preferences, but also on the overall performance of the network. In some cases, there exists a resource allocation game among the users as they are competing for limited network resources. In [19], the authors model the joint channel allocation and power control problem for D2D users as two games which are well solved by using

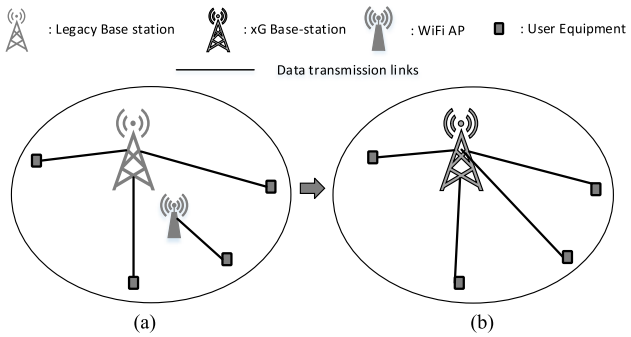


Fig. 1. Illustration of UE access mode in HetNets. (a) Legacy network accessHetNets with SBS and WiFi; (b) xG network access: Integrated WiFi AP and SBS.

multi-agent Q-learning strategy. In [18] and [20], network selection problems are modeled as an SMDP which is solved by using Q-learning algorithm. The results show that network performance can be significantly improved and RL performs well in a stochastic environment. However, the action space for each agent is substantially large, and thus a long search time is needed to obtain the global optimization results. Therefore, it is essential to reduce the computational complexity when using the RL algorithm, especially in a time-varying dynamic scenario.

III. System Model of LTE-WiFi Radio Level Aggregation

A. Scenario Description

We focus on the scenario of coexistence of Cellular Small Base Stations (SBS) and WiFi Access Points (AP), as shown in Fig. 1. We assume that the neighboring SBSs and WiFi APs use fixed and orthogonal channel resources to avoid interferences. UEs are equipped with both Cellular and WiFi interfaces which can request data from either RAT [21]. As shown in Fig. 1(a), the Legacy HetNet is used for comparison purpose, where SBS has no cooperation with WiFi AP, which implies that the data transmission is requested from either Cellular or WiFi interface. In Fig. 1(b), the next Generation Network Base Station (xGBS) aggregates the cellular SBS and WiFi AP by a standardized LTE-WLAN connection (standardized by 3GPP in Rel-13) [22]. We would like to mention that our proposed modeling and solution in this paper are not limited to LTE Rel-13. We use sub-channel of LTE in our model due to LTE is the 4G standard cellular RAT. Indeed, our proposed learning based multi-RAT access control mechanism can be applicable to general WiFi and cellular RAT scenario. xGBS behaves as an anchor node for both data and control planes, and it is connected to the core network via regular S1 interfaces. The authors of [6] propose a new entity alien access gateway (AAGW) which is deemed as a part of the small cells or WiFi segments towards integrated LTE-WiFi. Moreover, AAGW is available to be implemented in the 3GPP architecture without any other changes. In this paper, we assume this external entity is employed in LTE-WiFi aggregated scenario and the similar architecture is used.

Due to the randomness of UEs' behaviors in mobility and data request, the access to the network is indeed a time-varying dynamic scenario. We divide the time into frames for performing RAT selection and resource allocation decisions, which is defined as "Decision Time Frame (DTF)". In order not to violate the QoS requirements, the length of DTF is set to be less than the minimum latency tolerance of traffic flows. If the arrival flows are all not latency sensitive services, the length of DTF could be set longer. In other words, the length of DTF could be varying in different decision-making process, with the aim to guarantee the QoS requirements while improving channel resources utilization. Without loss of generality, we suppose that within a DTF, the network topology can be deemed as static. In this paper, BS is enabled to obtain the global network information including UEs services' QoS requirements and the channel load information. The integration of multiple RATs has to be done intelligently under the unified management of the xGBS, with aim of allocating resources from both RATs to users flexibly while taking into account the preferences of both the UEs and network operator.

B. Channel Access Model

On licensed band of LTE, medium access is performed by using Orthogonal Frequency Division Multiple Access (OFDMA) on the downlink. The spectrum on licensed band is divided into time-frequency radio blocks (RBs) which are referred to as sub-channels in the rest of this paper for convenience. Besides, we assume the transmission power is uniformly allocated to each sub-channel on the licensed band. On unlicensed band, we consider RAT of IEEE WiFi 802.11. Because of the Listen Before Talk (LBT) mechanism that used in IEEE WiFi 802.11, WiFi can only serve one terminal or aggregated MAC protocol data unit (A-MPDU) [23] to transmit data at a time, and thus UEs share the same WiFi band by way of TDM. To protect ongoing sessions in the dynamic network environment, an admission control policy is applied. In more details, in order not to compromise the QoS of ongoing sessions, new arrival flows need to wait until there are spare channel resources for transmission. We assume that there are λ LTE sub-channels and 1 WiFi channel in a XGBS in this paper. Flows are allowed to be split into multiple sub-flows which can use $\gamma(1 \leqslant \gamma \leqslant \lambda + 1)$ discontinuous (spectrum) channels for transmission to meet their QoS requirements. Furthermore, in the multi-RAT access control scheme, UEs are allowed to use multiple RATs for transmission in multiple DTFs via the dynamic orchestration of the scheduling algorithm.

IV. Problem Formulation and Analysis

In the multi-RAT access problem, the network operator pursues maximizing system throughput, while the UEs tend to maximize their transmission rate and satisfy their QoS requirements. Hence we formulate the optimal multi-RATs access problem from the perspective of the network operator and UEs respectively.

From the perspective of network operator, the network throughput can be improved by increasing the sum of the

downlink throughput over licensed and unlicensed bands for all UEs in every DTF. Consider m RATs which have $N_1, \cdots, N_m$ channels (or sub-channels), respectively. Each channel $j \in N_i$ (i means the ith RAT) has a bandwidth B_{ij} to provides U_{ij} throughput. The problem is to choose exactly one channel from each RAT such that the throughput is maximized subject to the bandwidth capacity $\hat{B}_i$. We can formulate the long-term system average throughput U maximization problem in the kth DTF as

$$\max \qquad U(k) = \sum_{i=1}^{m} \sum_{j \in N_i} U_{ij}(k), \tag{1}$$

$$\text{s.t.} \qquad \sum_{i=1}^{m} \sum_{j \in N_i} B_{ij} c_{ij}(k) \leq \hat{B}_i(k), \tag{1.1}$$

$$\sum_{i=1}^{m} \sum_{j \in N_i} c_{ij}(k) \geq 1, \tag{1.2}$$

$$c_{ij}(k) \in \{0,1\}, i = \{1, \cdots, m\}, j \in N_i, \tag{1.3}$$

where c_{ij} is a binary variable representing that if the jth channel of the ith RAT is used in the transmission. The value of m is fixed and equals to two due to that we consider RATs of LTE and WiFi which operate on licensed and unlicensed bands respectively. Constraint (1.1) states that the allocated bandwidth cannot exceed the spare bandwidth that can be provided by licensed and unlicensed bands respectively. Constraint (1.2) states that at least one channel (or sub-channel) would be selected to serve a traffic flow.

On the other hand, from the perspective of UEs, multi-RAT access problem is indeed a resource allocation problem, where a potential game exists among individual UEs, as the network resources become scarce. Since the network environment changes over time, a dynamic and flexible resource allocation mechanism is essential. Thus the optimal multi-RAT access problem from UE perspective could be formulated as maximizing the UE's average transmission rate with guaranteed QoS constraints:

$$\max \qquad \frac{1}{\mathcal{K}_n} \sum_{k \in [1, \mathcal{K}_n]} R_n(k), \tag{2}$$

$$\text{s.t.} \quad \mathcal{Q}_n \geqslant \mathcal{Q}_n \text{th}, \forall n \in [1, N], \tag{2.1}$$

where N is the number of UEs in the current DTF, $\mathcal{K}_n$ is the total number of DTFs for a specific application of UE n in the system operation, and $R_n(k) = (1 - L_n(k))S_n(k)/T_n(k)$ is the transmission rate for a specific application of UE n in the kth DTF where $S_n(k)$ is the size of flow, $T_n(k)$ is the flow's transmission time (including the queuing delay), and $L_n(k)$ is the packet loss probability. We use $H_n(k)$ to represent the set of LTE sub-channels which are chosen by the sub-flows of a flow in the kth DTF, where $H_n(k)$ equals to a null set means that the LTE sub-channels are not involved in the transmission process for a flow in the kth DTF. Besides, we use a binary value $\omega(k) = \{0, 1\}$ to indicate whether the WiFi channel is involved in the transmission process in the kth DTF or not. Therefore, $S_n(k)$ can be represented as $\sum_{h \in H_n(k)} s_h^L + \omega(k) \cdot s^W$, where s_h^L and s^W are the size of sub-flows which are allocated to LTE sub-channel h and the WiFi channel, respectively. The "smartness" of our scheduling mechanism is reflected in that the RAT access modes (using which LTE sub-channels or WiFi channel) of UEs are determined dynamically in different DTFs according to the time-varying environment of network. In constraint (2.1), $\mathcal{Q}_n$ is the QoS function mapping between the allocated resources and the expected performance, and $\mathcal{Q}_n$th is the threshold of UE n's QoS requirement.

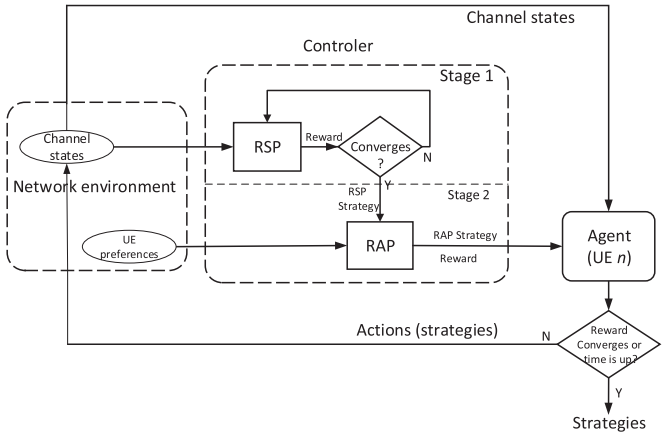


Fig. 2. SMDP based hierarchical decision framework (HDF).

We find that the multi-RAT access control problem which is the combination of problem (1) and (2) can be mapped to a Multiple Choice dimension Knapsack Problem (MCKP). MCKP is computationally harder than knapsack. Even for the case of dimensions equal to two, the problem does not have a $fully\ polynomial-time\ approximation scheme$ [24], therefore, our target problem is also NP-hard.

Considering the network dynamics, UEs movement, varying channel states and diverse QoS requirement, etc., it is very challenging to develop an optimal solution to this multi-RAT access control problem of maximizing long-term average system throughput subject to individual UE's QoS requirements. As this is a stochastic optimization problem with dynamic input, we employ a Semi-Markov Decision Process (SMDP) with multi-agent reinforcement learning approach to solve it in the subsequent sections.

V. Semi-Markov Decision Process for Multi-RAT Access Control

A. SMDP Based Hierarchical Decision Framework

As RAT selection and resource allocation in multi-RAT access can be deemed as two sequential processes, we propose an SMDP based hierarchical decision framework (HDF) to solve the joint RAT selection and resource allocation problem. Fig. 2 shows the block diagram of HDF. In HDF, once a decision (including the RAT selection and corresponding resources allocation) is made by an agent (UE), network states may change. The updated states and the reward derived from the evaluation function are feedback to the controller. This trial-and-go process ends until the reward converges or time is over, which drives the decisions towards the direction that the network expects.

Specifically, the controller employs a two stage decision process in the HDF, including RAT/Channel Selection Process (RSP) and Resource Allocation Process (RAP) for fulfilling multi-RAT access. The RSP game of Stage 1 is executed in the inner loop of HDF, where information of channel states is necessary to be signaled to controller (BS) for making decision. In RSP game, we try to reduce the feasible action space for each user by taking into account the constraint of (3), in order to expedite the convergence. After the Nash Q-learning in RSP game converges, a number of sets of feasible RAT/channels are obtained and output to RAP game of Stage 2 for another round of learning process. In RAP game of Stage 2, which is executed at the external loop of HDF, each agent employs P&V MCTS algorithm to search the resource allocation strategy with aim of maximizing its throughput, subject to the constraints on bandwidth resource and QoS requirements. In RAP, UE preferences and channel states information are also exploited. In the following, we first define network states, actions, state transition and then elaborate the evaluation functions of the SMDP model.

We define the state of channel i in kth DTF as vector $\boldsymbol{c}_i(k) = (\mu_i(k), N_i(k), l_i(k))$, where $\mu_i(k)$ is the instantaneous service rate of channel i at the kth DTF. Note that the value of $\mu_i(k)$ reflects the effect of channel states, including path-loss, fading. $N_i(k)$ is the number of sub-flows dwelling in channel i, and $l_i(k) = \sum_{j=1}^{N_i(k)} z_{i,j}$ represents the load of channel i, where $z_{i,j}$ is the size of sub-flow of flow j in channel i yet to be served. In addition, we let the overall channel state in kth DTF be denoted by $C(k) = (\boldsymbol{c}_1(k), \boldsymbol{c}_2(k), \cdots, \boldsymbol{c}_{N_c}(k))$.

An action in SMDP represents a strategy adopted by the controller at a specific network state. As channels are controlled by the xGBS, and the radio conditions are assumed unchanged within a DTF, at the beginning of a DTF, the queuing delay of a new arrival flow at channel i can be expressed as $l_i(k)/\mu_i(k)$. Feasible channels for a flow are sorted with a non-decreasing order $(1, 2, ..., N_c)$ of the access queuing time. We define the action for flow n as $\mathcal{A}_n \triangleq (\gamma_n, \Lambda_n), \forall n \in [1, N]$, where γ_n represents the number of sub-flows that we split it into, and Λ_n represents the channel selection strategies set.

In the multi-RAT model, the channel state changes as the time goes on, as well as the arrival and departure of the traffic flow. We can obtain that the state of channel i changes from $\boldsymbol{c}_i(k) = \{\mu_i(k), N_i(k), \sum_{j=1}^{N_i(k)} z_{i,j}\}$ to $\boldsymbol{c}_i(k+1) = \{\mu_i(k+1), N_i(k+1), \sum_{j=1}^{N_i(k+1)} z_{i,j}\}$. Accordingly, the overall network state changes from $C(k)$ to $C(k+1)$.

Note that an RSP or RAP strategy determines a reward by using an evaluation function. The evaluation function reflects our design objectives. We elaborate the RSP and RAP evaluation functions in the following.

B. Evaluation Function of RSP

As the data flow of a UE is allowed to use multiple channels for transmission, access collisions may happen if the same sub-channel or WiFi channel is selected at the same time, causing access failure. Moreover, due to the different states of channels, sub-flows that are transmitted via different sub-channels may arrive at the terminal out of order. We consider a flow-level packet reordering process, it is obvious that the frequent re-ordering of sub-flows at the terminal will cause non-negligible time delay. Our designed multi-RAT access control algorithm needs to cope with these problems. As mentioned before, in the process of RSP, a data flow could be split into several sub-flows. Let sub-flows $(1, 2, \ldots, \gamma_n)$ of UE n be defined as agents in our model which make decisions in RSP for selecting RAT and channels simultaneously. We define $\pi_n = (a_1, ..., a_{\gamma_n})$ as the joint channel selection strategies of these sub-flows, where a_j represents the sequence number of the channel which is selected by agent j. We use functions $coll_j(s, \pi_n)$ and $dis_j(s, \pi_n)$ to evaluate the strategy set π_n for agent j at state s, with aim of avoiding collisions and disorder respectively. Then we can define the RSP evaluation function, denoted by $RSP_j(s, \pi_n)$, for agent j as

$$RSP_j(s, \pi_n) = coll_j(s, \pi_n) + dis_j(s, \pi_n), \tag{3}$$

where

$$coll_j(s, \pi_n) = \begin{cases} -\sigma, & \text{if collision} \\ 0, & \text{otherwise} \end{cases},$$

$$dis_j(s, \pi_n) = \begin{cases} -\delta, & \text{if disorder} \\ 0, & \text{otherwise} \end{cases},$$

in which σ and δ are two constants which are greater than zero. On the one hand, for $coll_j(s, \pi_n)$, we assume that agents j and k adopt channel selection strategies a_j and a_k respectively. Then a collision happens and the access fails if $a_j = a_k$, and we feed back $-\sigma$ to agent j as the evaluation value of this cooperation strategy. On the other hand, for $dis_j(s, \pi_n)$, as we sort these channels by the queuing time each time before selecting channels, if agents j and $l (j < l)$ adopt actions a_j and $a_l (a_j > a_l)$ respectively, the arrivals of sub-flows may be out of order, incurring some expense for resequencing at the terminal and we thus feed back $-\delta$ to agent j as the evaluation value of this cooperation strategy.

C. Evaluation Function of RAP

As our design objective is to maximize the long-term network throughput with constraints of user QoS requirements, we design the evaluation function for RAP in Stage 2 with aim of driving the agent to select a strategy towards what the network expects. Let us begin with examining the instantaneous transmission rate of flow n in a specific DTF. We define $\hat{T}_n$ as the transmission delay of flow n, thus S_n/T_n in (2) can be represented as $(S_n/\hat{T}_n) \cdot (\hat{T}_n/T_n)$. We use B_n to denote $S_n/\hat{T}_n$, which is the bandwidth required by flow n. Furthermore, We take natural logarithm on both sides of (2) as

$$\ln R_n = \ln B_n + ln(\hat{T}_n/T_n) + \ln(1 - L_n). \tag{4}$$

We can see that the logarithm of UE transmission rate is translated into three different metrics. Next, we analyze these three metrics, to find their relationship with the QoS requirements on bandwidth, latency and packet loss rate of services, respectively.

In a DTF, for a specific traffic flow of UE n, LTE sub-channels or the WiFi channel could be selected for data transmission.

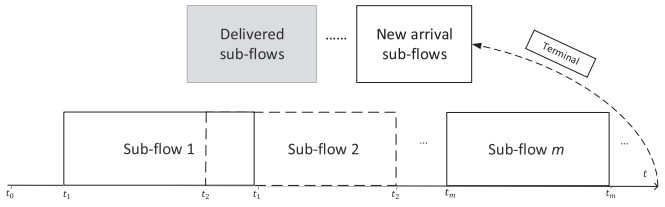


Fig. 3. Transmission process of arrival UE sub-flows on LTE sub-channels.

When LTE sub-channels are adopted, let H_n be the set of LTE sub-channels which are chosen by sub-flows of flow n, and the number of elements in H_n is the number of sub-flows of flow n, denoted by γ_n. Besides, we let r_h and t_h be the transmission rate and transmission delay on sub-channel h respectively, and $\hat{T}_n$ can be represented as $\max_{h\in H} t_h$. Obviously, the maximum transmission bandwidth B_n can be obtained if we let the transmission delay on each selected sub-channel be the same, and employ more sub-channels for transmission. When WiFi RAT is chosen for data transmission, we would like to mention that the involvement of LTE is not efficient. As the bandwidth of WiFi channel is much greater than that of LTE sub-channels, the system transmission capacity can only be increased slightly by using additional LTE sub-channels when WiFi is adopted. In addition, this LTE+WiFi transmission mode in one short DTF will greatly increase the complexity of the channel scheduling decision as the action space is significantly increased. We thus would compromise slightly on the system performance by using only one RAT in a DTF for significantly decreasing the complexity of our learning algorithm. On the other hand, in multiple DTFs, both RATs could be adopted in turns in the data transmission process by observing the network state, with aim to efficiently utilize the network resources.

Once a transmission mode is selected, $\hat{T}_n$ is obtained and fixed. We suppose that the xGBS can request data from the nearby MBS which has a large memory to store the entire content directory [25]. In this case, for flow n, the latency we consider is mainly composed of channel transmission delay and queuing delay. Then we analyze the relationship between $\hat{T}_n/T_n$ and the packet average time delay. Fig. 3 describes the transmission process of arrival UE sub-flows on the selected LTE sub-channels, where we define t'_0 as the beginning time of request, and t_m and t'_m $(1 \leqslant m \leqslant \gamma_n)$ as the beginning and the ending of the transmission time for sub-flow m respectively. Let $d_m = t_m - t'_{m-1}$ be the time interval between two sequential sub-flows. Thus the sum of the intervals between two sub-flows is denoted by $D_n = \sum_{m=1}^{\gamma_n} d_m$. Therefore, it is obvious that T_n increases with the value of D_n. Furthermore, when analyzing the average time delay and the packet loss rate, the results in the past DTFs will be taken into consideration to get a long-term performance optimization. In past $\mathcal{K}$ DTFs, we define $x_n(k)$ as the number of packets of flow n in the kth past DTF, and thus the long-term average time delay $\mathcal{T}_n$ is given by $\frac{\sum_{k=0}^{\mathcal{K}} T_n(k)}{\sum_{k=0}^{\mathcal{K}} x_n(k)}$. We can see that an available long-term average time delay can be obtained by orchestrating time intervals between two successive sub-flows in this current DTF.

We employ a congestion control mechanism in the data transmission process of SARA. As packets should be queued for transmission in congestion, the packets of a sub-flow could be blocked and even be dropped due to that the queuing delay exceeds the threshold time, denoted by η. We define that in the kth past DTF, the number of packets which are dropped due to that their queuing time exceeds η are $\rho_i(k)$. We can obtain the long-term packet loss rate $\mathcal{L}_n$ as $\frac{\sum_{k=0}^{\mathcal{K}} \rho_n(k)}{\sum_{k=0}^{\mathcal{K}} x_n(k)}$, and we use $\mathcal{P}_n = 1 - \mathcal{L}_n$ to denote the successful rate of packets transmission.

We use $\mathcal{B}_n$th, $\mathcal{T}_n$th, and $\mathcal{P}_n$th to denote the QoS requirements on bandwidth, packet average time delay and the packet loss rate which are signaled by flow n to the xGBS. To guarantee the flow's QoS requirements, we use a step function as

$$\varepsilon(t) = \begin{cases} 1, \text{ if } t \geqslant 0 \\ \epsilon, \text{ otherwise} \end{cases}$$

in the evaluation function design, where ϵ is close to 0^+ and let $B_n^* = \varepsilon(B_n - \mathcal{B}_n\text{th}) \cdot B_n, T_n^* = \varepsilon(\mathcal{T}_n\text{th} - \mathcal{T}_n) \cdot \hat{T}_n/T_n$, and $P_n^* = \varepsilon(\mathcal{P}_n - \mathcal{P}_n\text{th}) \cdot (1 - L_n)$. Thus the evaluation function for RAP is defined as

$$RAP_n(s,a) = e^{w_B \ln(B_n^*) + w_T \ln(T_n^*) + w_P \ln(P_n^*)}, \qquad (5)$$

where w_B, w_T, and w_P are involved in as weight coefficients to optimize the long-term sum of discounted reward we expect, to give designated weightage to individual QoS metrics. Note that the method of setting weight coefficients is to maximize the distinction of weightage between the preference metrics and others. This can roughly guide the algorithm towards the direction of the QoS metrics with larger weight coefficients. Moreover, if the weight coefficients are removed (or let $w_B = w_T = w_P$), the strategy output by SARA will be improved towards maximizing the system average throughput. By using evaluation function (5) in RAP, it is feasible to dynamically orchestrate the resource allocation strategies among the flows with constraint of the basic QoS requirements.

VI. Multi-Agent Reinforcement Learning for SARA

Given network state s and evaluation function $r(s,a)$, we employ reinforcement learning (RL) to solve the optimal multi-RAT access problem in a DTF. As it is not easy to obtain the transition probability $p(s,a)$ between network states, we resort to a model-free Q-learning algorithm for its appropriateness [26]. Our RSP and RAP are modeled under a Multi-Agent System (MAS) which is defined as a collection of agents that observe and act in a same environment. In our model, making decisions for a given traffic flow takes into account the decisions of other traffic flows in the same DTF. In MAS, we can observe the previous actions of agents and the current states, and the immediate rewards after each agent chooses their actions. The multi-RAT access decision process of multiple sub-flows in the MAS can be modeled as an n-agent stochastic game $\mathcal{G}$ [13] which is defined as:

Definition 1: An n-agent stochastic game $\mathcal{G}$ is a tuple $\langle S, A^1, \cdots, A^n, r^1, \cdots, r^n, p\rangle$, where S is the state space, A^i is the action space of agent $i (i = 1, \cdots, n)$, $r^i : S \times A^1 \times \cdots \times A^n \to R$ is the evaluation function for agent i, $p : S \times A^1 \times \cdots \times A^n \to \triangle(S)$ is the transition probability

map, where $\triangle(S)$ is the set of probability distributions over state space S.

Based on the proposed n-agent stochastic game $\mathcal{G}$, in more details, in the HDF as shown in Fig. 2, there exist a simultaneous game and a sequential game among the agents in RSP and RAP respectively. Based on the SMDP model proposed in Section V, we resort to Nash Q-learning Algorithm (NQA) and Policy and Value MCTS algorithm (P&V-MCTS) [11], variants of preliminary Q-learning algorithm, to solve these two games respectively.

A. Preliminary: Q-Learning

Q-learning algorithm uses transition experience of the form $\langle s, a, r, s' \rangle$ (from state s, action a resulted in reward r and next state s') to improve an estimate $\hat{Q}$ of the optimal state-action value function Q^*. Q-value is defined as the expected long-term discounted reward of state s, when policy set π is adopted. The estimate $\hat{Q}$ is given by

$$\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha(R + \beta^k \max_{a' \in A} \hat{Q}(s',a')) - \hat{Q}(s,a)),$$

where $\beta \in [0,1]$ is the discount factor, R is the instant rewards received when the agent adopts action a, and α is the learning rate $(0 < \alpha < 1)$, which determines to what extent the learned Q-value will update the old one. For example, when $\alpha = 0$, the network does not learn, while when $\alpha = 1$, the network considers only the recent Q-value. It has been proven in [27] that if the learning rate is decayed at an appropriate rate, $\hat{Q}(s,a)$ eventually converges to $\hat{Q}^*(s,a)$. In Q-learning algorithm, our objective is to find an optimal policy by iteration $\pi^*(s) = \text{argmax}_{a \in A(s)} Q(s,a), \forall s, \pi$.

B. Nash Q-Learnaing Algorithm for RSP

In Nash Q-learning algorithm, the agents at the same decision epoch attempt to learn their equilibrium Q-values, starting from an arbitrary guess. To this end, an agent maintains the Q-values of other agents, which are used to update its own Q-values by taking the best response actions in each state to other agents' actions. Eventually nash equilibrium (NE) is reached, where no agent is willing to alert their strategy anymore.

Definition 2: In stochastic game, a Nash equilibrium point is a tuple of N strategies $(\pi_*^1, \cdots, \pi_*^N)$ such that for all $s \in S$, $\pi^n \in \sqcap^n$ and $n = 1, \cdots, N$, $v^i(s, \pi_*^1, \cdots, \pi_*^N) \geqslant v^n(s, \pi_*^1, \cdots, \pi_*^{(n-1)}, \pi^n, \pi_*^{(n+1)}, \cdots, \pi_*^N)$, where $\sqcap^n$ is the set of strategies available to agent n.

For an n-agent system, the Q-function for any agent becomes $Q(s, a^1, \cdots, a^n)$, rather than $Q(s,a)$. Given the extended notion of Q-function, and NE as a solution concept, we define a Nash Q-value as the expected sum of discounted rewards when all agents follow specified NE strategies[13].

Definition 3: Agent $n's$ Nash Q-function is defined over $(s, a^1, \cdots, a^N)$, as the sum of agent $n's$ current reward plus its future rewards when all agents follow a joint equilibrium strategy, i.e., $Q_*^n(s, a^1, \cdots, a^N) = r^i(s, a^1, \cdots, a^N) + \beta \sum p(s'|s, a^1, \cdots, a^N) v^n(s', \pi_*^1, \cdots \pi_*^N)$, where $(\pi_*^1, \cdots, \pi_*^N)$ is the joint Nash equilibrium strategy and $r^n(s, a^1, \cdots, a^N)$ is the one-period reward of agent i in state s under joint action $(a^1, \cdots, a^N)$. $v^n(s', \pi_*^1, \cdots, \pi_*^N)$ is the total discounted reward of agent n over infinite periods starting from state s given that agents follow the equilibrium strategies.

Algorithm 1: Nash Q-learning Algorithm for RSP.

Input: Network state S; Action space $\mathcal{A}(s)$; reward $RSP(s,\pi)$ in (3); learning rate α; discount value β;
Output: Feasible RAT/channel selection strategies set $\prod_{n \in N}^{I}$.

```
∀s ∈ S, ∀n ∈ N : Q(s, a_n) ← 0, a_n ∈ A(s);
for n = 1, 2, ..., N do
    for k = 1, 2, ..., K do
        while Nash equilibrium described in definition 2 is not achieved do
            choose action (a^1, ..., a^{γ_n}) at random;
            Observe state s and s' ← s^{k+1};
            Q_n^{k+1}(s, a^1, ..., a^{γ_n}) ← (1 − α)Q_n^k(s, a^1, ...a^{γ_n}) + α[RSP_n^k(s, a^1, ..., a^{γ_n}) + β max_{π^1,...,π^{γ_n}} Q(s', π^1, ..., π^{γ_n})];
        end while
    end for
end for
```

In this paper, the process of RSP is limited to stationary strategies, where in a DTF, the number of strategies and agents are finite. It is proved by Fink (1964) that there exists at least one NE point in stationary strategies for n-player discounted stochastic game processes [13]. In order to simulate a realistic scenario, the network states are first randomly initialized. The Q-values for agents are first set to zero. At state s, the network randomly selects action set $(a^1, a^2, ..., a^{\gamma_n})$, where γ_n is the number of sub-flows of UE n. When network state changes to s', Q-value of sub-flows of UE n is updated according to the Nash Q-function. The learning process ends until that Nash equilibrium among the γ_n agents is achieved, and a set of feasible RAT/channel selection strategies are output to each agent. Then the network executes another round of learning process for next UE. We elaborate the Nash Q-learning for RSP in Algorithm 1.

C. Policy & Value MCTS Algorithm (P&V-MCTS) for RAP

In RAP, the sequential game among multiple agents is solved by employing P&V-MCTS algorithm. The adopted actions by each agent are subsets of the feasible strategies $\prod_{n \in N}^{I}$ output from the NQA in RSP. P&V-MCTS combines the policy and value network in an MCTS algorithm. For a specific UE, the policy network is used to sample actions, and the value network is used to evaluate the selected strategies. We suppose that in the Monte-Carlo search tree, each node s contains edges (s,a) of the tree for all actions, and each edge stores a set of statistic parameters as $\{r(s,a), N(s,a), Q(s,a)\}$, where $r(s,a)$ is the instant reward, $N(s,a)$ is the visit count and $Q(s,a)$ is the action-value comes from the value network, which is updated by Q-learning algorithm. Specifically, P&V-MCTS consists of three strategic steps, namely strategy selection, simulation and backpropagation, as shown in Fig. 4.

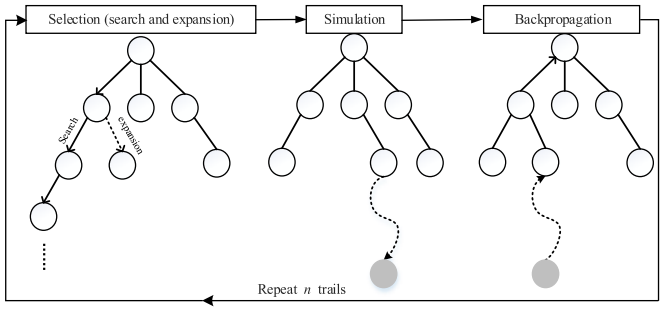


Fig. 4. Strategic steps of MCTS.

Selection: At each step, for a UE, a strategy set of LTE sub-channels or WiFi channel is selected according to an action selection strategy $a = \text{argmax}_a(Q(s,a) + u(s,a))$, by using a variant of the upper confidence intervals in trees (UCT) algorithm [28], where $u(s,a) = c \cdot P(s,a)\frac{\sqrt{\sum_{a \in \mathcal{A}(s)} N(s,a)}}{1+N(s,a)}$ is a bonus that encourages exploration and exploitation, in which c determines the level of exploration, $P(s,a)$ is the prior probability. In our policy network, we let the WiFi channel be selected with a high prior probability, and other LTE sub-channels are chosen with equal probability. The UCT algorithm leverages the upper confidence bond (UCB) algorithm to intelligently allocate search resources at each point in the tree. This search strategy initially prefers the actions with high Q-value and low visit count, but gradually prefers the actions with high action value.

Simulation: In the sequential game of RAP, agents make decisions in sequence according to the arrival time in a DTF. A simulation begins at the root (the first agent) of the search tree and finishes at a leaf node (the last agent). We define the simulation as a trial that tries to search the best strategy before the estimated Q-value converges to the maximum Q-value or time is up.

Backpropagation: When a simulation reaches a leaf node, a separate backward pass is initiated to update the traversed edges' statistic parameters, where the $Q(s,a)$ of each node is updated by Q-learning algorithm and the visit count of each visited node is increased by one.

The search process can be terminated at any time and output a set of current best strategies, in order to satisfy real-time requirements as in [11]. In other words, the learning time or the number of trials should be dynamically adjusted, thus to avoid violation of UEs' QoS requirements.

Algorithm 2 describes the P&V-MCTS algorithm for RAP, which aims at finding desirable resource allocation strategies for each traffic flow. As summarized in Algorithm 2, the Q-value and the number of visits on each node are first initialed to zero. The network state is randomly initialized. Agent n selects an action based on the UCT algorithm. When network state changes to s', Q-value of agent n is updated according to the preliminary Q-function, and the number of visits on the visited nodes increases by one. The learning process ends until the algorithm converges or the time is up, and then the corresponding resource allocation strategies are output to each agent.

Algorithm 2: P&V-MCTS algorithm for RAP.

Input: Network state S; Action space $\prod^I$; reward $RAP(s,a)$ in (5); learning rate α; discount value β; QoS requirement $\mathcal{Q}_n$th and size S_n of traffic flow $n \in N$;

Output: Desirable resource allocation strategies set $\prod_{n\in N}^{II}$;

1: $\forall s \in S, \forall n \in N, \forall a_n \in \prod_{n\in N}^{I}$: $N(s,a_n) \leftarrow 0$;
2: **for** $k = 1, 2, ..., \mathcal{K}$ **do**
3: $\quad \forall s \in S, \forall n \in N, \forall a_n \in \prod_{n}^{I}$: $Q(s,a_n) \leftarrow 0$;
4: $\quad$ **for** $n = 1, 2, ..., N$ **do**
5: $\quad\quad a_n \leftarrow \text{argmax}_{a_n}(Q(s,a_n) + u(s,a_n))$;
6: $\quad$ **end for**
7: $\quad Q^{k+1}(s,a) \leftarrow (1-\alpha)Q^k(s,a) + \alpha[RAP^k(s,a) + \beta \text{max}_{a'}Q(s',a')]$;
$\quad\quad N(s,a) \leftarrow N(s,a) + 1$;
8: **end for**

D. Computational Complexity of SARA

We now analyze the computational complexity of SARA. As SARA algorithm mainly consists of a Nash Q-learning algorithm and a P&V-MCTS algorithm, we analyze the complexity of the two algorithms separately. For the Nash Q-learning algorithm, we let $|S|$ be the number of states, and $|A|$ be the number of an agent's actions. We can easily obtain that the total number of trials is $\gamma_n|S| \cdot |A|^{\gamma_n}$. As aforementioned, γ_n is the number of agents (sub-flows) of flow n. We can see that the computational complexity of Nash Q-learning for flow n is a linear function of $|S|$, a polynomial function of $|A|$, but an exponential function of the number of agents.

Next, we discuss the computational complexity of the P&V MCTS algorithm, which combines the evaluation function and the UCT algorithm in the reinforcement learning process. The authors of [10] proposed a three-dimensional cone which compares four algorithms of Full tree search, Evaluation-function methods, MCTS methods and the P&V MCTS algorithm, respectively. The height and diameter of the cone represent the search depth and width respectively, and the top of the cone represents the optimal solution for the current state. Intuitively, the area of full tree search is the largest among these methods, which is too large to be examined. The search width of MCTS is narrower than that of the Full tree search, due to that MCTS use UCT algorithm for searching subset of actions that are worth considering. Moreover, as Evaluation-function method can approximate the value received at the end of a game, its visited area is expanded into a fixed depth. In this work, as explicit expression of evaluation functions are obtained by our designed SMDP, the P&V MCTS algorithm search continuously until the end of the game, directed by the evaluation functions and UCT algorithm. The combination of the evaluation functions and the UCT algorithm can provide a quantum leap in performance in the P&V MCTS algorithm, resulting in that the search area of the P&V MCTS algorithm is significantly decreased compared with the other three algorithms. In Section VII we examine the number of trials (or search times) of the P&V-MCTS in some cases by simulations.

TABLE I
PARAMETERS USED IN SIMULATIONS

Description/Parameters	LTE	WiFi
Number of UEs	Less than 8	
Number of BS(AP)	1	
Noise level	-106 dbm	
Maximum Tx Power	23 dbm	
Average data volume per flow	3 Mb	
Available spectrum per BS(AP)	3 Mhz	20 Mhz
Average DL PHY data rate	15 Mbps	35 Mbps
γ (number of sub-channels)	15	
β (discount factor)	0.99	
c (level of exploration)	2	
σ	100	
δ	100	
η	1200 ms	

TABLE II
TYPICAL SERVICES WITH CERTAIN QOS REQUIREMENTS

Example services	Bandwidth Requirement ($\mathcal{B}_{th}$)	Packet Loss Rate ($\mathcal{L}_{th}$)	Latency ($\mathcal{T}_{th}$)
Conversational voice	Low	10^{-2}	200 ms
Real time gaming	Medium	10^{-3}	50 ms
Video (buffered)	High	10^{-6}	300 ms
Video (live)	Very high	10^{-3}	100 ms
IMS signaling	Low	10^{-6}	100 ms
...	...	...	...

VII. PERFORMANCE EVALUATION

In this section, we validate the effectiveness of our proposed SARA algorithm by using simulation experiments. We consider a two-tier HetNet which consists of one LTE SBS (such as Femto), one WiFi AP and varying number of UEs. We assume that both the WiFi AP and LTE SBS are situated at the same place. We ignore the inter-cell interferences by using orthogonal radio channels. We consider that there are less than 8 UEs requesting for data transmission via WiFi and/or LTE Femto within a DTF, and we assume that the decision-making sequence of traffic flows are based on their sequence of the requesting time. Let the channel rate of WiFi be 35 Mbps (50% of which is for uplink), which is lower than the highest physical layer data rate of 72 Mbps for single-antenna 802.11n. Other system parameters used are similar to those used in related work [6], [16]. The system parameters in the performance evaluation are summarized in Table I.

The 3GPP has defined a series of services with different QoS requirements [29], and we list some typical services with certain QoS requirements in Table II, which are adopted in our simulations.

A. Comparison References

We use the following on-demand channel resources scheduling algorithms as comparison references in our performance evaluation: (i) Proportional Fair Scheduling for Multiple Transmission Systems (PFSMTS) [30]: PFSMTS assigns users to multiple carriers (e.g., sub-channels in LTE) with aim of maximizing the sum of logarithmic average user rates. It is widely used as a comparison reference in related work [31], [32]. (ii) LTE Assisted Algorithm (LAA): LTE band works as an assisted band that traffic flows are preferentially offloaded to the WiFi band. If WiFi is occupied currently, LTE sub-channels with the best channel quality will be selected for the data transmission. (iii) Online Learning Algorithm (OLA): SARA becomes an on-line learning algorithm when we set the DTF very small, within which UEs always choose the channels with the maximum reward for transmission, which can be deemed as a greedy algorithm.

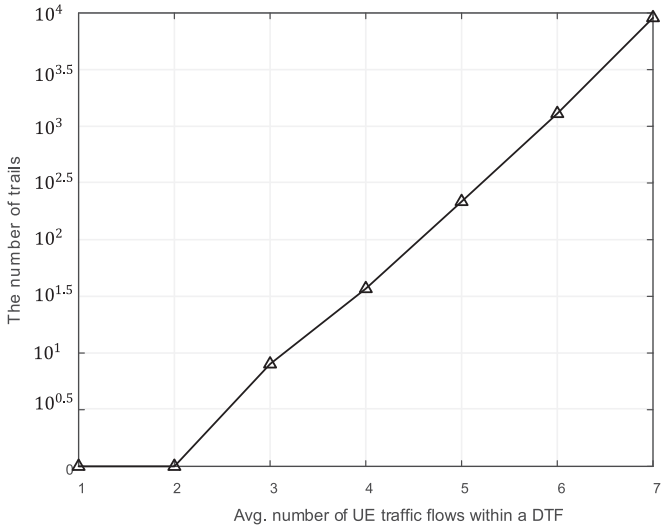
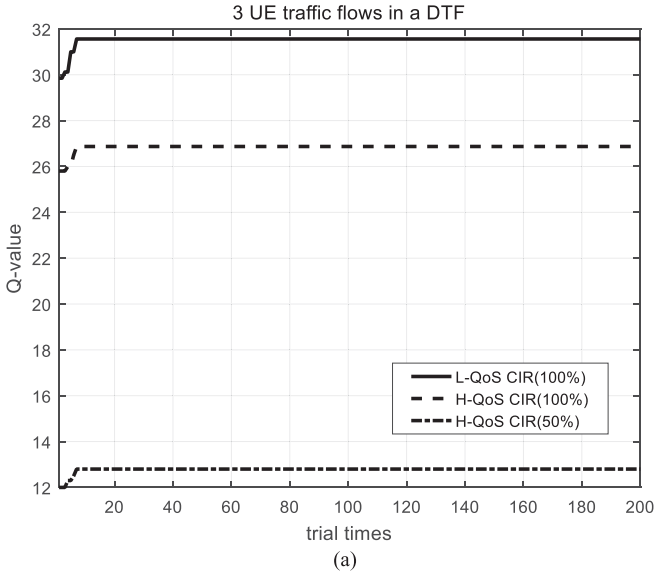


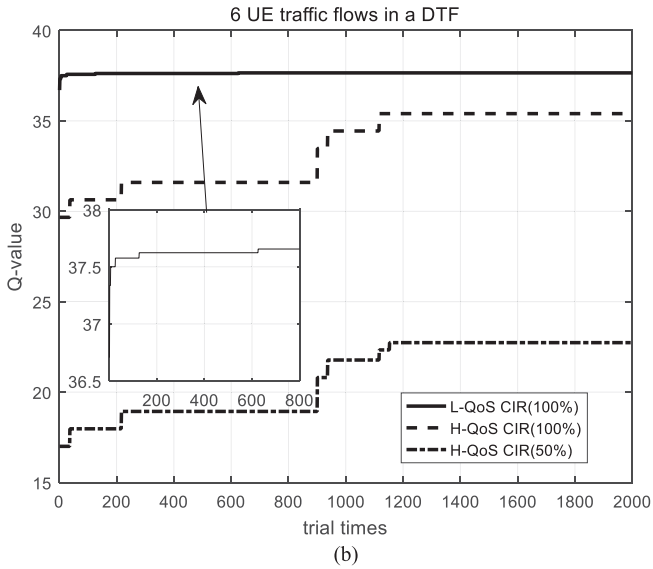
Fig. 5. The number of trials needed by SARA to get the best strategy.

B. Numerical Results and Discussions

In the first experiment, we examine the number of trials needed for SARA to reach the best strategy (maximum Q-value), with varying number of traffic flows. In the simulation, single thread processing is used in MCTS searching. At the beginning of SARA, the channel load states and the QoS types of traffic flows are randomly generated. Fig. 5 shows the curve of the number of trails which changes with the number of flows to be scheduled within a DTF. From Fig. 6 we can observe that when the number of UE traffic flows is small, the number of trials needed to reach the best strategy increases slowly. But it increases in an exponential fashion when the number of flows becomes larger, say 5. This is because that the breadth of the search tree increases sharply with the search depth. This implies that the corresponding computational complexity involved in MCTS search to obtain the best strategy increases rapidly with the number of flows, and may become intractable when there are large number of traffic flows. We consider two types of traffic flows: delay sensitive and delay tolerant sensitive flows. For delay sensitive traffic flows, a short DTF is set to accommodate a few number of flows (e.g., less than 5 as shown in Fig. 5) for the fast convergence of algorithm. On the other hand, for delay-tolerant sensitive traffic flows, more flows are allowed within a longer DTF, and a longer learning process is allowed for pursuing a higher performance gain. Besides, the search process in MCTS can be terminated at arbitrary time, outputting



(a)



(b)

Fig. 6. The impact of the mean number of UE traffic flows and CIR on Q-value.

a "current best" (sub-optimal) solution, in case that a global optimal result cannot be obtained in time.

Next, we examine the influence of QoS requirement level of flows and Channel Idle Ratio (CIR) on the Q-value and the number of trials needed. We define that a traffic flow has a high QoS requirement level (H-QoS) with at least one of the QoS parameters satisfying $\mathcal{B}_{th} \geqslant 1$ mbps, $\mathcal{T}_{th} \leqslant 200$ ms, and $\mathcal{L}_{th} \leqslant 10^{-5}$; otherwise it has a low QoS requirement level (L-QoS). We define the Channel Idle Ratio (CIR) as the ratio of idle channels over all channels at the beginning of an experiment, e.g., $CIR = 50\%$ means that there are approximately 50% channels used by on-going flows at the beginning. Let there be 3 and 6 traffic flows respectively in the simulation experiments. We conduct 6 experiments with parameter configurations as shown in Fig. 6 for comparison. Fig. 6 shows the obtained Q-value in the learning process with varying number of trails. Note that a higher Q-value in SARA corresponds a better system resources

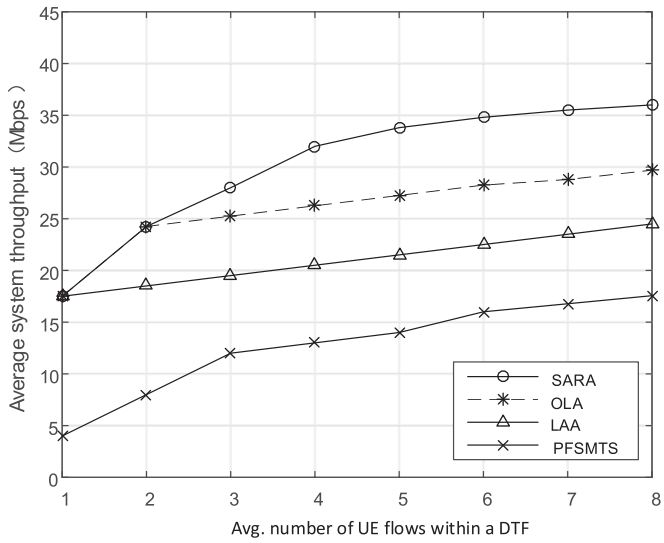


Fig. 7. Avg. system throughput of SARA, OLA, LAA and PFSMTS.

utilization. It is shown in Fig. 6 that CIR value affects the Q-value. Specifically, the Q-value with $CIR = 100\%$ is greater than that with $CIR = 50\%$. This is because that when the CIR is smaller, there may be no sufficient channel resources for arriving UE flows, resulting in a lower Q-value. For the same CIR at the beginning of experiment, we find that the maximum Q-value of H-QoS flows is lower than that of the L-QoS flows. This is because that the flows with H-QoS could compromise some performance on the throughput for guaranteeing their QoS requirements. Besides, as shown in Fig. 6(b), when there are 6 traffic flows, we can see that the number of trials needed for H-QoS flows is greater than that for L-QoS flows. This is because that more search trails are needed for SARA to find the strategies that satisfy flows' higher QoS requirements.

Next, we compare the long-term system average throughput of SARA with that of OLA, LAA, and PFSMTS algorithms with varying number of flows. In this experiment, we let the weight coefficients $w_B = w_T = w_P = 1/3$ such that the average system throughput totally increases with the Q-value. Fig. 7 shows the system average throughput with varying number of flows within a DTF. From Fig. 7, we can see that average system throughput of SARA is significantly higher than that of all other the other three scheduling algorithms for non-trivial number (say $\geqslant 3$) of UE traffic flows. Specifically, the improvement is approximately $12.3\% \sim 20\%$, $47.3\% \sim 61.9\%$, $138.4\% \sim 142.9\%$ when compared with OLA, LAA, and PFSMTS respectively. This implies that the smart integration of WiFi and LTE for data transmission by exploiting in SARA can significantly improve the system resource utilization in a dynamic network environment. Besides, we can see that the average system throughput of all four algorithms increases with the number of traffic flows. This is as expected as more channel resources are used with more transmission requests in a DTF, until the network becomes saturation.

In this simulation, we compare the fairness of SARA and PFSMTS. As referring to [30], we can see from the scheduling equation that PFSMTS can schedule multiple UEs by using

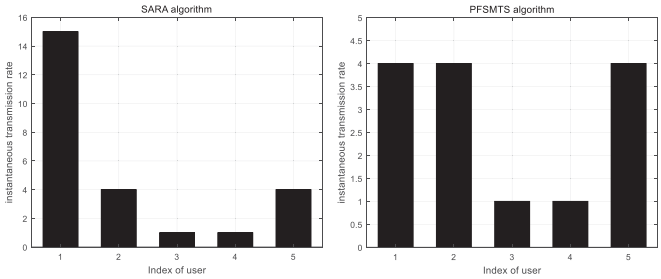


Fig. 8. Resources allocation of SARA and PFSMTS for different users.

multiple carriers (sub-channels in the OFDMA system). The SARA algorithm we propose in this paper poses the same nature of ideas that it can schedule multiple UEs by using multiple channels of LTE and WiFi. In PFSMTS and the proposed SARA, the allocation fairness is reflected in that resources are reasonably allocated to each user based on their requirements [33]. We simulate SARA and PFSMTS for examining the transmission rate. In this simulation, users 1, 2 and 5 are supposed to have high data-rate requirement, while users 3 and 4 have low data-rate requirement. Fig. 8 shows the transmission rate for different users of SARA and PFSMTS respectively. We can see that the number of channel resources that allocated to each user is based on their data-rate requirements. Moreover, the authors of [30] have proved PFSMTS's scheduling fairness for each UE. PFSMTS aims at achieving the long-term fairness, which is the same as the traditional PF algorithm. In SARA, it is expected that the short-term resource utilization is considerably improved, since multiple users can share the air interface resources at the same time based on the design of SARA. We have that the scheduling fairness of SARA and PFSMTS are guaranteed in different time scales.

Finally, we examine how the weight coefficients in (5) affect strategy selection, and thus the corresponding sum of the UE transmission rate (SUR) of SARA. We use the following weight coefficients configurations in the experiment: (i) $w_P = w_R = w_T = 1/3$; (ii) $w_P = w_R = 1/5$ and $w_T = 3/5$. At the beginning of the experiment, the channel load and the QoS types of traffic flows are randomly set. Fig. 9 shows the sum of UE transmission rate as a function of the number of flows within a DTF when weight coefficients changes. As shown in Fig. 9, we find that configuration (i) can always result in the maximum SUR. In comparison, significantly lower average throughput is achieved by using configuration (ii) when there are only 1 or 2 flows, while the maximum SUR is achieved when there are more than 2 flows. This is because that the SUR could be compromised for improving QoS with larger weight coefficients (e.g., latency). However, when the number of flows increases, the channel resources may become exhausted that there are no better strategy available for pursuing a higher QoS than threshold, and thus the resulted strategies can merely satisfy the minimum QoS requirements (threshold). In this case, the maximum SUR can be achieved. This result also implies that the evaluation function directs the output strategy toward improving QoS with larger weight coefficients (UE's QoS preference), when there are sufficient channel resources.

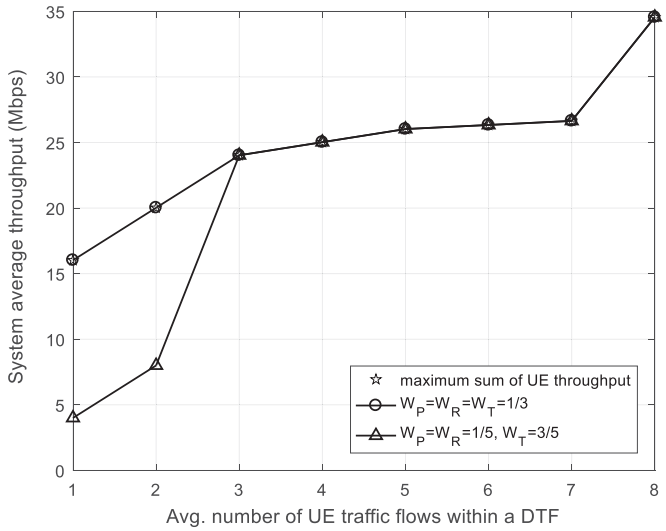


Fig. 9. Sum of transmission rate for various weight coefficient configurations.

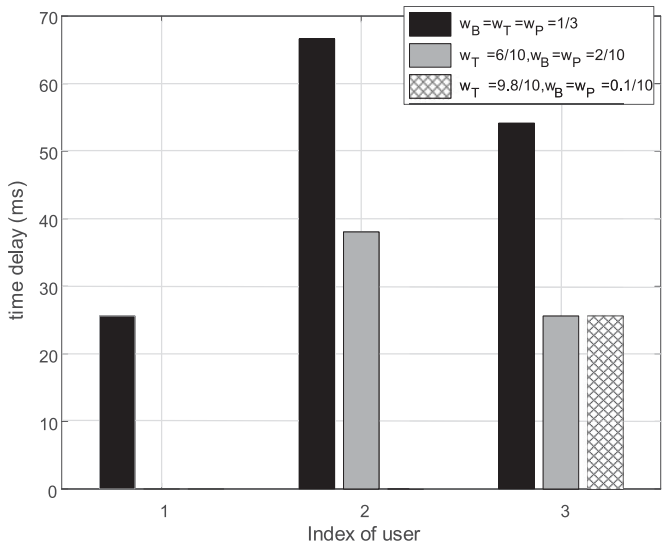


Fig. 10. The results of time delay that affected by different coefficients.

In more detail, to obtain a deeper understanding for the relationship between the performance and weight coefficients, we have conducted some additional experiments. The simulation settings remain the same as those of the previous simulations. We add one more weight configuration iii) $w_B = w_P = 0.1/10$ and $w_T = 9.8/10$ in this simulation. Fig. 10 shows the time delay of three UEs with different weightages. We can see that usually the greater a weight coefficient we set, the more the corresponding metric is improved. However, for the third UE, a greater coefficient does not result in a better result on time delay. This is because that there could be no more remaining resources or no better channel resource allocation to the user.

In our designed RAP reward function (5), for traffic flows with specific QoS requirements, the weight coefficients are set to be the same so that the system throughput improvement is the maximization objective while guaranteeing user QoS requirements. However, for traffic flows which have a very stringent QoS requirements (e.g., delay-sensitive traffic), we may set greater

TABLE III
RUNNING TIME OF ALGORITHMS

Number of UEs	Running time of one trial of SARA (or Online Learning)	LAA	PFSMTS
1	2.4 ms	3.1 ms	1.3 ms
2	2.9 ms		
3	3.3 ms		
4	3.9 ms		
5	4.4 ms		
6	5.2 ms		
7	5.8 ms		
8	6.6 ms		

weight coefficients for corresponding QoS metrics, in order to guide the channel selection strategies towards improving the corresponding QoS metrics. In other words, if we set a greater weight coefficient, the resulted RAP strategy tends to satisfy corresponding QoS metric requirement. Note that as the channel resources are finite and dynamic during the transmission process, improving the weight coefficients may not bring better strategies.

C. Further Discussions on SARA Complexity

To examine the complexity of SARA, we list the running time for the algorithms in Table III. From the results, we can see that our algorithm has higher running time than other three scheduling algorithms LAA, PFSMTS and OLA. Please note that we can readily limit the number of UEs involved in RSP and RAP by selecting appropriate the DTF length (for example 50 ms). In this way, we may control the number of trials needed to achieve the optimal resource allocation strategy, while meeting the delay requirements. Moreover, we would highlight again that we can terminate the search process (MCTS) at any time to meet the latency constraints, with a current best solution. Furthermore, in the studied Multi-RAT access control problem, flow-level access decision is to be made. Usually we need tens to hundreds of iterations for convergence when the number of UEs ≤ 5. The timeliness can be satisfied in general application scenarios in using the MCTS based Q-learning algorithm, by controlling the DTF length. Intuitively, there exist a tradeoff between the performance gain and the number of UEs involved in RAP and RSP. We would like to mention that even in the extreme case where there is a single UE (i.e., SARA reduced to Online Learning), some performance gain can still be obtained, compared with LAA, PFSMTS algorithms, which can be observed in Fig. 7.

VIII. CONCLUSION

Multi-RAT is a very promising technology for future wireless networks. However, due to the diverse QoS requirements of user applications and the time-varying characteristics of network environment, cross-domain management for multi-RAT access becomes very challenging. In this paper, we have proposed a Smart Aggregated RAT Access (SARA) strategy based on multi-agent reinforcement learning to address the multi-RAT access problem. We have formulated the access issue as an SMDP model with a multi-agent system. In SARA, we have used the multi-agent learning method to address this SMDP problem, with the aim of maximizing the average system throughput while satisfying the UEs QoS preferences. We have carried out computer simulations to verify the effectiveness of our proposed SARA. Numerical results indicate that significant performance improvement can be achieved by employing our proposed SARA. From the results, we find that the global optimization can be achieved with reasonable number of searches in most cases by using our proposed SARA. SARA outperforms several existing scheduling mechanisms in terms of average system throughput and system resource utilization while satisfying the UEs QoS preferences.

REFERENCES

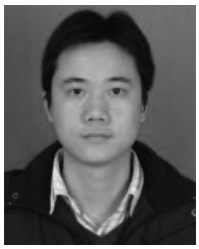
[1] T. J. Barnett, A. Sumits, S. Jain, and U. Andra, "Cisco visual networking index (VNI) update global mobile data traffic forecast," Cisco, San Jose, CA, USA, Tech. Rep., 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html
[2] D. Lopez-Perez *et al.*, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, Jun. 2011.
[3] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.
[4] Y. Li, H. Celebi, M. Daneshmand, C. Wang, and W. Zhao, "Energy-efficient femtocell networks: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 20, no. 6, pp. 99–105, Dec. 2013.
[5] Q. Li, G. Wu, and R. Q. Hu, "Analytical study on network spectrum efficiency of ultra dense networks," in *Proc. IEEE 24th Annu. Int. Symp. Personal, Indoor, Mobile Radio Commun.*, 2013, pp. 2764–2768.
[6] O. Galinina *et al.*, "5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1224–1240, Jun. 2015.
[7] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Comput. Netw.*, vol. 50, no. 13, pp. 2127–2159, 2006.
[8] B. Cao, F. He, Y. Li, C. Wang, and W. Lang, "Software defined virtual wireless network: Framework and challenges," *IEEE Netw.*, vol. 29, no. 4, pp. 6–12, Jul./Aug. 2015.
[9] P. Simon, *Too Big to Ignore: The Business Case for Big Data*. Hoboken, NJ, USA: Wiley, 2013.
[10] M. L. Littman, "Reinforcement learning improves behaviour from evaluative feedback." *Nature*, vol. 521, no. 7553, pp. 445–451, 2015.
[11] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 1, pp. 484–489, 2016.
[12] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman, "PAC model-free reinforcement learning," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 881–888.
[13] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, no. 6, pp. 1039–1069, 2003.
[14] R. Coulom, "Efficient selectivity and backup operators in Monte-Carlo tree search," in *Proc. 5th Int. Conf. Comput. Games*, 2007, vol. 4630, pp. 72–83.
[15] LTE-U Forum, "Coexistence Study for LTE-U SDL," 2015.
[16] F. Liu, E. Bala, E. Erkip, M. C. Beluri, and R. Yang, "Small-cell traffic balancing over licensed and unlicensed bands," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5850–5865, Dec. 2015.
[17] A. R. Elsherif, W. P. Chen, A. Ito, and Z. Ding, "Adaptive small cell access of licensed and unlicensed bands," in *Proc. IEEE Int. Conf. Commun.*, 2013, pp. 6327–6332.
[18] M. El Helou, M. Ibrahim, S. Lahoud, K. Khawam, D. Mezher, and B. Cousin, "A network-assisted approach for RAT selection in heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1055–1067, Jun. 2015.
[19] H. Tabrizi, G. Farhadi, and J. Cioffi, "Dynamic handoff decision in heterogeneous wireless systems: Q-learning approach," in *Proc. IEEE Int. Conf. Commun.*, 2012, pp. 3217–3222.

[20] S. Maghsudi and S. Stanczak, "Joint channel allocation and power control for underlay D2D transmission," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 2091–2096.
[21] J. A. Kong, *Progress in Electromagnetics Research*, vol. 1. Amsterdam, The Netherlands: Elsevier, 2012.
[22] B. Nelson, "4G americas LTE aggregation unlicensed spectrum white paper," Tech. Rep. Nov., 2015.
[23] E. Perahia and R. Stacey, *Next Generation Wireless LANs: 802.11 n and 802.11 ac*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
[24] A. Kulik and H. Shachnai, "There is no EPTAS for two-dimensional knapsack," *Inf. Process. Lett.*, vol. 110, no. 16, pp. 707–710, 2010.
[25] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1382–1393, May 2017.
[26] P. Dayan, "Technical note Q,-learning," *Mach. Learn.*, vol. 292, pp. 279–292, 1992.
[27] C. J. Watkins and P. Dayan, *Q-Learning*, vol. 8. New York, NY, USA: Springer, 1992.
[28] C. D. Rosin, "Multi-armed bandits with episode context," *Ann. Math. Artif. Intell.*, vol. 61, no. 3, pp. 203–230, 2011.
[29] W. Paper, "Quality of service (QoS) and policy management in mobile data networks," Tech. Rep. Dec. 2013.
[30] H. Kim, K. Kim, Y. Han, and S. Yun, "A proportional fair scheduling for multicarrier transmission systems," in *Proc. IEEE 60th Veh. Technol. Conf.*, 2004, vol. 1, pp. 409–413.
[31] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems—Part I: Chunk allocation," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2734–2744, Sep. 2009.
[32] C. Ú. Castellanos *et al.*, "Performance of uplink fractional power control in UTRAN LTE," in *Proc. IEEE Veh. Technol. Conf.*, 2008, pp. 2517–2521.
[33] C. Wengerter, J. Ohlhorst, and A. V. Elbwart, "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," in *Proc. 2005 IEEE 61st Veh. Technol. Conf.*, 2005, vol. 3, no. 2, pp. 1903–1907.

Mu Yan received the B.S. degree in electronic engineering from the Beijing Jiaotong University, Beijing, China, in 2014. He is currently working toward the Ph.D. degree at the National Key Lab of Science and Technology on Communications, University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests include next generation cellular networks, access control, multirate transmission, and machine learning, etc.

Gang Feng (M'01–SM'06) received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1986 and1989, respectively, and the Ph.D. degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 1998. He joined the School of Electric and Electronic Engineering, Nanyang Technological University in December 2000 as an Assistant Professor and was promoted as an Associate Professor in October 2005. He is currently a Professor with the National Laboratory of Communications, UESTC. He has extensive research experience and has published widely in computer networking and wireless networking research. His research interests include resource management in wireless networks, next generation cellular networks, etc.

Jianhong Zhou received the M.Eng. degree in electronics and electrical engineering from the Nanyang Technological University, Singapore in 2008, and the Ph.D. degree in computer software and theory from the University of Chinese Academy of Sciences, Beijing, China, in 2016. He is currently a Lecturer with the School of Computer and Software Engineering, Xihua University, Chengdu, China and a Postdoctor with the National Key Lab of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China, respectively. Her research interests include next generation cellular networks, low latency ultrareliable communication, etc.



Shuang Qin received the B.E. degree in electronic information science and technology and the Ph.D degree in communication and information system both from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2006 and 2012, respectively. He is currently an Associate Professor with the National Key Laboratory of Science and Technology on Communications, UESTC. His research interests include cooperative communication in wireless networks, data transmission in opportunistic networks, and green communication in heterogeneous networks.