This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

A Realization of Fog-RAN Slicing via Deep Reinforcement Learning

Hongyu Xiang, Shi Yan, *Member, IEEE*, and Mugen Peng, *Fellow, IEEE*

***Abstract*—To meet the wide range of 5G use cases in a cost-efficient way, network slicing has been advocated as a key enabler. Unlike the core network slicing in a virtualized environment, radio access network (RAN) slicing is still in its infancy and the corresponding realization is challenging. In this paper, we investigate the realization approach of fog RAN slicing, where two network slice instances for hotspot and vehicle-to-infrastructure scenarios are concerned and orchestrated. In particular, the framework for RAN slicing is formulated as an optimization problem of jointly tackling content caching and mode selection, in which the time-varying channel and unknown content popularity distribution are characterized. Due to the different users' demands and the limited resources, the complexity of original optimization problem is significant high, which makes traditional optimization approaches hard to be directly applied. To deal with this dilemma, a deep reinforcement learning algorithm is proposed, whose core idea is that the cloud server makes proper decisions on the content caching and mode selection to maximize the reward performance under the dynamical channel state and cache status. The simulation results demonstrate the performance in terms of hit ratio and sum transmit rate can be significantly improved by the proposal.**

***Index Terms*—fog radio access network, network slicing, deep reinforcement learning.**

I. Introduction

To meet emerging applications and services in fifth generation (5G) [1], network slicing has been advocated as a key technique by both academia and industry [2], [3]. In the concept of network slicing [4], the common physical network is sliced into multiple isolated networks automatically. Each specific network (i.e. network slice) is composed by 5G network functions and specific radio access technology settings. By orchestrating the network functions and providing a customized network, network slicing provides operators the needed networks on an as-a-service basis and satisfies stringent performance requirements of various service types. As one of the innovative techniques in future wireless networks, network slicing enables the 5G architecture flexible to accommodate diverse use cases in a cost-efficient way. Based on the service requirements and the network status, an end-to-end network slice instance is implemented and maintained [5], which usually includes the core network part and the radio access network part. In summary, network slicing is characterized with automation, isolation, customization, and programmability, which can be further referred in [6].

As an important part of network slicing, radio access network (RAN) slicing attracts the attention of many parties. Unlike the core network slicing based on the centrally positioned cloud infrastructure, the RAN slicing architecture can be either central or distributed. The RAN slicing in the ultra-dense networks is discussed in [7], [8]. In [7], the ultra-dense RAN is sliced into traditional communication slice and mobile-edge computing slice. Both slices are cooperated to improve the quality of computation experience for mobile devices. Inspired by the cost efficiency of network slicing, RAN slicing in ultra-dense networks to reduce the deployment and operational costs is investigated in [8]. Specially, a number of RAN runtime services are provided by constructing a novel multi-service RAN runtime environment, which supports slice orchestration and management, and flexible slice customization as per-tenant needs. The applications of RAN slicing in the cloud RAN are also studied. In [9], the cloud RAN infrastructure and resources are shared among multiple wireless operators. To maximize the profits of the network infrastructure provider, a two-stage resource allocation among multiple tenant slices for wireless operators is performed. Similarly, dynamic network slicing in multi-tenant heterogeneous cloud RANs is investigated in [10]. To design an efficient network slicing scheme, multiple factors like tenants' priority, baseband resources, fronthaul and backhaul capacities, quality of service, and interference are taken into consideration. A novel network slicing architecture featuring RAN abstraction is introduced in [11]. wherein a two-level MAC scheduler is utilized to abstract and share the physical resources among slices. Using the FlexRAN protocol to control and configure the eNodeB, the performance improvements of proposed network slicing architecture are validated via simulations results.

The aforementioned works typically concentrate on the realization in traditional RAN architectures. The prospect of network slicing in the 5G RAN should be jointly exploited. Unlike slicing in traditional RAN architectures, network slicing in the 5G RAN is distinct. Specially, the 5G RAN is characterized with multiple networking topologies and heterogeneous network functions [12]. Different topologies and network functions in the 5G RAN perform in a different way. Combining the network slicing with RAN architecture would be beneficial for the performance improvement and flexible

This paper was supported in part by the National Natural Science Foundation of China under No. 61921003, 61925101, 61831002, and 61901044, the State Major Science and Technology Special Project (Grant No. 2018ZX03001025 and 2018ZX033001023), the 03 special project and 5G program of Jiangxi Province (20192ABC03A041), the Beijing Natural Science Foundation under No. JQ18016, and the National Program for Special Support of Eminent Professionals. (Corresponding author: Mugen Peng, Shi Yan.)

H. Xiang (e-mail: xhyou@bupt.edu.cn), S. Yan (e-mail: yanshi01@bupt.edu.cn), and M. Peng (e-mail: pmg@bupt.edu.cn) are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

deployment. The fog RAN (F-RAN) [13], which incorporates the fog computing into the RAN, has been studied extensively as a revolutionary paradigm to tackle the various performance requirements in 5G networks. Based on the fog computing [14], computing, storage, control, and networking functions are distributed closer to users, like at fog access points (F-APs) and fog user equipments (F-UEs). Hence local signal processing, cooperative radio resource management, and distributed storing are available in the edge of F-RANs. According to different cache and channel conditions, UEs can not only access to F-UEs and F-APs for a decreased transmission delay, but also remote radio units (RRUs) for a cooperative resource management. With advanced edge caching and adaptive mode selection schemes, a high spectral efficiency and energy efficiency while maintaining a low latency level can be achieved [15]. Hence, a hierarchical architecture for network slicing in F-RANs is introduced in [16], which consists of a centralized orchestration layer and a slice instance layer to make RAN slicing adaptively implementable.

One of the key benefits of network slicing in F-RANs is the performance improvement of slice instances, by jointly allocating the caching and radio resource in F-RANs. With popular content cached in F-APs, duplicated content transmissions can be efficiently reduced, which further decreases the burdens on the constrained fronthaul and improves the performance of network slices. Extensive work under different scenarios has been done to prove the benefits. To improve the average requested content data rates, a hybrid content caching approach considering finite service latency is proposed in [17]. In the proposed approach, the content caching locations are optimized, which includes original content servers, central cloud units and base stations. Compared with the general popularity based caching algorithm and the independent caching algorithm, the hybrid content caching approach outperforms in terms of average end-to-end service latency and backhaul/fronthaul load reduction ratios. In [18], the tradeoff between the content diversity gain and the cooperative transmission gain is identified. Based on the stochastic geometry analysis, a probabilistic content placement scheme is proposed with the cache hit ratio and the rate coverage probability considered. Analysis and numerical results reveal that the proposed content placement outperforms the conventional caching schemes in terms of the average content delivery success probability. In [19], the joint cache placement and physical-layer transmission schemes is investigated. To improve the spectrum efficiency of content delivery, both centralized and distributed transmission-aware cache placement strategies are developed. Simulation results show that the proposed transmission aware caching algorithms can greatly reduce the users' average download delay due to the exploitation in caching and cooperation gains. In [20], the joint design of cloud and edge processing for the cache utilization improvement is studied and the delivery rate maximization problems for different fronthauling strategies are tackled. Numerical results show that there is a trade-off between fronthaul and caching resources. With a large cache, the impact of a small fronthaul capacity can be compensated, especially for a skewed popularity distribution.

Despite the advantages of network slicing in F-RANs, there remain technical challenges to be solved. The mechanism that efficiently exploits the utilization of caching and radio resource in a real wireless environment should be developed. Particularly, the content popularity distribution may change with spatio-temporal variation [21] and the channel is variant in the real wireless environment. To make the system model more realistic, the unknown-popularity feature of the contents and the time-varying nature of wireless channels should be addressed. In this paper, it is assumed that there is no knowledge of the content popularity distribution. The content popularity estimation and the resource allocation are considered jointly. Besides, instead of a block-fading channel, the channel is formulated as a finite-state Markov channel, which has been widely accepted to feature the correlation structure of the fading process [22]–[24]. Under these assumptions, we investigate the realization of RAN slicing in F-RANs where the content caching and mode selection are addressed. Selecting the transmission mode for the UE to download the desired content involves the content caching, which results in the coupling between the content caching and mode selection and makes the problem more challenging.

To deal with the aforementioned issues, the deep reinforcement learning (DRL) [25] is adopted, which is particularly advantageous in handling complicated control problems [26]–[28]. In [29], the virtualized network function (VNF) chain scaling problem in network slicing is investigated. Based on the current deployment of the VNF chain and the predicted flow rate, the DRL is utilized to produce VNF deployment decisions across a geographic span. The trace-driven simulation showed that the framework adapts quickly to traffic dynamics online and achieves lower system costs. To deal with the dynamic forwarding in slice-enabled industrial wireless networks, a latency-aware service orchestration based on deep Q-network (DQN) is designed in [30]. Through deep Q-learning, the fluctuation of link quality and node workload are aware, which enables timely forwarding adjustments. Simulation results demonstrate that the solution can not only provide scalable and flexible infrastructure for industrial wireless networks, but also ensure the low-latency delivery in network slices. To improve slices' performance in a edge computing enabled-RAN, the optimal computation offloading policy is investigated in [7]. The problem is formulated as a Markov decision process, which faces the curse of high dimensionality. To deal with the problem, a double DQN-based strategic computation offloading algorithm is proposed, which learns the optimal policy without knowing a priori knowledge of network dynamics. Numerical experiments show that the algorithms achieve a significant improvement in computation offloading performance compared with the baseline policies.

Besides applications in the network slicing, the DRL is also utilized in the wireless communication network optimization [31], like the joint allocation of caching and communication resources. The joint caching and interference alignment is studied in [32], wherein the mutual interferences in multi-user wireless networks are addressed. To deal with the interference under realistic time-varying channels, a DRL-based solution is proposed to select optimal interference alignment users, which achieves a significantly improved performance gain.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

To improve the performance of next generation vehicular networks, a dynamic orchestration of networking, caching, and computing resources is investigated in [33]. With the gains of not only networking but also caching and computing considered, the complexity of the formulated problem is very high. Therefore, a novel DRL approach is proposed, the effectiveness of which is validated via simulation results. To investigate the problem of proactive caching for multi-view 3D videos in 5G networks, a DRL-based joint views selection and local memory allocation scheme is proposed in [34]. Specially, a Markov decision model is constructed for the proactive caching problems. Since it is difficult to solve the problem with traditional dynamic programming, the DRL is adopted to find effective proactive caching policy. Simulation results show that the proposed solution effectively maintains high-quality user experience, especially for high-mobility 5G users moving among small cells. To tackle challenges with network dynamics, resource diversity, and the coupling of resource management with mode selection, a DRL-based joint mode selection and resource management approach is proposed in [35]. With a trained DQN to decide communication modes and computing resource allocation intelligently, the long-term system power consumption under dynamics of edge cache states is minimized. In [36], the joint communication, caching and computing design problem in the vehicular networks is investigated. To tackle the high complexity, a DRL with multi-timescale framework is developed. Compared with the ordinary DQN method, the presented approach has an improved stability and performance due to the introduction of double DQN and dueling DQN.

In summary, the existing works on the DRL applications have validated the benefits of DRL in handling complicated problems and improving network performance. This motivates us to utilize the DRL to realize the RAN slicing in F-RANs, which is formulated as an optimization problem of jointly tackling content caching and mode selection.

Our main contributions can be summarized as follows.

1) A realization approach of F-RAN slicing is investigated, wherein two examples of network slice instances for the hotspot scenario and the vehicle-to-infrastructure (V2I) scenario respectively are concerned and orchestrated. In particular, an optimization framework for the corresponding RAN slicing is presented, where the bit rate requirement of hotspot UEs and transmission delay requirement of V2I UEs are taken into account. Furthermore, the time-varying channel and unknown content popularity distribution are considered to make the system model more realistic.
2) To manage the complex problem and implement the presented optimization framework, a new solution based on the DRL is proposed. In the proposed solution, the cloud server decides the most appropriate content caching and mode selection with the constraint of fronthaul capacity and the F-AP capability. Followed with the content caching and mode selection, a sub-problem on the power allocation considering the inter-slice and intra-slice interference is derived according to the Perron-Frobenius theorem and proximal theory.
3) The proposed strategies are evaluated and the simulation results have validated the benefits of the proposed strategy: it can provide performance improvement with respect to the existing content caching and mode selection strategies, especially under the time-varying channel and unknown content popularity distribution. Simulation results also exhibit the performance under different network parameters, and show the influence of the fronthaul capacity and F-AP capability constraint on the performance.

The rest of this paper is organized as follows. Section II describes the system model and service scenarios. In Section III, the constraints that need to be taken into consideration are introduced and DRL-based solutions are proposed. The simulation results are provided in Section IV. A summarization on this paper and the future work are given in Section V.

II. System model of RAN slicing in F-RAN

As illustrated in Fig. 1, we present a system model of the RAN slice instances in the F-RAN. Suppose there are two kinds of single-antenna UEs: the hotspot UEs $\mathcal{K}_0$ and the V2I UEs $\mathcal{K}_1$. The K_0 hotspot UEs are crowded in a zone and demand high rates of the enhanced mobile broadband service, while the K_1 V2I UEs are moving at high speed and require delay guaranteed data transmission. To serve the hotspot UEs $\mathcal{K}_0$ and V2I UEs $\mathcal{K}_1$, the customized RAN slice instances are deployed by utilizing the F-RAN common infrastructures, and DQN trained in the cloud server will provide the corresponding management.

In the hotspot slice instance, M_0 RRUs with only radio functions are distributed around the UEs and connected to the cloud server via fronthaul. Since the capacity-constrained fronthaul limits the performance, M_1 F-APs with popular content cached are located in the targeted zones to alleviate the heavy burden. The F-APs are connected with the cloud server via Xn interface for a centralized content caching. The F-AP mode and RRU mode are also available for the V2I UEs. In the V2I slice instance, the F-AP can provide a much shorter transmission delay, since the delay in the fronthaul can be omitted due to the local service.

A. Communication model

Here we assume that V2I UEs and hotspot UEs are sharing the same resource block and the number of modes available is larger than the number of UEs, $M_1 + 1 >= K_0 + K_1$. Define UEs served by m-th F-AP at slot t as $\mathcal{S}_m(t)$ and UEs served by RRUs as $\mathcal{S}_0(t)$, we have $\mathcal{K}_0 \cup \mathcal{K}_1 = \cup_{m=0}^{M_1} \mathcal{S}_m(t)$. Denote the mode selection of UE k as $s_k(t) \in \{0, 1, 2, ..., M_1\}$, where $s_k(t) = 0$ means UE k is connected to all the RRUs, and $s_k(t) = m$ represents that UE k is connected to m-th F-AP. Once the mode selection of UE k has been given, namely $s_k(t) = m$, then UE k belongs to the UE set $\mathcal{S}_m$ and $k \in \mathcal{S}_m(t), m = 0, 1, 2, ..., M_1$.

Specially, suppose L_0 and L_1 antennas are equipped at the RRU and F-AP, respectively. The received signal at the k-th

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

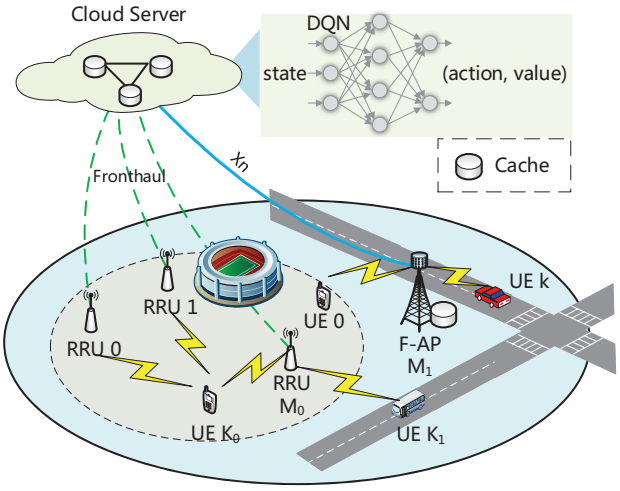

Fig. 1. A system model of RAN slice instances in F-RANs, wherein the DQN is utilized to make a decision to perform the content caching in F-APs and mode selection for UEs.

UE when $s_k(t) = 0$ can be written as

$$\begin{aligned} y_{0,k}(t) =& \sqrt{P_0(t)}\mathbf{h}_{0,k}(t)\mathbf{v}_{0,k}(t)x_k(t) \\ &+ \sqrt{P_0(t)} \sum_{j\in\mathcal{S}_0(t), j\neq k} \mathbf{h}_{0,k}(t)\mathbf{v}_{0,j}(t)x_j(t) \\ &+ \sum_{m=1}^{M_1} \sum_{j\in\mathcal{S}_m(t)} \sqrt{P_m(t)}\mathbf{h}_{m,k}(t)\mathbf{v}_{m,j}(t)x_j(t) \\ &+ z_k(t), \end{aligned} \tag{1}$$

where the first term at the right side represents the expected signal, the second term implies the interference from the UEs served by the RRUs, the third term means the inter interference from the F-APs, and the last term is the noise. $\mathbf{h}_{0,k}(t)$ and $\mathbf{v}_{0,k}(t)$ are the $1 \times M_0L_0$ vector of channel coefficients from the M_0 RRUs to the k-th UE and the $M_0L_0 \times 1$ unitary precoding vector of the k-th user over the time slot t, respectively. $\mathbf{h}_{m,k}(t)$ and $\mathbf{v}_{m,k}(t)$ are the $1 \times L_1$ vector of channel coefficients from the m-th F-AP to the k-th UE and the $L_1 \times 1$ unitary precoding vector of the k-th user over the time slot t, respectively. $x_k(t)$ and $z_k(t)$ are the transmitted signal and the additive white Gaussian noise with zero mean and σ^2 variance at the k-th receiver, respectively. P_0 and P_m are the transmission power of RRUs and m-th F-AP, respectively. Similarly, the received signal at the k-th UE when $s_k(t) = m$ is

$$\begin{aligned} y_{m,k}(t) =& \sqrt{P_m(t)}\mathbf{h}_{m,k}(t)\mathbf{v}_{m,k}(t)x_k(t) \\ &+ \sum_{j\in\mathcal{S}_m(t), j\neq k} \sqrt{P_m(t)}\mathbf{h}_{m,k}(t)\mathbf{v}_{m,j}(t)x_j(t) \\ &+ \sum_{m'=0, m'\neq m}^{M_1} \sum_{j\in\mathcal{S}_{m'}(t)} \sqrt{P'_m(t)}\mathbf{h}_{m',k}(t)\mathbf{v}_{m',j}(t)x_j(t) \\ &+ z_k(t), \end{aligned} \tag{2}$$

where the first term at the right side represents the expected signal, the second term implies the interference from the UEs served by the same F-AP, the third term means the inter interference from the RRUs and other F-APs, and the last term is the noise.

In this paper, the realistic time-varying channel in the network is modeled as first-order finite-state Markov channel (FSMC). Specifically, the channel coefficient $\|\mathbf{h}_{m,k}(t)\|^2$ is modeled as a Markov random variable with partitioned and quantized levels. Each level corresponds to a state of the Markov channel, the channel coefficient of which varies from one level to another level when one time slot elapses. The relation between the transformation can be modeled as

$$\mathbf{h}_{m,k}(t) = \rho\mathbf{h}_{m,k}(t-1) + \sqrt{1-\rho^2}\delta, \tag{3}$$

where δ has the same distribution with $\mathbf{h}_{m,k}(t)$ and $\rho(0 < \rho < 1)$ is the normalized autocorrelation function of a fading channel with motion at a constant velocity. For example, in the case of Rayleigh fading, we have $\rho = J_0(2\pi f_d)$ [37], where $J_0(\cdot)$ is the zeroth order Bessel function of the first kind, and f_d is the Doppler frequency.

B. Cache model

In this F-RAN, the F-APs can cache multiple content files from the centralization cache library in the cloud server, which stored a total of N contents that may be requested by UEs. Each of the content has a size of B and a unique content ID. Each F-AP can cache $C(C < N)$ contents, and the list of contents is $\mathcal{C}(t) = \{\Lambda_c | c = 1, 2, ..., C\}, \Lambda_c \in \{1, 2, ..., N\}$. Denote the list of all UEs' requests at slot t as $\mathbf{Req}(t) = \{Req_k | k = 1, 2, ..., K_0 + K_1\}, Req_k \in \{1, 2, ..., N\}$. We assume that a new request would not be generated until the transmission of the UE's previous requested content has been completed. The requested content Req_k of UE k may be cached in the F-AP (i.e., $Req_k \in \mathcal{C}(t)$) or not. Note that for the V2I UE $k \in \mathcal{K}_1$, there is a constraint on the delay and the requested content Req_k should be transmitted within T_{th} slots. We denote $T_k(t)$ as the residual slots of V2I UE $k \in \mathcal{K}_1$ to transmit the residual bits, the initial value of which is T_{th}.

Since the F-APs with popular contents cached can alleviate the burden on the fronthaul and provide a shorter transmission delay, the content caching is important for the network. Based on the requests $\mathbf{Req}(t)$, it is assumed that there is a decision $c(t) \in \{0, 1, 2, ..., C\}$ at each time slot t. When $c(t) = 0$, the currently requested content $Req^*(t)$ which is most popular would not be stored, otherwise it is stored by replacing the $c(t)$-th content $\Lambda_{c(t)}$ in the cache space of all the F-APs. $Req^*(t)$ is the most popular content that has not been cached, which is defined as follows.

$$Req^*(t) = \underset{n\in\mathcal{N}-\mathcal{C}(t)}{\arg\max} \sum_{k\in\mathcal{K}_0\cup\mathcal{K}_1} \mathbb{1}\{Req_k = n\}, \tag{4}$$

where $\mathbb{1}\{Req_k = n\}$ is an indicator function with value 1 when $Req_k = n$ holds, otherwise value 0.

In the RRU mode, we assume that the RRUs work in a full-duplex transmission way, where the RRUs forward the content Req_k to UE k while receiving the bits of the content from the cloud. Specially, the transmission procedure in the RRUs mode can be divided into two phases: phase I and phase

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

II. In the phase I, the data are transmitted from the cloud to the RRUs; while in the phase II, the data are transmitted from the RRUs to the UE. Suppose the fronthaul capacity R^{F1} is shared between UEs in $\mathcal{S}_0(t)$, the fronthaul rate allocated to UE k is $\frac{R^{F1}}{\|\mathcal{S}_0(t)\|}$ and the residual bits to be received from the cloud are

$$B_k^{res0}(t) = \max\{B_k^{res0}(t-1) - \frac{R^{F1}}{\|\mathcal{S}_0(t)\|}, 0\}, \tag{5}$$

where the initial value of $B_k^{res0}(t)$ is B. In the phase II, the residual bits to be transmitted are

$$\begin{aligned} B_k^{res1}(t) =& B_k^{res0}(t-1) - B_k^{res0}(t) \\ &+ \max\{B_k^{res1}(t-1) - R_{0,k}(t), 0\}, \end{aligned} \tag{6}$$

where the initial value of $B_k^{res1}(t)$ is 0 and the transmission rate between the RRUs and k-th UE at slot t is

$$\begin{aligned} R_{0,k}(t) &= \log(1 + \frac{P_0(t)\|\mathbf{h}_{0,k}(t)\|^2}{I_{0,k}^{intra}(t) + I_{0,k}^{inter}(t) + \sigma^2}), \\ I_{0,k}^{intra}(t) &= \sum_{j \in \mathcal{S}_0(t), j \neq k} P_0(t)\|\mathbf{h}_{0,k}(t)\|^2, \\ I_{0,k}^{inter}(t) &= \sum_{m=1}^{M_1} \sum_{j \in \mathcal{S}_m(t)} P_m(t)\|\mathbf{h}_{m,k}(t)\|^2. \end{aligned} \tag{7}$$

Note that for the V2I UE $k \in \mathcal{K}_1$ with $s_k(t) = 0$, the slots cost on phase I and phase II need to be less than the threshold T_{th}.

When the requested content Req_k has been cached in the F-AP, i.e., $Req_k \in \mathcal{C}(t)$, the UE k may connect to the F-AP and the transmission rate of UE k with mode selection $s_k(t) = m$ is

$$\begin{aligned} R_{m,k}(t) &= \log(1 + \frac{P_m(t)\|\mathbf{h}_{m,k}(t)\|^2}{I_{m,k}^{intra}(t) + I_{m,k}^{inter}(t) + \sigma^2}), \\ I_{m,k}^{intra}(t) &= \sum_{j \in \mathcal{S}_m(t), j \neq k} P_m(t)\|\mathbf{h}_{m,k}(t)\|^2, \\ I_{m,k}^{inter}(t) &= \sum_{m'=0, m' \neq m}^{M_1} \sum_{j \in \mathcal{S}_{m'}(t)} P'_m(t)\|\mathbf{h}_{m',k}(t)\|^2. \end{aligned} \tag{8}$$

Since the requested content has been cached, $Req_k \in \mathcal{C}(t)$, there is no need for the F-AP to receive the bits from the cloud, which implies that $B_k^{res1}(t) = 0$ holds for UE selecting the F-AP mode. UE k would obtain the content from the F-AP m directly, the residual bits of which to be transmitted are

$$B_k^{res0}(t) = \max\{B_k^{res0}(t-1) - R_{m,k}(t), 0\}, \tag{9}$$

With the current slot passed, the residual slots $T_k(t)$ for the V2I UE $k \in \mathcal{K}_1$ to transmit the residual bits $B_k^{res0}(t)$ are $T_k(t+1) = T_k(t) - 1$.

As it is shown, the RRU mode may provide a higher data rate than the F-AP mode due to the collaborative signal processing of the cloud. However, the F-AP mode could provide a shorter transmission delay because of the cache in F-APs. In this paper, we investigate the mode selection and content caching $\{s_k(t), c(t)\}$ in the proposed slice instances, where the FSMC and unknown content popularity distribution are characterized.

III. Problem Formulation and DRL-based Solution

In this section, the problem formulation and considered constraints for the coexisting RAN slice instances are presented. Since the demands of UEs are different and resource available are limited, the complexity of the original problem is very high, which makes it difficult to be solved using traditional approaches. To handle with this challenge, we would use a novel DRL-based approach and corresponding DRL-based solution is given. Based on the actions obtained from the DQN, a joint mode selection and content caching approach is derived to decide which mode should be selected by the UEs and which content should be cached in the F-APs.

A. Constraints Statement

Based on the principle of network slicing, the respective service requirement of each RAN slice instance should be assured. The rate requirement R_{th} of hotspot slice instance and the delay guarantee T_{th} of V2I slice instance are formulated as follows

$$\begin{aligned} &C1: R_{s_k,k}(t) >= R_{th}, \quad \forall k \in \mathcal{K}_0, \\ &C2: T_k(t) >= 0, \quad \forall k \in \mathcal{K}_1, \end{aligned} \tag{10}$$

where C1 implies that for any UE in the hotspot slice instance, the rate should be greater than the threshold R_{th}; and C2 represents that the transmission of the requested content Req_k should be completed before the residual slots T_k of V2I UE k count to 0.

Besides the service requirement of each RAN slice instance, the isolation between RAN slice instances should also be considered. Here, the isolation is ensured by limiting the inter-slice interference.

$$\begin{aligned} &C3: \sum_{j \in \mathcal{K}_1} P_{s_j}(t)\|\mathbf{h}_{s_j,k}(t)\|^2 <= I_{th0}, \quad \forall k \in \mathcal{K}_0, \\ &C4: \sum_{j \in \mathcal{K}_0} P_{s_j}(t)\|\mathbf{h}_{s_j,k}(t)\|^2 <= I_{th1}, \quad \forall k \in \mathcal{K}_1, \end{aligned} \tag{11}$$

where C3 means that the interference from the V2I slice instance to the hotspot UE $k \in \mathcal{K}_0$ should be under the threshold I_{th0}, and C4 means that the interference from the hotspot slice instance to the V2I UE $k \in \mathcal{K}_1$ should be under the threshold I_{th1}.

Since the RAN slice instances are constructed based on the same infrastructures, there are also constraints on the available resource in F-RANs. Specially, the fronthaul capacity R^{F1} between the cloud and RRUs is limited. Instead of regarding the maximum sum data rate transmitted on the fronthaul as constraints, we consider the constraints are related to the maximum number of UEs choosing RRU mode S_{max}^0 [38]. Similarly, since the signaling and coordination overhead in the F-AP increase with the number of the connected UEs, the UEs that the F-AP can support are assumed to be constrained due to the F-AP's limited capability.

$$\begin{aligned} &C5: \|\mathcal{S}_0(t)\| <= S_{max}^0, \\ &C6: \|\mathcal{S}_m(t)\| <= S_{max}^1, \quad \forall m \in \{1, 2, ..., M_1\}. \end{aligned} \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

In addition to the case of connecting to the RRUs directly, the F-APs with cache offer another transmission mode for hotspot and V2I UEs. In particular, UEs in the hotspot slice instance require a very high bit rate and, correspondingly, a huge burden on the fronthaul is generated if connected directly to the RRUs. Similarly, for UEs in the V2I slice instance choosing RRU mode, the delay from the cloud to the connected RRUs and the delay from the RRUs to UE should be both taken into consideration. In contrast, if connected to the F-APs, UEs would benefit from an alleviation on the burdened fronthaul and a reduction in the transmission delay at the expense of an elaborated mode selection and content caching. Therefore, we make an action at each time slot to identify the best way to associate the UEs with the F-APs/RRUs and manage the available caching storage in the F-APs. Based on the definition of mode selection and content caching, the following constraints should also be considered:

$$\begin{aligned} &C7: c(t) \in \{0,1,2,...,C\}, \\ &C8: s_k(t) \in \{0,1,2,...,M_1\}, \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1, \\ &C9: s_k(t) <= M_1 * \mathbb{1}\{Req_k \in \mathcal{C}(t)\}, \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1, \\ &C10: s_k(t) <= M_1 * \mathbb{1}\{B_k^{res1}(t) = 0\}, \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1. \end{aligned} \tag{13}$$

Constraint C7 implies that at each slot there is at most one content to be replaced in the cache of F-APs. Constraint C8 reflects the fact that UE k can only be connected to one F-AP or the RRUs. Besides, the UE's mode selection is also limited by C9 and C10. UE k can be assigned to connect the F-AP if and only if the requested content Req_k has been cached in the F-AP, which is stated in constraint C9. And C10 means UE k can reselect a mode after the residual bits in the RRUs have been transmitted.

B. Problem Formulation and the DRL-based Solution

Note that a problem with aforementioned constraints is a nonlinear optimization problem and falls within the category of integer programming. There have been a lot of different methods (e.g., branch-and-bound and genetic algorithms) that can be utilized to solve the integer programming problem. However, these solutions perform a centralized optimization at a given time and require simultaneously considering all UEs, F-APs, and RRUs. A high complexity would be generated as the number of UEs/F-APs/RRUs increases, other than the complexity resulting from the time-varying channel assumptions and the cache model assumptions. Moreover, there is a coupling effect among the variables in the constraints, which implies additional complexity would arise when computing the problem due to the iterative numerical analysis.

To overcome the given limitations, a unified study on content caching and mode selection is in essential. A learning-based and model-free approach is a promising candidate to manage complicated network scenarios and numerous optimization variables, especially by using deep neural networks. Considering definition constraints on content caching $\{c(t)\}$ and mode selection $\{s_k(t)\}$, the problem we investigated is discrete and low dimensional action space. Hence, we prefer the DQN to solve the targeted problem, in which a DQN is constructed in the centralized cloud via the historical information[1]. Through the DQN which learns the contents' popularities, a real-time content caching policy can be obtained. Similarly, mode selections for UEs in the slice instances would also be derived. The main advantage of the DRL-based solution is that it improves significantly the learning speed, especially in the problems with large state and action spaces. Through learning and building knowledge about the communication and caching environment, the solution of sophisticated network optimizations can be obtained, which provides autonomous decision-making for the network with minimum information exchanged.

Specially, the cloud server is responsible for acquiring each UE's channel state and F-APs' cache status to construct the system state $\mathbf{s}(t)$. Based on the state $\mathbf{s}(t)$, the agent DQN obtains an action $\mathbf{a}(t)$ to imply content caching and mode selection. A traditional ρ-greedy policy is utilized to balance the exploration and exploitation, i.e., to balance the reward maximization based on the knowledge already known with trying new actions to obtain knowledge unknown. After the action is performed, the reward can be obtained according to the reward function defined in the follow and the system will transfer to a new state. The current state, selected action, corresponding reward and the next state consist of the agent's experience, which is stored in the replay memory inside the DQN. The experience samples are utilized to train the approximate value function $Q(\mathbf{s}, \mathbf{a}; \omega)$ and target value function $\hat{Q}(\mathbf{s}, \mathbf{a}; \hat{\omega})$. As described in Algorithm 1, the value function $Q(\mathbf{s}, \mathbf{a})$ is trained towards the target value function $\hat{Q}(\mathbf{s}, \mathbf{a})$ by minimizing the loss function $Loss(\omega)$ at each iteration. The loss function can be written as

$$Loss(\omega) = \mathbb{E}[r_t + \tau \max_{\mathbf{a}' \in \mathcal{A}} \hat{Q}(\mathbf{s}_{t+1}, \mathbf{a}'; \hat{\omega}) - Q(\mathbf{s}_t, \mathbf{a}_t; \omega)]^2, \tag{14}$$

where τ is a discount factor. The explanation of the action, state and reward function in the DRL model are given as follows.

1) System State: The current system state $\mathbf{s}(t)$ is jointly determined by the channel coefficient $\mathbf{h}_{m,k}(t)$, residual bits and slots $\{B_k^{res0}(t), B_k^{res1}(t), T_k(t)\}$, cache feature $f_{z,c}(t)$, cache state $\mathcal{C}$ and the current request Req_k, which is defined as an vector with $\{(4 + M_1) * (K_0 + K_1) + K_1 + 4C + 3\}$ dimensions,

$$\begin{aligned} \mathbf{s}(t) = &\{\|\mathbf{h}_{m,k}(t)\|^2 | m = 0,1,...,M_1, k = 1,2,...,K_0 + K_1\} \\ &\times \{B_k^{res0}(t), B_k^{res1}(t) | k = 1,2,...,K_0 + K_1\} \\ &\times \{T_k(t) | k = 1,2,...,K_1\} \times \mathcal{C}(t) \times \mathbf{Req}(t) \\ &\times \{f_{z,c}(t) | z = 0,1,2, c = 0,1,...,C\}. \end{aligned} \tag{15}$$

Here, the cache feature $f_{z,c}(t)$ implies the popularity of content Λ_c within a specific term z, where $z \in \{0,1,2\}$ means short-term, medium-term and long-term, respectively. Specifically, the cache feature $f_{z,c}$ represents the total number of requests for content Λ_c in a specific short-, medium-, long-term, respectively, which varies as the cache state is updated.

[1]In this paper, we assume enormous computation resource are available in the cloud, which helps to train the DQN efficiently.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

Algorithm 1 The DRL-based solution for F-RAN slicing.

Initialize replay memory D to capacity L;
Initialize value function $Q(\mathbf{s}, \mathbf{a}; \omega)$ with weights ω;
Initialize target value function $\hat{Q}(\mathbf{s}, \mathbf{a}; \hat{\omega})$ with weights $\hat{\omega}$;
for episode $= 1, M$ **do**
  Initialize state $\mathbf{s}_0$ and action $\mathbf{a}_0$;
  for $t = 1, T$ **do**
    With probability ρ select a random action denoted as $\mathbf{a}_t^*$; otherwise select action $\mathbf{a}_t^* = \underset{\mathbf{a}_t \in \mathcal{A}}{\operatorname{argmax}} Q(\mathbf{s}_t, \mathbf{a}_t; \omega)$, where $Q(\cdot)$ is estimated by the DQN with parameter ω;
    Execute action $\mathbf{a}_t^*$ in emulator;
    Observe the reward r_t and the new state $\mathbf{s}_{t+1}$;
    Store the state transition $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ into D;
    Randomly sample a minibatch of state transitions $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ with a size of D_B from D;
    Set $y_t = r_t + \tau \max_{\mathbf{a}' \in \mathcal{A}} \hat{Q}(\mathbf{s}_{t+1}, \mathbf{a}'; \hat{\omega})$;
    Perform a gradient descent step on $(y_t - Q(\mathbf{s}_t, \mathbf{a}_t; \omega))^2$ with respect to the network parameters ω;
    Every preset steps reset $\hat{Q} = Q$
  end for
end for

For example, the short-term cache feature $f_{0,c}$ is the number of request for content Λ_c within the last request; while the long-term $f_{2,c}$ represents the number of requests for content Λ_c within the recent 100 requests. Since only the features from cached contents and the currently most popular content are considered, the index c ranges from 0 to the cache capacity C.

As shown in the system state $\mathbf{s}(t)$, the residual slots T_k of V2I UE k should be larger than 0 to meet constraint C2, otherwise the corresponding reward under any action would be 0. Besides, the number of possible system states can be very large as the number of F-APs, UEs and contents increases. Therefore it is difficult for traditional approaches to handle the problem because of the curse of dimensionality. Fortunately, the DQN has been proven to offer significant advantages in learning directly from high-dimensional inputs and solving non-convex and complex problems, thus it is proper to be used in our system.

2) System Action: In the system, the central cloud has to decide which content to be cached and which mode should be selected. Denote the system action space as $\mathcal{A}$, wherein its element $\mathbf{a}(t)$ is defined as follows.

$$\mathbf{a}(t) = \{s_k(t) | k = 1, 2, ..., K_0 + K_1\} \times \{c(t)\}. \tag{16}$$

In order to limit the action size, the value of $c(t)$ is restricted to vary from 0 to the cache capacity C, which meets the constraint C7. Therefore the choose of $c(t)$ directly affect cache state $\mathcal{C}$: the agent can only replace one selected cached content by the currently most popular content, or keep the cached contents the same. Besides, once content caching $c(t)$ is behaved, cache features $f_{z,c}(t)$ would be updated.

Similarly, the value of $s_k(t)$ ranges from 0 to M_1 to guarantee the constraint C8. Due to the constraint C5 and C6, the following conditions on the $s_k(t)$ must also be satisfied,

$$\begin{aligned} &\sum_{k \in \mathcal{K}_0 \cup \mathcal{K}_1} \mathbb{1}\{s_k(t) = 0\} <= S_{max}^0, \\ &\sum_{k \in \mathcal{K}_0 \cup \mathcal{K}_1} \mathbb{1}\{s_k = m\} <= S_{max}^1, \quad \forall m \in \{1, 2, ..., M_1\}, \end{aligned} \tag{17}$$

which further decrease the size of the action space $\mathcal{A}$.

By defining the action $\mathbf{a}(t)$ elaborately, parts of the constraints can be satisfied and residual constraints need to be taken into consideration when computing the reward.

3) Reward Function: The system reward represents the optimization objective, and the goal of the mode selection and content caching employing DQN is to get the content delivered to the UE with maximum reward.

To address constraints C1 $\sim$ C10, the reward $r(t)$ is defined as a function which reflects the degree of fulfillment of the optimization target and the constraints. In particular, the reward $r(t)$ is assigned a value of 0 whenever the constraints are not satisfied. The reasons for the constraints are unsatisfied may be:

– The propagation conditions in the selected mode do not allow achieving the rate threshold R_{th} (C1), or the inter-slice interference guarantee (C3 and C4);

– UE k is assigned to connect the F-AP without requested content Req_k (C9);

– UE k reselect a mode while the residual bits in the RRUs have not been transmitted (C10).

If the constraints C1 $\sim$ C10 are satisfied, the reward function is defined as weighted reward sum of hotspot UEs and V2I UEs,

$$r(t) = g(t) + \alpha_1 \sum_{k \in \mathcal{K}_0} r_k^0(t) + \alpha_2 \sum_{k \in \mathcal{K}_1} r_k^1(t), \tag{18}$$

where $r_k^0(t)$ and $r_k^1(t)$ are slice specific rewards, α_1 and α_2 are their weights ranging from 0 to 1. $g(t)$ is a cache-related reward and is defined as the weighted sum of the short and long-term cache hit ratios, which can be written as

$$g(t) = g^s(t) + \varphi g^l(t), \tag{19}$$

where φ is the weight to balance the short and long-term cache hit ratios. The short-term cache hit ratio $g^s(t)$ is the number of requests for local content within the next request. The long-term cache hit ratio $g^l(t)$ is the number of requests for local content within the next 10 requests. The $g(t)$ has a hidden relationship with the states, since the requests generated in the past and future are with the same unknown distribution. In particular, DQN is used to learn the hidden relationship from the recorded features in the states automatically. With the reward $g(t)$, the agent is compelled to choose the relatively more popular contents to cache.

For the hotspot UE, the reward $r_k^0(t)$ is defined as

$$r_k^0(t) = 1 - \frac{B_k^{res0}(t) + B_k^{res1}(t)}{B}, \tag{20}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

where $r_k^0(t)$ is a reward function of residual bits and implies that the less residual bits UE k have, the higher reward of UE k is.

For the V2I UE, the reward $r_k^1(t)$ is defined as

$$r_k^1(t) = 1 - \frac{B_k^{res0}(t) + B_k^{res1}(t)}{B} + \frac{T_k(t)}{T_{th}}, \tag{21}$$

where $r_k^1(t)$ implies that the less residual slots (more residual bits) UE k have, the lower reward of UE k is.

The central cloud gets reward $r(t)$ in system state $\mathbf{s}(t)$ when the action $\mathbf{a}(t)$ is performed in time slot t. The goal of applying DQN into the system model is to find a policy π that maximizes the expectation of cumulative reward, and the cumulative reward is the discounted sum of accumulated reward, which can be expressed as

$$CulRew = \max_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \epsilon^t r(t+1)] \tag{22}$$

where ϵ ranges from 0 to 1 and ϵ^t approaches to zero when t is large enough. In our work, a threshold for terminating the process is set. From the definition of expectation and the law of large numbers, the optimal policy π^* has the highest average cumulative reward given sufficient episodes.

Based on the definitions of rewards, it is noted that less residual bits $\{B_k^{res0}(t)+B_k^{res1}(t)\}$ to be transmitted and more residual slots $T_k(t)$ available would result in a higher reward $r(t)$. While under given $\{s_k^*(t), c^*(t)\}$, a smaller $\{B_k^{res0}(t)+B_k^{res1}(t)\}$ and larger $T_k(t)$ could be achieved via a higher transmission rate, since a higher rate can transmit more data in a slot. To address the impact of transmission rate on the reward, we leverage the method in [39] to derive the optimal rate under given $\{s_k^*(t), c^*(t)\}$. Consider the power threshold $\{P_{s_k^*}^{th} | k = 1, 2, ..., K_0+K_1\}$ for the selected RRUs and F-APs, the optimization problem is as follows.

$$\max_{\{P_{s_k^*}(t)\}} R_{TOT} = \sum_{k \in \mathcal{K}_0 \cup \mathcal{K}_1} R_{s_k^*(t),k}(t) \tag{23}$$

subject to the following constraints:

$$\begin{aligned} &C1, C3, C4, \\ &P_{s_k^*}(t) <= P_{s_k^*}^{th}, \quad k = 1, 2, ..., K_0 + K_1. \end{aligned} \tag{24}$$

Obviously, the problem (23) is non-convex due to the inter and intra interference. To make it more obvious to understand, we reformulate the problem (23) into a matrix-related form. Specially, we first define the transmit power matrix of the $K = K_0 + K_1$ users as

$$\mathbf{P} = \left(P_{s_1^*}(t), P_{s_2^*}(t), ..., P_{s_K^*}(t)\right)^{\mathrm{T}}. \tag{25}$$

We also define the $K \times K$ nonnegative matrix $\mathbf{F}$ with entries

$$\mathbf{F}_{i,j} = \begin{cases} 0 & , \text{if} \quad i = j \\ \frac{\|\mathbf{h}_{s_i^*(t),j}(t)\|^2}{\|\mathbf{h}_{s_j^*(t),j}(t)\|^2} & , \text{if} \quad i \neq j \end{cases} \tag{26}$$

and the $K \times 1$ vector $\mathbf{z}$

$$\mathbf{z} = \left(\frac{\sigma^2}{\|\mathbf{h}_{s_1^*(t),1}(t)\|^2}, \frac{\sigma^2}{\|\mathbf{h}_{s_2^*(t),2}(t)\|^2}, \cdots, \frac{\sigma^2}{\|\mathbf{h}_{s_K^*(t),K}(t)\|^2} \right)^{\mathrm{T}}. \tag{27}$$

Denote the signal to interference and noise ratios (SINRs) of all UEs as a $K \times 1$ vector $\mathbf{SINR}$, where the k-th element represents the SINR of UE k and $\mathbf{SINR}_k = \frac{P_{s_k^*(t)}(t)}{(\mathbf{FP}+\mathbf{z})_k}$. Similarly, denote the rates of all UEs as a $K \times 1$ vector $\mathbf{R}$, and the k-th element $\mathbf{R}_k$ is the rate of UE k, which can be expressed as:

$$\mathbf{R}_k = \log(1 + \mathbf{SINR}_k). \tag{28}$$

Rewrite problem (23) as:

$$\max_{\mathbf{P}} \sum_{k=1}^{K} \mathbf{R}_k \tag{29}$$

subject to the following constraints:

$$\mathbf{R}_k >= R_{th}, \quad \forall k \in \mathcal{K}_0, \tag{30}$$

$$(\boldsymbol{\theta}_k^1)^{\mathrm{T}} \mathbf{P} <= I_{th0}, \quad \forall k \in \mathcal{K}_0, \tag{31}$$

$$(\boldsymbol{\theta}_k^0)^{\mathrm{T}} \mathbf{P} <= I_{th1}, \quad \forall k \in \mathcal{K}_1, \tag{32}$$

$$\boldsymbol{\eta}_k^{\mathrm{T}} \mathbf{P} \leq P_{s_k^*(t)}^{th}, \quad \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1, \tag{33}$$

where the $K \times 1$ vector $\boldsymbol{\eta}_k$ is the k-th column of the $K \times K$ identity matrix, the $K \times 1$ vector $\boldsymbol{\theta}_k^0$ and $\boldsymbol{\theta}_k^1$ are defined as shown in the top of the next page.

As it is shown in problem (29), the targeted function is non-convex on the power but convex with respect to rate variable $\mathbf{R}_k$. The same applies to constraint (30). To deal with the residual constraints (31), (32) and (33) which are non-convex on $\mathbf{R}_k$, a new theorem is presented as follows.

Theorem 1: *Let*

$$\mathbf{B}_k^0 = \mathbf{F} + \frac{1}{I_{th0}} \mathbf{z} (\boldsymbol{\theta}_k^1)^{\mathrm{T}}, \quad \forall k \in \mathcal{K}_0 \tag{36}$$

$$\mathbf{B}_k^1 = \mathbf{F} + \frac{1}{I_{th1}} \mathbf{z} (\boldsymbol{\theta}_k^0)^{\mathrm{T}}, \quad \forall k \in \mathcal{K}_1 \tag{37}$$

$$\mathbf{B}_k^2 = \mathbf{F} + \frac{1}{P_{s_k^*(t)}^{th}} \mathbf{z} \boldsymbol{\eta}_k^{\mathrm{T}}, \quad \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1 \tag{38}$$

$$\mathbf{q} = \mathbf{FP} + \mathbf{z} \tag{39}$$

then constraint (31)*,* (32) *and* (33) *are equal to:*

$$\mathbf{B}_k^0 \mathrm{diag}(\exp(\mathbf{R})) \mathbf{q} \leq \left(\mathbf{I} + \mathbf{B}_k^0\right) \mathbf{q}, \quad \forall k \in \mathcal{K}_0, \tag{40}$$

$$\mathbf{B}_k^1 \mathrm{diag}(\exp(\mathbf{R})) \mathbf{q} \leq \left(\mathbf{I} + \mathbf{B}_k^1\right) \mathbf{q}, \quad \forall k \in \mathcal{K}_1, \tag{41}$$

$$\mathbf{B}_k^2 \mathrm{diag}(\exp(\mathbf{R})) \mathbf{q} \leq \left(\mathbf{I} + \mathbf{B}_k^2\right) \mathbf{q}, \quad \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1 \tag{42}$$

■

Proof: Similar proof can be found in the Appendix B of [39]. ■

Based on the above theorem, the power variable $\mathbf{P}$ in the original constraints is replaced by rate variable $\mathbf{R}_k$. Further transformation on the non-convex constraints (40), (41) and (42) are done based on the Perron-Frobenius theorem.

Theorem 2: *If the following non-negative matrixes hold,*

$$\tilde{\mathbf{B}}_k^0 = \left(\mathbf{I} + \mathbf{B}_k^0\right)^{-1} \mathbf{B}_k^0, \tag{43}$$

$$\tilde{\mathbf{B}}_k^1 = \left(\mathbf{I} + \mathbf{B}_k^1\right)^{-1} \mathbf{B}_k^1, \tag{44}$$

$$\tilde{\mathbf{B}}_k^2 = \left(\mathbf{I} + \mathbf{B}_k^2\right)^{-1} \mathbf{B}_k^2, \tag{45}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

$$\boldsymbol{\theta}_k^{\mathbf{0}} = \left(\|\mathbf{h}_{s_1^*(t),k}(t)\|^2, \|\mathbf{h}_{s_2^*(t),k}(t)\|^2, \cdots, \|\mathbf{h}_{s_{K_0}^*(t),k}(t)\|^2, \underbrace{0, ..., 0}_{K_1} \right)^{\mathrm{T}} \tag{34}$$

$$\boldsymbol{\theta}_k^{\mathbf{1}} = \left(\underbrace{0, ..., 0}_{K_0}, \|\mathbf{h}_{s_{K_0+1}^*(t),k}(t)\|^2, \|\mathbf{h}_{s_{K_0+2}^*(t),k}(t)\|^2, \cdots, \|\mathbf{h}_{s_K^*(t),k}(t)\|^2 \right)^{\mathrm{T}} \tag{35}$$

the constraints (40)*,* (41) *and* (42) *can be transformed into a convex form as:*

$$\log\left(\psi\left(\tilde{\mathbf{B}}_k^0 \mathrm{diag}(\exp(\mathbf{R}))\right)\right) \leq 0, \forall k \in \mathcal{K}_0, \tag{46}$$

$$\log\left(\psi\left(\tilde{\mathbf{B}}_k^1 \mathrm{diag}(\exp(\mathbf{R}))\right)\right) \leq 0, \forall k \in \mathcal{K}_1, \tag{47}$$

$$\log\left(\psi\left(\tilde{\mathbf{B}}_k^2 \mathrm{diag}(\exp(\mathbf{R}))\right)\right) \leq 0, \forall k \in \mathcal{K}_0 \cup \mathcal{K}_1, \tag{48}$$

where $\psi(\cdot)$ *is the Perron-Frobenius eigenvalue of a nonnegative matrix.* ■

Proof: Take the (40) as an example. Since $\left(\mathbf{I} + \mathbf{B}_k^0\right)^{-1}$ exists, multiply both sides of constraint (40) by $\left(\mathbf{I} + \mathbf{B}_k^0\right)^{-1}$, and the following equation can be derived as:

$$\mathbf{B}_k^0 \mathrm{diag}(\exp(\mathbf{R}))\mathbf{q} \leq \left(\mathbf{I} + \mathbf{B}_k^0\right)\mathbf{q} \tag{49}$$

$$\Leftrightarrow \quad \left(\mathbf{I} + \mathbf{B}_k^0\right)^{-1} \mathbf{B}_k^0 \mathrm{diag}(\exp(\mathbf{R}))\mathbf{q} \leq \mathbf{q} \tag{50}$$

$$\Leftrightarrow \quad \tilde{\mathbf{B}}_k^0 \mathrm{diag}(\exp(\mathbf{R}))\mathbf{q} \leq \mathbf{q}. \tag{51}$$

According to the subinvariance theorem [40], if a nonnegative matrix $\mathbf{A}$, a positive number a and a nonnegative vector $\mathbf{v}$ satisfy $\mathbf{A}\mathbf{v} \leq a\mathbf{v}$, then $\psi(\mathbf{A}) \leq a$ and the equality holds if and only if $\mathbf{A}\mathbf{v} = a\mathbf{v}$.

Let $\mathbf{A} = \tilde{\mathbf{B}}_k^0 \mathrm{diag}(\exp(\mathbf{R}))$, $a = 1$ and $\mathbf{v} = \mathbf{q}_n$ and rewrite the constraint as $\psi\left(\tilde{\mathbf{B}}_k^0 \mathrm{diag}(\exp(\mathbf{R}))\right) \leq 1$, which is the same as $\log\left(\psi\left(\tilde{\mathbf{B}}_k^0 \mathrm{diag}(\exp(\mathbf{R}))\right)\right) \leq 0$. ■

Due to the log-convexity property of the Perron-Frobenius eigenvalue, the constraints (46), (47) and (48) are convex. Now the concerned optimization problem in (29) is a convex optimization problem as follows

$$\begin{aligned} &\max_{\mathbf{R}} \sum_{k=1}^{K} \mathbf{R}_k \\ &s.t. (30), (46), (47), (48), \end{aligned} \tag{52}$$

which can be solved in polynomial time. Suppose $\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\upsilon}$ are the Lagrange multipliers corresponding to the constraints, the corresponding algorithm is as follows.

Algorithm 2 Sub-gradient method based iteration algorithm for the rate optimization

1: Set the iteration index $i = 1$.
2: Initialize the Lagrange multipliers $\boldsymbol{\lambda}^{(i)}, \boldsymbol{\beta}^{(i)}, \boldsymbol{\mu}^{(i)}, \boldsymbol{\upsilon}^{(i)}$, the step size $\boldsymbol{\xi}_{\boldsymbol{\lambda}}^{(i)}, \boldsymbol{\xi}_{\boldsymbol{\beta}}^{(i)}, \boldsymbol{\xi}_{\boldsymbol{\mu}}^{(i)}, \boldsymbol{\xi}_{\boldsymbol{\upsilon}}^{(i)}$, the maximum number of iterations $I_{\max}$ and the iteration threshold ν.
3: **for** $1 \leq i \leq I_{\max}$
4: Obtain $\mathbf{R}^{(i)}$ by solving the Lagrange function through the proximal gradient method in [41];
5: Update the Lagrange multipliers $\boldsymbol{\lambda}^{(i+1)}$, $\boldsymbol{\beta}^{(i+1)}$, $\boldsymbol{\mu}^{(i+1)}$,
6: $\boldsymbol{\upsilon}^{(i+1)}$ with $\boldsymbol{\lambda}^{(i)}$, $\boldsymbol{\beta}^{(i)}$, $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\upsilon}^{(i)}$ and the step sizes
7: $\boldsymbol{\xi}_{\boldsymbol{\lambda}}^{(i)}$, $\boldsymbol{\xi}_{\boldsymbol{\beta}}^{(i)}$,$\boldsymbol{\xi}_{\boldsymbol{\mu}}^{(i)}$, $\boldsymbol{\xi}_{\boldsymbol{\upsilon}}^{(i)}$;
8: **if** $\left|\boldsymbol{\lambda}^{(i+1)} - \boldsymbol{\lambda}^{(i)}\right| + \left|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\right| + \left|\boldsymbol{\mu}^{(i+1)} - \boldsymbol{\mu}^{(i)}\right|$
9: $+ \left|\boldsymbol{\upsilon}^{(i+1)} - \boldsymbol{\upsilon}^{(i)}\right| \leq \nu$
10: break out;
11: **end if**
12: $i = i + 1$;
13: **end for**
14: return $\mathbf{R} = \mathbf{R}^{(i)}$.

IV. Simulation Results

To demonstrate the performance of the proposed DRL-based solution to the realization of RAN slicing, the proposed approach has been evaluated through simulations. In this section, we first introduce simulation settings. Then, simulation results are discussed.

A. Simulation Setup

The performance evaluations are carried out through simulations in the following scenario: In the F-RAN scenario, $M_0 = 3$ RRUs and $M_1 = 7$ F-APs are configured initially and deployed in an area of 400 m $\times$ 400 m. Based on the scenario, we make a direct comparison between the proposed DRL-based solution and the existing content caching and mode selection solutions. The details on the mode selection and content caching in the comparison solutions will be introduced in the following subsection.

Assume that the considered UEs are allocated on a same subcarrier, the UEs are divided into two kinds according to the service type: the hotspot UEs and the V2I UEs. $K_0 = 5$ hotspot UEs and $K_1 = 3$ V2I UEs are considered. In the system, there are a total of $N = 500$ contents with size $B = 40$ bytes. The requests $\mathbf{Req}(t)$ are generated according to a Zipf distribution. Specially, the V2I UE needs to download the requested content within a time duration $T_{th} = 5$ slots.

The proposed DRL-based solution is tested under different simulation assumptions, especially under the condition of time-varying channel and unknown content popularity distribution. We first analyze the performance under different content popularity distribution, which includes the fixed content popularity distribution assumption (Zipf distribution with a 1.3 Zipf parameter), and a varying content popularity distribution assumption. In the latter case, the data set is generated with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

the same Zipf parameter, while content popularity rank is changing. We also consider the effects of different ρ values, wherein the channel coefficients are quantized and uniformly partitioned into 5 levels. The transition probability of remaining in the same state is related to the value of ρ.

The other considered parameters in the simulated scenario are as follows: The bandwidth is 200 kHz and the noise power spectral density is $N_0 = 10^{-12}$ W/Hz. The maximum transmission power of the RRU is 10 W and that of the F-AP is 40 W. The short-, medium-, long-term features are the number of requests for a file within the most recent 1, 10, 100 requests.

The whole parameters in the simulation are summarized as shown in Table I.

TABLE I
SIMULATION PARAMETERS

Noise power spectral density N_0	10^{-12} W/Hz
Number of antennas L_0, L_1	$2, 4$
Maximum connected UEs S^0_{max}, S^1_{max}	$3, 2$
Rate threshold R_{th}, time threshold T_{th}	4 bytes/slot,5 slots
Total contents N, content size B	$500, 40$ bytes
Replay memory capacity L	2000

B. Performance Comparison

To analyze the performance of our solution, we evaluate the performance and provide comparisons with other existing mode selection and content caching strategies. Note that the existing works on DRL applications [26]− [36] can not be applied directly to the scenario investigated in this work, we consider the following content caching solutions:

1) Least Recently Used (LRU) [42]: In this policy, the system makes a decision on the cache replacement based on the most recent requests for every cached content. When the cache storage is full, the least recently requested content in the cache would be replaced by the content $Req^*(t)$.

2) Least Frequently Used (LFU) [43]: In this policy, the system makes a decision on the cache replacement based on the number of requests for every cached content. When the cache storage is full, the least frequently requested content in the cache would be replaced by the content $Req^*(t)$.

3) First In First Out (FIFO) [44]: In this policy, the system makes a decision on the cache replacement based on the time when the content is cached. When the cache storage is full, the earliest stored content in the cache would be be replaced by the content $Req^*(t)$.

Also, the following mode selection under the content caching provided by the proposed DRL-based solution is considered:

1) All to Nearest AP (ANA): In this policy, the UE is connected to the F-AP or RRUs with the lowest propagation loss. Note that only when the F-AP has cached the desired content, the UE can select the F-AP. The performance of this approach may be bounded by the constrained capability of F-APs and fronthaul.

Fig. 2 shows the convergence performance of the proposed solution with different learning rates μ_l and batch sizes D_b. As it is shown, the average cumulative reward of the first several episodes is low since the DQN is at the beginning of the learning process. The performance then increases with the episode added until it reaches a relatively high and stable value. It is also observed that the value of different learning rates and batch sizes has effects on the convergence performance. Specifically, the solution converges faster when the learning rate μ_l is larger and the stable value is higher when the batch size D_b is smaller. Therefore, the learning rate and batch size are important to obtain the global optimum result, which should be chosen elaborately for a specific fog RAN slicing problem.

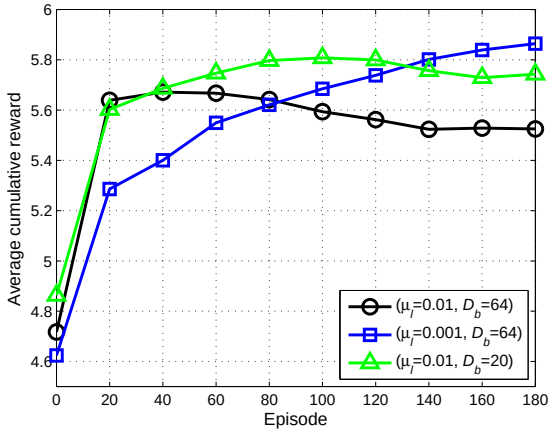

Fig. 2. Convergence performance of the proposed solution with different learning rates μ_l and batch sizes D_b.

To validate the benefits of proposed fog RAN slicing solution, the average cache hit ratio achieved by our solution and other caching solutions introduced above are compared in Fig. 3. Note that the simulation experiment is run for a total of 2000 time slots, the "average" cache hit ratio for different cache capacity C is estimated by performing tests over 2000 time slots and then taking the average. Under an unknown fixed Zipf distribution, the proposed realization solution for fog RAN slicing provides a higher average cache hit ratio for all cache capacity values. As the cache capacity increases, the average cache hit ratio increases until saturated. This is because that the contents with higher popularity have been cached and the rest contents provide limited gain. A larger cache capacity would not improve the cache hit ratio effectively and the cache hit ratio now is more effected by the popularity distribution of the contents.

To address the impact of a varying content popularity distribution, we provide Fig. 4 to demonstrate the performance of the fog RAN slicing solution. Specially, the requests are generated by a Zipf distribution, and the content popularity rank would change every 10 time slots. As it is shown in Fig. 4, the average cache hit ratio of the LRU, LFU, and FIFO are relatively low. However, the proposed DRL-based solution can learn from the environment and adapt its value function $Q(\mathbf{s}, \mathbf{a})$ quickly. After short time slots to modify the weights ω of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

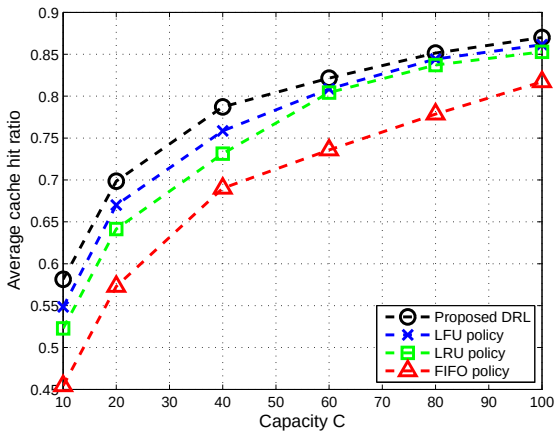

Fig. 3. The average cache hit ratio versus the cache capacity under different caching solution, wherein the proposed DRL-based solution outperforms than the other content caching policies.

the value function, the proposed DRL-based solution provides an increased average cache hit ratio and maintains it at a relatively stable value. Both Fig. 3 and Fig. 4 have validated that the proposed DRL-based solution outperforms the other benchmark approaches, and provides a better performance for hotspot UEs and V2I UEs.

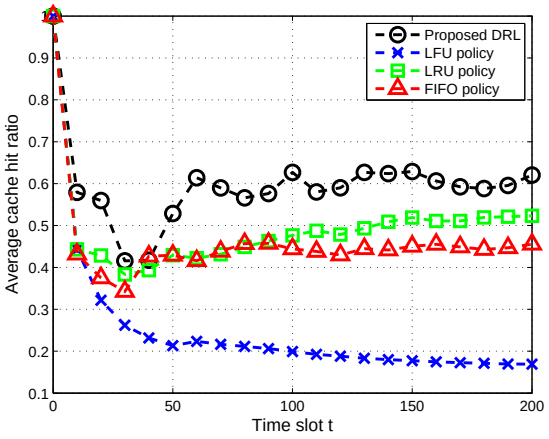

Fig. 4. Comparison of the average cache hit ratio under a Zipf distribution with changing content popularity rank. The proposed DRL-based solution is compared with the benchmark schemes

Besides the investigation on the impact of the content caching, we also study the effects of the time-varying channels. In the realization approach of fog RAN slicing, we made a performance comparison between different mode selection solutions, wherein invariant channels are assumed for the ANA policy. As it is shown in Fig. 5, the network's reward increases with the increase of ρ in different solutions. This is because a higher ρ value means more accurate channel state information. When the ρ is 1, the channels become invariant. It can be observed that the proposed DRL-based solution with larger cache has better performance compared to other solutions, because the proposed solution considers the impact of the time-varying channels and takes advantage of the larger cache size. The proposed DRL-based solution can provide a better mode selection in the time-varying channel environment via the trained DQN, which further maintains the performance of hotspot UEs and V2I UEs. With the transition probability increasing, the gap between the DRL and the other solutions is getting smaller. And the solutions with same cache capacity perform the same when the channel becomes invariant.

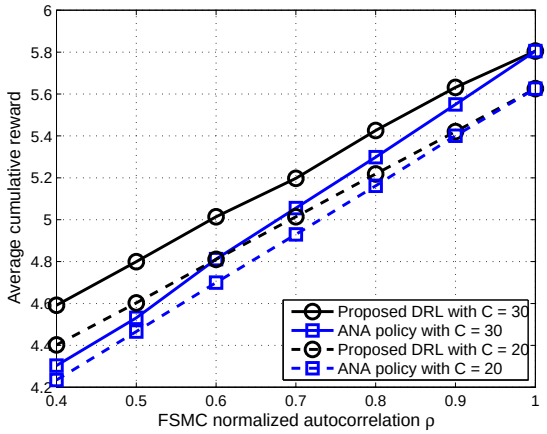

Fig. 5. The average cumulative reward versus ρ under different mode selection, wherein the proposed DRL-based solution outperforms than the ANA policy with same cache capacity. ρ is the normalized autocorrelation function and is related to the transformation probability between FSMC states.

Fig. 6 shows the effects of different fronthaul capacity and F-AP capability on the performance of fog RAN slicing solution. As discussed in the aforementioned work, the fronthaul capacity S^0_{max} in C5 and the F-AP capability S^m_{max} in C6 limit the number of UEs selecting the RRU mode and F-AP mode. From Fig. 6, we can see that the network's reward increases with the increase of fronthaul capacity S^0_{max} and F-AP capability S^m_{max}, where we assume the F-AP capability is the same for all F-APs in the simulation. The gain of the reward is because a higher value of the capacity will provide more flexible options for UEs, which may result in higher sum reward in the network.

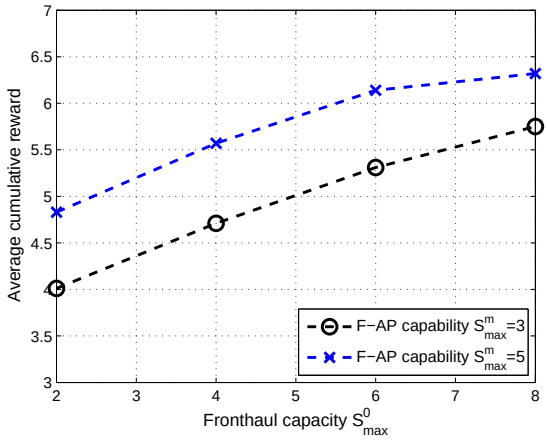

Fig. 6. Comparison of the average cumulative reward with different values of the total fronthaul capacity S^0_{max} and F-AP capability S^m_{max}.

V. Conclusion and Future Work

In this paper, we have investigated the realization of F-RAN slicing. Specially, the coexistence of the hotspot scenario and vehicle-to-infrastructure scenario are considered and two corresponding network slice instances are implemented. With the difference of users demands and the limitation of F-RAN resources considered, a content caching and mode selection optimization problem is formulated. To solve this high complexity problem efficiently, a DRL-based approach has been developed. Given the current system state, the cloud server provides an action via the well trained DRL model. With the content caching and mode selection derived from the obtained actions, the cumulative reward of the system, which is defined by the cache hit ratio and performance of slice instances, is maximized. By simulations, the impacts of learning rate and batch size have been shown. Moreover, compared with the baselines, the proposed approach can achieve significant performance improvements and robustness against the time-varying channels and unknown content popularity distribution.

There are still some other topics to be researched in the realization of network slicing via DRL. For example, the extend of our work to the realistic application scenario, wherein the network dynamics are non-Markovian time-variant. Some other key challenges for DRL techniques are also appealing to be solved. Robustness to model drift, handle with outliers, safe-learning, and generalization of algorithms should be further investigated. For example, when the number of UEs is significantly large, distributed algorithms like asynchronous advantage actor-critic will be more scalable, and thus worth exploring. To deal with the huge amounts of raw data, we would furthermore explore the fog computing to pre-process the raw data. With the intermediate data, the model training in the centralized cloud would speed up. Besides, the trained models can be deployed in the edge, which improves the response time of different services. Besides, content caching and mode selection will become more involved in a F-RAN that satisfies various application demands. The differentiated demands raise the problem of complicated system design that needs to address issues like QoS provisioning, computing, caching and radio resource allocation. Obviously, this is far beyond the extent of current research and worth in exploring.

References

[1] ITU-R, "IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," Sep. 2015.

[2] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94-100, May 2017.

[3] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Commun. Standards Mag.*, vol. 2, no. 1, pp. 60-65, Mar. 2018.

[4] R. Hattachi and J. Erfanian, "5G white paper," Feb. 2015.

[5] R. Wen, G. Feng, J. Tang, and *et. al.*, "On robustness of network slicing for next generation mobile networks, " *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 430-444, Jan. 2019.

[6] I. Afolabi, T. Taleb, K. Samdanis, and *et. al.*, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions, " *IEEE Commun. Surveys & Tuts.*, vol. 20, no. 3, pp. 2429-2453, 3rd Quart., 2018.

[7] X. Chen, H. Zhang, C. Wu, and *et. al.*, "Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4005-4018, Jun. 2019.

[8] C. Chang, N. Nikaein, O. Arouk, and *et. al.*, "Slice orchestration for multi-service disaggregated ultra-dense RANs," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 70-77, Aug. 2018.

[9] V. Ha and L. Le, "End-to-end network slicing in virtualized OFDMA-based cloud radio access networks," *IEEE Access*, vol. 5, pp. 18675-18691, Sep. 2017.

[10] Y. Loong, J. Loo, T. Chuah, and L. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146-2161, Apr. 2018.

[11] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102-108, Jun. 2017.

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

[13] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog computing based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46-53, Jul. 2016.

[14] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud Radio Access Networks: A New Perspective for Enhancing Spectral and Energy Efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[15] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003-5009, Aug. 2016.

[16] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, "Network slicing in fog radio access networks: Issues and challenges," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 110-116, Dec. 2017.

[17] J. Kwak, Y. Kim, L. Le, and S. Chong, "Hybrid content caching in 5G wireless networks: Cloud versus edge caching," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3030-3045, May. 2018.

[18] S. Chae, T. Quek, and W. Choi, "Content placement for wireless cooperative caching helpers: A tradeoff between cooperative gain and content diversity gain," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6795-6807, Oct. 2017.

[19] J. Liu, B. Bai, J. Zhang, and K. Letaief, "Cache placement in fog-RANs: From centralized to distributed algorithms, " *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039-7051, Nov. 2017.

[20] S. Park, O. Simeone, and S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks, " *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621-7632, Nov. 2016.

[21] J. Song, M. Sheng, T. Quek, and *et. al.*, "Learning-based content caching and sharing for wireless networks, " *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4309-4324, Oct. 2017.

[22] H. Wang and P. Chang, "On verifying the first-order Markovian assumption for a Rayleigh fading channel model," *IEEE Trans. Veh. Tech.*, vol. 45, no. 2, pp. 353-357, May 1996.

[23] W. Wang, H. Yang, and M. Alouini, "Wireless transmission of big data: A transmission time analysis over fading channel, " *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4315-4325, Jul. 2018.

[24] E. Vinogradov, A. Bamba, W. Joseph, and C. Oestges, "Physical-statistical modeling of dynamic indoor power delay profiles, " *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6493-6502, Oct. 2017.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, and *et. al.*, "Human-level control through deep reinforcement learning, " *Nature*, vol. 518, no. 7540, pp. 529-533, 2015.

[26] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks, " *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1207 - 1221, May 2016.

[27] Y. Sun, M. Peng, Y. Zhou, and *et. al.*, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys & Tuts.*, vol. 21, no. 4, pp. 3072 - 3108, Jun. 2019,

[28] X. Zhang and M. Peng, "Testbed design and performance emulation in fog radio access networks," *IEEE Netw.*, vol. 33, no. 3, pp. 49 - 57, Jun. 2019.

[29] Z. Luo, C. Wu, Z. Li, and W. Zhou "Scaling geo-distributed network function chains: A prediction and learning framework," *IEEE J. Select. Areas Commun.*, vol. 37, no. 8, pp. 1838-1850, Aug. 2019.

[30] M. Li, C. Chen, C. Hua, and X. Guan, "Intelligent latency-aware virtual network embedding for industrial wireless networks," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7484 - 7496, Oct. 2019.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.2965927, IEEE Transactions on Wireless Communications

[31] N. Luong, D. Hoang, S. Gong, and *et. al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys & Tuts.*, vo. 21, no. 4, pp. 3133 - 3174, May 2019.
[32] Y. He, Z. Zhang, F. Yuand, and *et. al.*, "Deep reinforcement learning-based optimization for cache enabled opportunistic interference alignment wireless networks, " *IEEE Trans. Veh. Tech.*, vol. 66, no. 11, pp. 10433-10445, Nov. 2017.
[33] Y. He, N. Zhang, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Tech.*, vol. 67, no. 1, pp. 44-55, Jan. 2018.
[34] Z. Zhang, Y. Yang, M. Hua, and *et. al.*, "Proactive caching for vehicular multi-view 3D video streaming via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2693-2706, May 2019.
[35] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning based mode selection and resource management for green fog radio access networks, " *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1960-1971, Apr. 2019.
[36] L. Tan and R. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Trans. Veh. Tech.*, vol. 67, no. 11, pp. 10190-10203, Nov. 2018.
[37] R. H. Clarke, "A statistical theory of mobile radio reception," *Bell Syst. Tech. J.*, vol. 47, no. 6, pp. 957-1000, Jul./Aug. 1968.
[38] J. Tang, W. Tay, T. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371-3384, May 2017.
[39] K. Zhang, M. Peng, P. Zhang, and X. Li, "Secrecy-optimized resource allocation for device-to-device communication underlaying heterogeneous networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 2, pp. 1822-1834, Feb. 2017.
[40] C. Tan, S. Friedland, and S. Low, "Nonnegative matrix inequalities and their application to nonconvex power control optimization," *SIAM J. Matrix Analysis Appl*, vol. 32, no. 3, pp. 1030-1055, 2011.
[41] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123-231, 2013.
[42] M. Ahmed, S. Traverso, P. Giaccone, and *et. al.*, "Analyzing the performance of LRU caches under non-stationary traffic patterns," *Computer Science*, vol. 155, no. 1, pp. 110-114, 2013.
[43] A. Jaleel, K. Theobald, S. Steely, and J. Emer, "High performance cache replacement using re-reference interval prediction (RRIP)," in *ACM SIGARCH Computer Architecture News*, vol. 38, pp. 60-71, ACM, 2010.
[44] D. Rossi and G. Rossini, "Caching performance of content centric networks under multi-path routing (and more)," *Relatorio tecnico, Telecom ParisTech*, pp. 1-6, 2011.

Hongyu Xiang is currently pursuing the Ph.D. degree at the Beijing University of Posts & Telecommunications (BUPT). He received the B.S. degree in communication engineering from the Fudan University, China, in 2013. His research interests are radio access network slicing, cooperative radio resource management and collaboration radio signal processing in large-scale networks like the fog radio access networks.

Mugen Peng (M'05-SM'11-F'20) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. Afterward, he joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering since 2012. In 2014, he was an Academic Visiting Fellow with Princeton University, Princeton, NJ, USA. He leads a Research Group focusing on wireless transmission and networking technologies with the State Key Laboratory of Networking and Switching Technology, BUPT. He has authored/coauthored over 100 refereed IEEE journal papers and over 300 conference proceeding papers. Dr. Peng was a recipient of the 2018 Heinrich Hertz Prize Paper Award, the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in the JCN 2016 and IEEE WCNC 2015. He is on the Editorial/Associate Editorial Board of the IEEE Communications Magazine, IEEE Internet of Things Journal, IEEE Access, IET Communications, and China Communications. He is the Fellow of IEEE and IET.

Shi Yan (M'19) received the Ph.D. degree in communication and information engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2017. He is currently an assistant professor in the key laboratory of universal wireless communications (Ministry of Education) at BUPT. In 2015, he was an Academic Visiting Scholar with Arizona State University, Tempe, AZ, USA. His research interests include game theory, resource management, deep reinforcement learning, stochastic geometry and fog radio access networks.