Wireless Resource Scheduling in Virtualized Radio Access Networks Using Stochastic Learning

Xianfu Chen, *Member, IEEE*, Zhu Han, *Fellow, IEEE*, Honggang Zhang, *Senior Member, IEEE*, Guoliang Xue, *Fellow, IEEE*, Yong Xiao, *Senior Member, IEEE*, and Mehdi Bennis, *Senior Member, IEEE*

Abstract—How to allocate the limited wireless resource in dense radio access networks (RANs) remains challenging. By leveraging a software-defined control plane, the independent base stations (BSs) are virtualized as a centralized network controller (CNC). Such virtualization decouples the CNC from the wireless service providers (WSPs). We investigate a virtualized RAN, where the CNC auctions channels at the beginning of scheduling slots to the mobile terminals (MTs) based on bids from their subscribing WSPs. Each WSP aims at maximizing the expected long-term payoff from bidding channels to satisfy the MTs for transmitting packets. We formulate the problem as a stochastic game, where the channel auction and packet scheduling decisions of a WSP depend on the state of network and the control policies of its competitors. To approach the equilibrium solution, an abstract stochastic game is proposed with bounded regret. The decision making process of each WSP is modeled as a Markov decision process (MDP). To address the signalling overhead and computational complexity issues, we decompose the MDP into a series of single-agent MDPs with reduced state spaces, and derive an online localized algorithm to learn the state value functions. Our results show significant performance improvements in terms of per-MT average utility.

Index Terms—Software-defined networking, radio access networks, network virtualization, multi-user resource scheduling, stochastic games, Markov decision process, learning

1 Introduction

The proliferation of mobile devices and data-intensive applications is driving up a significant demand for wireless services. It's predicted that the mobile data traffic will double every year from 2011 to 2020 [1]. The network operators (NOs) witness the exponential growth in data traffic, which has led to an unremitting increase in network provisioning. Network densification has been considered as the dominant theme towards future mobile networking [2]. As mobile data traffic evolves from being voice-dominated to being video-dominated and data-dominated, the revenue per unit-data for NOs is declining at an unhealthy rate. These trends are pushing NOs to look for more cost-effective ways to provide wireless services. Network sharing emerges as a disruptive mechanism to control both the capital expenditure (CapEx) and the operational expenditure (OpEx) [3]. This has started breaking up the vertically integrated nature of the conventional cellular system. Radio access network (RAN) sharing among NOs has recently received considerable attention due to the potential performance gain and inherent cost-saving benefits provided by utilizing economies of scale and avoiding duplicated investment on network infrastructure [4], [5], [6]. The worldwide combined CapEx and OpEx savings can be up to $60 billion through RAN sharing over a period of five years [7].

However, the RAN sharing activities are mostly based on long-term business agreements between the NOs, and most existing network sharing solutions have the drawbacks of separating both data and control planes among NOs, accommodating customized wireless services, and capability of adapting to dynamic network statistics in practice. It is important to design a fundamental framework that can boost network performances and reduce NOs' expenses by allowing more efficient and flexible network sharing. Network virtualization is emerging as a key enabler for RAN sharing, with which the traditional single ownerships of network infrastructure and spectrum resources can be decoupled from the wireless services [8], [9], [10], [11]. Consequently, the same physical network infrastructure is able to host multiple wireless service providers (WSPs) [12]. Although network virtualization is a promising technology for next generation RANs, one unique research challenge lies in wireless resource scheduling across mobile terminals (MTs) of

- *X. Chen is with the VTT Technical Research Centre of Finland Ltd, Oulu 90571, Finland. E-mail: xianfu.chen@vtt.fi.*
- *Z. Han is with the Department of Electrical and Computer Engineering and the Department of Computer Science, University of Houston, Houston, TX 77004 and the Department of Computer Science and Engineering, Kyung Hee University, Seoul 02447, South Korea. E-mail: zhan2@mail.uh.edu.*
- *H. Zhang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China. E-mail: honggangzhang@zju.edu.cn.*
- *G. Xue is with the Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ 85281. E-mail: xue@asu.edu.*
- *Y. Xiao is with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721. E-mail: xyong.2012@gmail.com.*
- *M. Bennis is with the Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland. E-mail: bennis@ee.oulu.fi.*

Manuscript received 9 Oct. 2016; revised 11 Aug. 2017; accepted 15 Aug. 2017. Date of publication 22 Aug. 2017; date of current version 2 Mar. 2018. (Corresponding author: Xianfu Chen.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2017.2742949

different WSPs. The diversified traffic from MTs in an RAN makes it difficult to achieve the optimal tradeoff between service flexibility and network scalability.

A software-defined networking (SDN) framework has been proposed to coordinate the complex radio access in cellular networks [13], [14]. SDN simplifies network management by decoupling the control plane from the data plane and is a natural platform for realizing network virtualization in RANs. The authors in [15], [16] presented software-defined cellular network architectures, allowing a remote controller to perform wireless resource slicing without modifying the MAC scheduler of a base station (BS). CloudIQ constitutes the first step towards wireless SDN by centralizing all data and control plane processing [17]. However, pushing all data processing to a central location imposes a huge demand on bandwidth and latency on backhaul. In [18], the authors proposed a proof-of-concept illustration of a software-defined RAN architecture where the independent physical BSs are virtualized as a centralized network controller (CNC), which performs centralized control plane operations.

In this paper, we are primarily concerned with a SDN-enabled virtualized RAN, where the CNC manages a limited set of channels and multiple competing WSPs bid the channel access opportunities for their MTs according to the network dynamics. The dynamics in a wireless network can be the results of the environmental disturbances and the interactions among the WSPs. For example, the packet arrival rates and the channel states can change from time to time due to the environmental disturbances. In the virtualized network, the CNC schedules channel usage upon collecting the bids from WSPs. The fairness during this centralized auction process is regulated through a Vickrey-Clarke-Groves (VCG) pricing mechanism[1] [19]. After receiving the channel scheduling outcomes, each MT of a WSP then proceeds to schedule the packets in the queue to optimize the expected long-term performance. Due to the competitive and stochastic nature of the channel auction and packet scheduling process, a general approach is to model the problem as a stochastic game [20], [21]. Stochastic game has already been widely adopted in the literature to analyze the dynamic optimization problem of wireless resource scheduling [21], [22]. Most of these results are, however, not scalable and with high implementation complexity, which hinder their applications in dense networks.

The contributions of this work are three-folded. First, we model the interactions among the CNC, WSPs and MTs in a SDN-enabled virtualized RAN with time-varying packet arrivals and channel states as a stochastic game, where each WSP aims to achieve its maximum expected long-term payoff. Second, due to the lack of complete information of the global network states and the control policies of all participating WSPs, we transform the original stochastic game into an abstract stochastic game with bounded performance regret. The decision making process of each WSP is modeled as a single-agent Markov decision process (MDP). Finally, we propose a linear decomposition approach for the per-WSP MDP to reduce the computational complexity at each WSP and the signalling overheads between the WSP and the subscribed MTs. Our proposed decomposition approach allows each MT to locally compute its individual state value functions. Without a priori statistical knowledge of channel states and packet arrivals, we propose a reinforcement learning [23] based algorithm for a MT to learn the state value functions by restructuring the immediate utility function in each scheduling slot. Compared to most existing works in literature, our proposed wireless resource scheduling scheme is simple and inherently self-organized, and can be deployed in dense networks with a large number of MTs.

1. One promising advantage of VCG is that the dominant policy for a WSP is to bid the true valuations for the spectrum resources.

TABLE 1
Major Notations Used in the Paper

$W/\mathcal{W}$	number/set of WSPs
$K/\mathcal{K}$	number/set of cells
$\lvert\mathcal{N}_{k,w}\rvert/\mathcal{N}_{k,w}$	number/set of MTs in cell k subscribed to WSP w
δ	scheduling slot duration
$H_{k,n}$, $H^t_{k,n}$	channel state of MT n in cell k
$Q_{k,n}$, $Q^t_{k,n}$	queue state of MT n in cell k
$Q^{(max)}$	maximum queue length
μ	packet size
$\varphi_{k,n}$, $\varphi^t_{k,n}$	channel allocation variable for MT n in cell k
$A_{k,n}$, $A^t_{k,n}$	packet arrivals of MT n in cell k
$\boldsymbol{\chi}_{k,n}$, $\boldsymbol{\chi}^t_{k,n}$	state of MT n in cell k
$\boldsymbol{\rho}_{k,n}$, $\boldsymbol{\rho}^t_{k,n}$	channel allocation of MT n in cell k
$R^t_{k,n}$	scheduled packets of MT n in cell k at slot t
$P^t_{k,n}$	transmit power of MT n in cell k at slot t
$P^{(max)}$	maximum transmit power for MTs
$U_{k,n}$	utility of MT n in cell k
$\phi_{k,n}$	weight of transmit power for MT n in cell k
$\alpha_{k,n}$	utility price for MT n in cell k
$\mathbb{U}_{k,n}$	expected long-term utility of MT n in cell k
F_w	payoff of WSP w
$\boldsymbol{\beta}_w$, $\boldsymbol{\beta}^t_w$	bid of WSP w
ν_w, ν^t_w	valuation of WSP w
$\boldsymbol{o}_w$, $\boldsymbol{o}^t_w$	channel request profile of WSP w
τ_w, τ^t_w	payment of WSP w
s_w, s^t_w	abstract state at WSP w
$\boldsymbol{\theta}$, $\boldsymbol{\theta}^t$	winner determination vector
$\boldsymbol{\pi}_w$, $\hat{\boldsymbol{\pi}}_w$	control policy of WSP w
$\pi^{(c)}_w$, $\hat{\pi}^{(c)}_w$	channel auction policy of WSP w
$\boldsymbol{\pi}^{(p)}_w$	packet scheduling policy for MTs of WSP w
V_w, $\hat{V}_w$	expected long-term payoff of WSP w
$\mathbb{V}_w$, $\hat{\mathbb{V}}_w$	optimal state value function of WSP w
$\mathbb{Q}_w$	state-channel allocation $\mathbb{Q}$-function of WSP w
$\mathbb{U}_w$	expected long-term payment of WSP w

The rest of this paper is organized as follows. In Section 2, we introduce the system model and the assumptions used throughout the paper. In Section 3, we formulate the problem of wireless resource scheduling as a stochastic game and discuss the corresponding game theoretic solution. In Section 4, we propose an abstract game based model to approximate the stochastic game and derive a learning scheme to solve the wireless resource scheduling problem online. We present the numerical results in Section 5 to evaluate the performance achieved by our proposed scheme. Finally, Section 6 draws the conclusion. In Table 1, we summarize the major notations of this paper.

2 System Model

As depicted in Fig. 1, we consider a SDN-enabled virtualized RAN, where multiple heterogeneous wireless services are supported over a common physical network infrastructure owned by a physical NO. The shared RAN covers a certain area. Different WSPs provide different wireless services, and each MT subscribes to only one WSP. Depending on the service types, the MTs belong to W groups, each of which

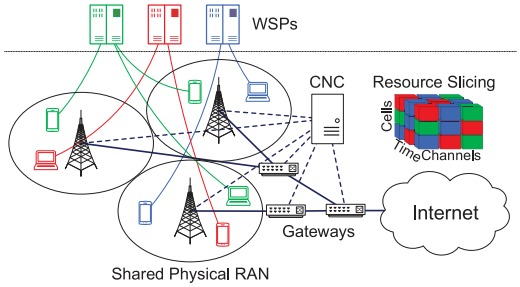

Fig. 1. System model of a virtualized radio access network (RAN) (WSP: Wireless service provider; CNC: Centralized network controller.). The mobile terminals (MTs) of different WSPs, which are shown in the same colors, are geographically distributed across the cells in the network. Over the time horizon, the CNC allocates the limited radio resource slices to MTs based on the bids that are announced by their respective subscribing WSPs.

corresponds to one service that is provided by a WSP $w \in \mathcal{W} \equiv \{1, \ldots, W\}$. Meanwhile, we assume the service area of interest consists of K cells and can be represented by a topology graph $\mathcal{TG} = \langle \mathcal{K}, \mathcal{D} \rangle$, where $\mathcal{K} = \{1, \ldots, K\}$ and $\mathcal{D} = \{d_{k,k'} : k \neq k', k, k' \in \mathcal{K}\}$ represents the relative locations between the cells with

$$d_{k,k'} = \begin{cases} 1, & \text{if cells } k \text{ and } k' \text{ are neighbours;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let $\mathcal{N}_{k,w}$ be the set of MTs in a cell $k \in \mathcal{K}$ that are subscribed to WSP w. Accordingly, $\mathcal{N}_w = \cup_{k \in \mathcal{K}} \mathcal{N}_{k,w}$ denotes the set of MTs of WSP w, $\mathcal{N}_k = \cup_{w \in \mathcal{W}} \mathcal{N}_{k,w}$ denotes the set of MTs in cell k, and $\mathcal{N} = \cup_{w \in \mathcal{W}} \mathcal{N}_w$ (or $\mathcal{N} = \cup_{k \in \mathcal{K}} \mathcal{N}_k$) denotes the set of all MTs in the network.

The system is operated over discrete scheduling slots of equal time duration δ (seconds) and the spectrum band used in the network is divided into J non-overlapping orthogonal channels with the same bandwidth M (Hz). Over the time horizon, the WSPs compete for the limited number of channels to serve their MTs. Each of them announces at the beginning of each scheduling slot $t \in \mathbb{N}_+$ to the CNC a bid, which is a vector $\underline{\boldsymbol{\beta}}_w^t = [\underline{\nu}_w^t, \underline{\boldsymbol{o}}_w^t]$, $\forall w \in \mathcal{W}$. Note that $\underline{\boldsymbol{\beta}}_w^t$ is not necessarily equal to $\boldsymbol{\beta}_w^t = [\nu_w^t, \boldsymbol{o}_w^t]$, where $\boldsymbol{o}_w^t = [o_{k,w}^t : k \in \mathcal{K}]$ with each $o_{k,w}^t$ being the exact number of requested channels in a cell k and ν_w^t is the true valuation of obtaining $\boldsymbol{o}_w^t$ from the perspective of WSP w. Upon receiving the auction bids $\underline{\boldsymbol{\beta}}^t = [\underline{\boldsymbol{\beta}}_w^t : w \in \mathcal{W}]$ from all WSPs, the CNC schedules the channels and computes the payment $\tau_w(\underline{\boldsymbol{\beta}}^t)$ for each WSP w. Let $\boldsymbol{\theta}(\underline{\boldsymbol{\beta}}^t) = [\theta_w(\underline{\boldsymbol{\beta}}^t) : w \in \mathcal{W}]$ be the winner determination vector at scheduling slot t, where $\theta_w(\underline{\boldsymbol{\beta}}^t) = 1$ if WSP w wins the channel auction while $\theta_w(\underline{\boldsymbol{\beta}}^t) = 0$ means no channel is allocated to the MTs of WSP w during the slot. For the channel scheduling at the CNC during a single slot, we assume that a channel cannot be allocated to adjacent cells in order to ensure that they do not interfere with each other, and in each cell, a MT can be assigned at most one channel and a channel can be assigned to at most one MT. Let $\boldsymbol{\rho}_{k,n}^t = [\rho_{k,n,j}^t : j \in \mathcal{J}]$ be the channel scheduling outcome for a MT $n \in \mathcal{N}_{k,w}$ at slot t, where $\mathcal{J} = \{1, \ldots, J\}$ and

$$\rho_{k,n,j}^t = \begin{cases} 1, & \text{if channel } j \text{ is allocated to} \\ & \text{MT } n \in \mathcal{N}_{k,w} \text{ at scheduling slot } t; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

For brevity, we may also write $\tau_w(\underline{\boldsymbol{\beta}}^t)$ and $\boldsymbol{\theta}(\underline{\boldsymbol{\beta}}^t)$ as τ_w^t and $\boldsymbol{\theta}^t = [\theta_w^t : w \in \mathcal{W}]$, respectively. Mathematically, $\forall t \in \mathbb{N}_+$, the constraints for a feasible channel scheduling decision at the CNC are given by

$$\left(\sum_{w \in \mathcal{W}} \sum_{n \in \mathcal{N}_{k,w}} \rho_{k,n,j}^t \right) \left(\sum_{w \in \mathcal{W}} \sum_{n \in \mathcal{N}_{k',w}} \rho_{k',n,j}^t \right) = 0,$$

$$\text{if } d_{k,k'} = 1, \forall d_{k,k'} \in \mathcal{D}, \forall j \in \mathcal{J}; \tag{3}$$

$$\sum_{n \in \mathcal{N}_{k,w}} \varphi_{k,n}^t = \theta_w^t o_{k,w}^t, \forall k \in \mathcal{K}, \forall w \in \mathcal{W}; \tag{4}$$

$$\sum_{w \in \mathcal{W}} \sum_{n \in \mathcal{N}_{k,w}} \rho_{k,n,j}^t \leq 1, \forall k \in \mathcal{K}, \forall j \in \mathcal{J}. \tag{5}$$

In (4), $\varphi_{k,n}^t = \sum_{j \in \mathcal{J}} \rho_{k,n,j}^t$ is a channel allocation variable that equals 1 if MT n is assigned a channel and 0 otherwise.

We assume a block-fading model for the channels within the spectrum band. In addition, we assume that the states of all channels experienced by a MT are identical during the period of each scheduling slot. Let $H_{k,n}^t \in \mathcal{H}$ be the channel state of a MT $n \in \mathcal{N}_{k,w}$ ($\forall k \in \mathcal{K}$ and $\forall w \in \mathcal{W}$) at slot t, where $\mathcal{H}$ is the set of possible channel states for all MTs. Then $H_{k,n}^t$ is quasi-static within each slot,[2] and is independent and identically distributed (i.i.d.) between different slots. The distribution of $H_{k,n}^t$ is assumed to be not known a priori. At every MT, a data queue is maintained for buffering the packets that arrive at the end of a scheduling slot.[3] The arriving packets get queued until transmission and we assume that every packet has a constant size of μ (bits). Let $Q_{k,n}^t$ and $A_{k,n}^t$ be, respectively, the queue length and the random new packet arrivals for MT n at slot t. The packet arrival process is assumed to be independent among the MTs and i.i.d. across scheduling slots. Like $H_{k,n}^t$, we assume that the distribution of $A_{k,n}^t$ is unknown as well. Let $R_{k,n}^t$ be the number of targeted packets that are to be removed from the queue at slot t. The number of packets that are eventually transmitted then is $\varphi_{k,n}^t R_{k,n}^t$, and the queue evolution of MT n can be written as

$$Q_{k,n}^{t+1} = \min\left\{ Q_{k,n}^t - \varphi_{k,n}^t R_{k,n}^t + A_{k,n}^t, Q^{(max)} \right\}, \tag{6}$$

where $Q^{(max)}$ is the maximum buffer size that restricts $Q_{k,n}^t \in \mathcal{Q} = \{0, \ldots, Q^{(max)}\}$. The structure of one scheduling slot is shown in Fig. 2.

Following the discussions in [24], the required power for a MT $n \in \mathcal{N}_{k,w}$ ($\forall k \in \mathcal{K}$ and $\forall w \in \mathcal{W}$) to reliably transmit $\varphi_{k,n}^t R_{k,n}^t$ error-free packets can be computed as

$$P_{k,n}^t = \frac{M \varsigma^2}{H_{k,n}^t} \left(2^{\frac{\mu \varphi_{k,n}^t R_{k,n}^t}{M \delta}} - 1 \right), \tag{7}$$

where ς^2 is the noise power spectral density. We can observe from (7) that given $H_{k,n}^t$, the transmit power $P_{k,n}^t$ is a strictly monotonically increasing function of $\varphi_{k,n}^t R_{k,n}^t$. Let $P^{(max)}$ be the maximum transmit power for all MTs, namely,

2. In practical cellular scenarios, the channel coherence time is much longer than a typical scheduling slot duration [22], [25].

3. The developed wireless resource scheduling scheme in this paper can also be extended to the downlink case.

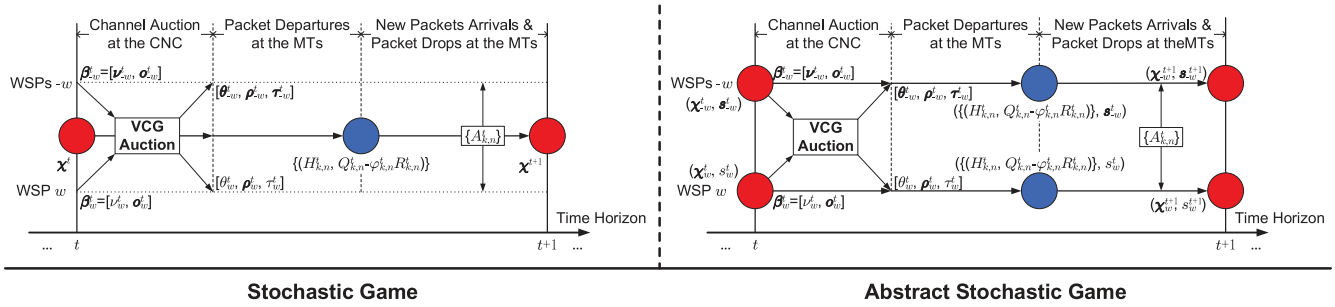

Fig. 2. The comparison between a stochastic game and the corresponding abstract stochastic game. In the original stochastic game $\mathcal{G}$, the WSPs need to acquire full network dynamics; while in the abstract stochastic game $\tilde{\mathcal{G}}$, each WSP behaves based on the abstraction of other competing WSPs' private information and the local network dynamics.

$P_{k,n}^t \leq P^{(max)}$, $\forall t$. Therefore, $0 \leq R_{k,n}^t \leq \min\{Q_{k,n}^t, Q_{k,n}^{(max),t}\}$, where $Q_{k,n}^{(max),t}$ is jointly determined by $H_{k,n}^t$ and $P^{(max)}$. We next need a payoff function to reward the WSPs for winning the channel auction. The instantaneous payoff of a WSP $w \in \mathcal{W}$ at slot t can be defined as

$$F_w(\boldsymbol{\chi}^t, \boldsymbol{\varphi}_w^t, \mathbf{R}_w^t) = \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}_{k,w}} \alpha_{k,n} U_{k,n}\left(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t\right) - \tau_w^t, \quad (8)$$

where $\boldsymbol{\chi}_{k,n}^t = (H_{k,n}^t, Q_{k,n}^t) \in \mathcal{X} \equiv \mathcal{H} \times \mathcal{Q}$ denotes the local network state at a MT n that encapsulates the local channel and queue state information, $\boldsymbol{\chi}^t = (\boldsymbol{\chi}_{k,n}^t : k \in \mathcal{K},\ n \in \mathcal{N}_k) \in \mathcal{X}^{|\mathcal{N}|}$ characterizes the global network state with $|\mathcal{N}|$ denoting the cardinality of the set $\mathcal{N}$, $\boldsymbol{\varphi}_w^t = [\varphi_{k,n}^t : k \in \mathcal{K}, n \in \mathcal{N}_{k,w}]$ and $\mathbf{R}_w^t = [R_{k,n}^t : k \in \mathcal{K}, n \in \mathcal{N}_{k,w}]$ are the channel allocation and the packet scheduling decisions for MTs of WSP w, and $\alpha_{k,n} \in \mathrm{IR}_+$ can be viewed as the unit price to charge MT n for realizing utility $U_{k,n}(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t)$ from consuming transmit power to schedule the queued packets to avoid the packet overflows (due to the limited buffer size), which is chosen to be

$$\begin{aligned} U_{k,n}\left(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t\right) &= \phi_{k,n} U_{k,n}^{(1)}\left(P_{k,n}^t\right) \\ &\quad + U_{k,n}^{(2)}\left(Q_{k,n}^t\right) + U_{k,n}^{(3)}\left(\eta_{k,n}^t\right). \end{aligned} \quad (9)$$

In (9), $\eta_{k,n}^t = \max\{Q_{k,n}^t - \varphi_{k,n}^t R_{k,n}^t + A_{k,n}^t - Q^{(max)}, 0\}$ defines the number of packet drops, and we assume that $U_{k,n}^{(1)}(\cdot)$, $U_{k,n}^{(2)}(\cdot)$ and $U_{k,n}^{(3)}(\cdot)$ are positive monotonically decreasing functions. $U_{k,n}^{(2)}(\cdot)$ and $U_{k,n}^{(3)}(\cdot)$ measure the satisfactions of the buffer delay and the packet drops, respectively. The buffer delay can be termed as the immediate queue occupancy [26] and the packet drops occur when the queue vacancy is less than the number of arriving packets. $\phi_{k,n} \in \mathrm{IR}_+$ is a constant weighting factor that balances the importance of the transmit power consumption. With slight abuse of notations, we rewrite $U_{k,n}^{(1)}(P_{k,n}^t)$, $U_{k,n}^{(2)}(Q_{k,n}^t)$ and $U_{k,n}^{(3)}(\eta_{k,n}^t)$ as $U_{k,n}^{(1)}(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t)$, $U_{k,n}^{(2)}(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t)$ and $U_{k,n}^{(3)}(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t)$.

3 Problem Formulation and Game Theoretic Approach

In this section, we first formulate the problem of wireless resource scheduling (namely, the competitive channel auction and packet scheduling) as a stochastic game and then discuss the game theoretic solution.

3.1 Stochastic Game Formulation

We design a stationary control policy $\boldsymbol{\pi}_w = (\pi_w^{(c)}, \boldsymbol{\pi}_w^{(p)})$ for each WSP $w \in \mathcal{W}$, where $\pi_w^{(c)}$ and $\boldsymbol{\pi}_w^{(p)} = (\pi_{k,n}^{(p)} : k \in \mathcal{K}, n \in \mathcal{N}_{k,w})$ are the channel auction and the packet scheduling policies, respectively. Note that the packet scheduling policy $\pi_{k,n}^{(p)}$ is MT specified. Then $\boldsymbol{\pi}_w^{(p)}$ depends only on $\boldsymbol{\chi}_w^t = (\boldsymbol{\chi}_{k,n}^t : k \in \mathcal{K}, n \in \mathcal{N}_{k,w}) \in \mathcal{X}_w = \mathcal{X}^{|\mathcal{N}_w|}$. The joint control policy of all WSPs is represented by $\boldsymbol{\pi} = (\boldsymbol{\pi}_w : w \in \mathcal{W})$. With $\boldsymbol{\pi}_w$, WSP w at each slot t announces the bid $\underline{\boldsymbol{\beta}}_w^t$ to the CNC for channel scheduling and decides the number of packets $\mathbf{R}_w^t$ to be transmitted after observing $\boldsymbol{\chi}^t \in \mathcal{X}^{|\mathcal{N}|}$, i.e., $\boldsymbol{\pi}_w(\boldsymbol{\chi}^t) = (\pi_w^{(c)}(\boldsymbol{\chi}^t), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w^t)) = (\underline{\boldsymbol{\beta}}_w^t, \mathbf{R}_w^t) \in \mathcal{Y}_w$. The CNC subsequently performs centralized channel scheduling with the VCG mechanism that maximizes the "social welfare",[4]

$$\begin{aligned} &\max_{\boldsymbol{\theta}^t} \sum_{w\in\mathcal{W}} \theta_w^t \underline{v}_w^t \\ &\text{s.t.} \quad \text{constraints (3), (4) and (5),} \end{aligned} \quad (10)$$

and results in the payment for each WSP w,

$$\tau_w^t = \max_{\boldsymbol{\theta}_{-w}^t} \sum_{w'\in\mathcal{W}\setminus\{w\}} \theta_{w'}^t \underline{v}_{w'}^t - \max_{\boldsymbol{\theta}^t} \sum_{w'\in\mathcal{W}\setminus\{w\}} \theta_{w'}^t \underline{v}_{w'}^t, \quad (11)$$

where $-w$ denotes all the other WSPs in set $\mathcal{W}$ without WSP w. Scheduling packets over the assigned channels leads to utilities, $U_{k,n}(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t)$, $\forall k \in \mathcal{K}$, $\forall n \in \mathcal{N}_{k,w}$. In particular, the VCG auction for channel scheduling at a slot t possesses the following economic properties:

- *Efficiency* – When all WSPs announce their true bids, the CNC schedules channels to maximize the sum of valuations, resulting in efficient channel utilization.
- *Individual Rationality* – Each WSP w can expect a non-negative payoff $\underline{v}_w^t - \tau_w^t$ at any slot t.
- *Truthfulness* – No WSP can improve its payoff by bidding different from its true valuation, which implies that the optimal bid at any scheduling slot t is $\underline{\boldsymbol{\beta}}_w^t = \boldsymbol{\beta}_w^t$, $\forall w \in \mathcal{W}$.

The VCG mechanism has been proven to be efficient for channel scheduling in one slot [28]. In the following sections, we put our efforts on investigating the potential of adapting VCG to the wireless resource scheduling problem in a virtualized RAN with time-varying network states, that

4. Other fairness rules (e.g., [27]) can also be implemented and do not affect the proposed wireless resource scheduling scheme in this paper.

is, the problem of stochastic channel auction and packet scheduling, which is formulated as a stochastic game $\mathcal{G}$ to be detailed in the following. In the game $\mathcal{G}$, W WSPs are the players and there are a set $\mathcal{X}^{|\mathcal{N}|}$ of global network states and a collection of decision-making sets, $\mathcal{Y}_w, \forall w \in \mathcal{W}$.

The stationary joint control policy $\boldsymbol{\pi}$ induces a probability distribution over the sequence of global network states $\{\boldsymbol{\chi}^t : t \in \mathrm{IN}_+\}$, and hence a probability distribution over the sequences of per-slot payoffs $\{F_w(\boldsymbol{\chi}^t, \boldsymbol{\varphi}_w^t, \mathbf{R}_w^t) : t \in \mathrm{IN}_+\}$, $\forall w \in \mathcal{W}$. From assumptions on the channel states and the packet arrivals, the random process, $\boldsymbol{\chi}^t, t \in \mathrm{IN}_+$, is Markovian with the following transition probability

$$\begin{aligned}
&\mathsf{Pr}\Big\{\boldsymbol{\chi}^{t+1}|\boldsymbol{\chi}^t, \boldsymbol{\varphi}\Big(\boldsymbol{\pi}^{(c)}\big(\boldsymbol{\chi}^t\big)\Big), \boldsymbol{\pi}^{(p)}\big(\boldsymbol{\chi}^t\big)\Big\} \\
&= \mathsf{Pr}\Big\{\mathbf{Q}^{t+1}|\mathbf{Q}^t, \boldsymbol{\varphi}\Big(\boldsymbol{\pi}^{(c)}\big(\boldsymbol{\chi}^t\big)\Big), \boldsymbol{\pi}^{(p)}\big(\boldsymbol{\chi}^t\big)\Big\}\mathsf{Pr}\big\{\mathbf{H}^{t+1}\big\} \\
&= \prod_{k\in\mathcal{K}}\prod_{n\in\mathcal{N}_k} \mathsf{Pr}\Big\{Q_{k,n}^{t+1}|Q_{k,n}^t, \varphi_{k,n}\Big(\boldsymbol{\pi}^{(c)}\big(\boldsymbol{\chi}^t\big)\Big), \pi_{k,n}^{(p)}\Big(\boldsymbol{\chi}_{k,n}^t\Big)\Big\} \\
&\times \prod_{k\in\mathcal{K}}\prod_{n\in\mathcal{N}_k} \mathsf{Pr}\Big\{H_{k,n}^{t+1}\Big\},
\end{aligned} \tag{12}$$

where $\mathsf{Pr}\{\cdot\}$ denotes the probability of an event, $\boldsymbol{\varphi} = [\boldsymbol{\varphi}_w : w \in \mathcal{W}]$ is the global channel allocation, while $\boldsymbol{\pi}^{(c)} = (\pi_w^{(c)} : w \in \mathcal{W})$ and $\boldsymbol{\pi}^{(p)} = (\boldsymbol{\pi}_w^{(p)} : w \in \mathcal{W})$ are the joint channel auction and the joint packet scheduling policies. Taking expectation with respect to the sequence of per-slot payoffs, the expected long-term payoff of a WSP w for a given initial state $\boldsymbol{\chi}^1 = \boldsymbol{\chi}$ can be expressed as

$$\begin{aligned}
&V_w(\boldsymbol{\chi}, \boldsymbol{\pi}) = (1-\gamma) \\
&\times \mathsf{E}\left[\sum_{t=1}^{\infty}(\gamma)^{t-1} F_w\Big(\boldsymbol{\chi}^t, \boldsymbol{\varphi}_w\Big(\boldsymbol{\pi}^{(c)}\big(\boldsymbol{\chi}^t\big)\Big), \boldsymbol{\pi}_w^{(p)}\big(\boldsymbol{\chi}_w^t\big)\Big)|\boldsymbol{\chi}\right],
\end{aligned} \tag{13}$$

where $\gamma \in [0, 1)$ is the discount factor and $(\gamma)^{t-1}$ denotes the discount factor to the $(t-1)$th power. $V_w(\boldsymbol{\chi}, \boldsymbol{\pi})$ is also named as the state value function of WSP w in state $\boldsymbol{\chi}$ under joint policy $\boldsymbol{\pi}$. Note that for the non-ergodic Markov system[5] in this paper, we define an expected infinite-horizon discounted payoff function in (13) as the optimization objective. Nevertheless, the expected infinite-horizon undiscounted payoff

$$\overline{F}_w(\boldsymbol{\pi}) = \mathsf{E}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} F_w\Big(\boldsymbol{\chi}^t, \boldsymbol{\varphi}_w\Big(\boldsymbol{\pi}^{(c)}\big(\boldsymbol{\chi}^t\big)\Big), \boldsymbol{\pi}_w^{(p)}\big(\boldsymbol{\chi}_w^t\big)\Big)\right], \tag{14}$$

can be approximated by (13) when γ approaches 1 [29]. The aim of each WSP w is to find a best-response control policy $\boldsymbol{\pi}_w^*$ that maximizes $V_w(\boldsymbol{\chi}, \boldsymbol{\pi}_w, \boldsymbol{\pi}_{-w})$ for any given initial network state $\boldsymbol{\chi}$, which can be formally formulated as

$$\boldsymbol{\pi}_w^* = \arg\max_{\boldsymbol{\pi}_w} \; V_w(\boldsymbol{\chi}, \boldsymbol{\pi}_w, \boldsymbol{\pi}_{-w}), \forall \boldsymbol{\chi} \in \mathcal{X}^{|\mathcal{N}|}. \tag{15}$$

In stochastic game $\mathcal{G}$, a Nash equilibrium (NE) is generally accepted as a solution concept to describe the most rational decision makings by the WSPs and is defined as follows.

5. Due to the competition among the WSPs, the control policies, $\boldsymbol{\pi}_w$, $\forall w \in \mathcal{W}$, are not unichain.

Definition 1. *In stochastic game $\mathcal{G}$, a NE is a tuple of policies $(\boldsymbol{\pi}_w^* : w \in \mathcal{W})$, where each $\boldsymbol{\pi}_w^*$ of a WSP w is the best response to the other WSPs' control policies $\boldsymbol{\pi}_{-w}^*$.*

The result below shows that for the considered stochastic game, there always exists a NE in stationary policies.

Theorem 1. *Every W-player stochastic game with discounted payoffs has at least one stationary NE [30].*

Remark 1. From (13), we can observe that the expected long-term payoff of a WSP $w \in \mathcal{W}$ depends on information of both the global network state and the joint control policies of all WSPs. In other words, the channel auction and the packet scheduling decision makings of WSPs are closely coupled with each other in our game model.

3.2 Best-Response Joint Channel Auction and Packet Scheduling

Assume that each WSP $w \in \mathcal{W}$ obtains the global network state information and all the other WSPs play the NE policies $\boldsymbol{\pi}_{-w}^* = (\boldsymbol{\pi}_{-w}^{(c),*}, \boldsymbol{\pi}_{-w}^{(p),*})$, WSP w's best-response policy can be obtained by solving (16), $\forall \boldsymbol{\chi} \in \mathcal{X}^{|\mathcal{N}|}$, where $\mathbb{V}_w(\boldsymbol{\chi}) = V_w(\boldsymbol{\chi}, \boldsymbol{\pi}_w^*, \boldsymbol{\pi}_{-w}^*)$ is the optimal state value function and $\boldsymbol{\chi}'$ is the subsequent global network state. The channel scheduling and the packet scheduling decisions are made independently at the CNC and the WSP sides, which motivates us to define a state-channel allocation $\mathbb{Q}$-function (17). The best-response control policy $\boldsymbol{\pi}_w^*$ is composed of the channel auction policy $\pi_w^{(c),*}$ and the packet scheduling policy $\boldsymbol{\pi}_w^{(p),*}$. With any given channel auction policy $\pi_w^{(c)}$, $\boldsymbol{\pi}_w^{(p),*} = (\pi_{k,n}^{(p),*} : k \in \mathcal{K}, n \in \mathcal{N}_{k,w})$ of WSP w is the solution to (18) and $\pi_w^{(c),*}$ can be of

$$\pi_w^{(c),*}(\boldsymbol{\chi}) = \arg\max_{\pi_w^{(c)}(\boldsymbol{\chi})} \; \mathbb{Q}_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big)\Big). \tag{19}$$

Remark 2. From the above analysis, we can observe that achieving the NE is still a technically challenging task. Each WSP $w \in \mathcal{W}$ in the network has to know complete information of the global network dynamics (20), $\forall \boldsymbol{\chi}, \boldsymbol{\chi}' \in \mathcal{X}^{|\mathcal{N}|}$, which incurs exponential computation complexity even for a network of reasonable size. More importantly, the private queue state transition probabilities and channel state distributions at other competing WSPs, i.e., the $\mathsf{Pr}\{Q'_{k,n'}|Q_{k,n'}, \varphi_{k,n'}(\pi_w^{(c)}(\boldsymbol{\chi}), \pi_{-w}^{(c),*}(\boldsymbol{\chi})), \pi_{k,n'}^{(p),*}(\boldsymbol{\chi}_{k,n'})\}$ and the $\mathsf{Pr}\{H'_{k,n'}\}$, $n' \in \mathcal{N}_{k,w'}$, $\forall k \in \mathcal{K}$ and $\forall w' \in \mathcal{W} \setminus \{w\}$, are impossible to be known.

4 Abstract Stochastic Game Reformulation and Online Localized Learning

In this section, we elaborate on how to determine the best-response control policies in a computationally efficient way, by which the channel auction and the packet scheduling decisions can be made with limited information. More specifically, at each scheduling slot, the WSPs announce to the CNC the bids based on the their own local abstract network information. After receiving the channel scheduling outcomes from the CNC, the MTs decide the number of packets to be delivered.

$$\mathbb{V}_w(\boldsymbol{\chi}) = \max_{\boldsymbol{\pi}_w(\boldsymbol{\chi})} \Bigg\{ (1-\gamma) F_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big) + \gamma \sum_{\boldsymbol{\chi}' \in \mathcal{X}^{|\mathcal{N}|}} \mathsf{Pr}\Big\{ \boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\varphi}\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \Big(\boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w), \boldsymbol{\pi}_{-w}^{(p),*}(\boldsymbol{\chi}_{-w})\Big) \Big\} \mathbb{V}_w(\boldsymbol{\chi}') \Bigg\} \tag{16}$$

$$\mathbb{Q}_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big)\Big) = \max_{\boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)} \Bigg\{ (1-\gamma) F_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big) + \gamma \sum_{\boldsymbol{\chi}' \in \mathcal{X}^{|\mathcal{N}|}} \mathsf{Pr}\Big\{ \boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\varphi}\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \Big(\boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w), \boldsymbol{\pi}_{-w}^{(p),*}(\boldsymbol{\chi}_{-w})\Big) \Big\} \mathbb{V}_w(\boldsymbol{\chi}') \Bigg\} \tag{17}$$

$$\mathbb{Q}_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big)\Big) = \max_{\boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)} \Bigg\{ (1-\gamma) F_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big) + \gamma \sum_{\boldsymbol{\chi}' \in \mathcal{X}^{|\mathcal{N}|}} \mathsf{Pr}\Big\{ \boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\varphi}\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \Big(\boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w), \boldsymbol{\pi}_{-w}^{(p),*}(\boldsymbol{\chi}_{-w})\Big) \Big\} \max_{\pi_w^{(c)}(\boldsymbol{\chi}')} \mathbb{Q}_w\Big(\boldsymbol{\chi}', \boldsymbol{\varphi}_w\Big(\pi_w^{(c)}(\boldsymbol{\chi}'), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi}')\Big)\Big) \Bigg\} \tag{18}$$

$$\begin{aligned} &\mathsf{Pr}\Big\{ \boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\varphi}\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \Big(\boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w), \boldsymbol{\pi}_{-w}^{(p),*}(\boldsymbol{\chi}_{-w})\Big) \Big\} \\ &= \prod_{k \in \mathcal{K}} \prod_{n \in \mathcal{N}_{k,w}} \mathsf{Pr}\Big\{ Q'_{k,n} | Q_{k,n}, \varphi_{k,n}\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n}) \Big\} \prod_{k \in \mathcal{K}} \prod_{n \in \mathcal{N}_{k,w}} \mathsf{Pr}\Big\{ H'_{k,n} \Big\} \\ &\times \prod_{w' \in \mathcal{W} \setminus \{w\}} \prod_{k \in \mathcal{K}} \prod_{n' \in \mathcal{N}_{k,w'}} \mathsf{Pr}\Big\{ Q'_{k,n'} | Q_{k,n'}, \varphi_{k,n'}\Big(\pi_w^{(c)}(\boldsymbol{\chi}), \boldsymbol{\pi}_{-w}^{(c),*}(\boldsymbol{\chi})\Big), \pi_{k,n'}^{(p),*}(\boldsymbol{\chi}_{k,n'}) \Big\} \prod_{w' \in \mathcal{W} \setminus \{w\}} \prod_{k \in \mathcal{K}} \prod_{n' \in \mathcal{N}_{k,w'}} \mathsf{Pr}\Big\{ H'_{k,n'} \Big\} \end{aligned} \tag{20}$$

4.1 Stochastic Game Abstraction

To avoid the high computational cost due to the large dimensionality of the network state space $\mathcal{X}^{|\mathcal{N}|}$ and to capture the impacts of other WSPs' decision makings in a non-cooperative networking environment, there is an imperative need for a WSP to approximate the inter-WSP couplings. One possible solution is to construct an abstract version of the original stochastic game $\mathcal{G}$ [31]. In an abstract stochastic game $\hat{\mathcal{G}}$, all WSPs behave based on their own local network states and abstractions of other competing WSPs' local states. We denote $\hat{\mathcal{X}}_w = \mathcal{X}_w \times \mathcal{S}_w$ as the abstract network state space of each WSP $w \in \mathcal{W}$, where $\mathcal{S}_w = \{1, \ldots, |\mathcal{S}_w|\}$ is an abstraction of the state space $\mathcal{X}_{-w}$ of other WSPs.

Definition 2. *For each WSP $w \in \mathcal{W}$ in the abstract stochastic game $\hat{\mathcal{G}}$, the abstract states regarding other competing WSPs are constructed according to a surjective aggregation function $l_w : \mathcal{X}_{-w} \rightarrow \mathcal{S}_w$. This implies,*

$$\mathcal{S}_w = \{ l_w(\boldsymbol{\chi}_{-w}) : \boldsymbol{\chi}_{-w} \in \mathcal{X}_{-w} \}, \forall w \in \mathcal{W}. \tag{21}$$

Remark 3. The NE of the abstract game $\hat{\mathcal{G}}$ can be matched back to the original stochastic game $\mathcal{G}$ given the surjective aggregation functions for all WSPs [33]. But the mapping from $\mathcal{X}_{-w}$ to $\mathcal{S}_w$ in (21), $\forall w \in \mathcal{W}$, requires the complete local network dynamics from the competitors. Moreover, the state-of-art state abstraction algorithms [32], which are based on various similarity criteria of defining the aggregation functions, are NP-complete [33].

Developing algorithms of constructing the surjective aggregation functions for such an abstract game requires the WSPs to exchange the local network dynamics, which is daunting for our non-cooperative scenario. On the other hand, the environmental feedbacks received by the WSPs (i.e., the payment (11) or the payoff (8)) also imply the behavioral couplings in the game $\mathcal{G}$. We hence propose that each WSP $w \in \mathcal{W}$ in the abstract game $\hat{\mathcal{G}}$ builds an internal $\mathcal{S}_w$ by classifying the value intervals of the payments (or the payoffs[6]), the procedure of which will be explained in Section 5.1. Herein, $\mathcal{S}_w$ can be treated as an approximation of $\mathcal{X}_{-w}$ with $|\mathcal{S}_w| \ll |\mathcal{X}_{-w}|$. In line with the discussions, the WSPs' self-organizing behaviours are then not constrained by the form of l_w as well as the $\mathcal{X}_{-w}$. We will see in Section 4.3 that the statistics of $\mathcal{S}_w$ can be learned from the history of the game $\hat{\mathcal{G}}$. Based on the abstract network state space $\hat{\mathcal{X}}_w$, we have the abstract action space $\hat{\mathcal{Y}}_w$, and the payoff function $\hat{F}_w$ in game $\hat{\mathcal{G}}$ is accordingly defined over $\hat{\mathcal{X}}_w$ and $\hat{\mathcal{Y}}_w$ for each WSP w. Let $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_w : w \in \mathcal{W})$ be a stationary joint control policy employed by the WSPs in the abstract game $\hat{\mathcal{G}}$, where $\hat{\boldsymbol{\pi}}_w = (\hat{\pi}_w^{(c)}, \boldsymbol{\pi}_w^{(p)})$ with $\hat{\pi}_w^{(c)}$ being the abstract channel auction policy. Likewise, the abstract state value function for WSP w under $\hat{\boldsymbol{\pi}}$ can be expressed as (22), $\forall \hat{\boldsymbol{\chi}}_w = (\boldsymbol{\chi}_w, s_w) \in \hat{\mathcal{X}}_w$ with $s_w \in \mathcal{S}_w$,

6. The two criteria are equivalent since the only difference is the exact utility values from the MTs which can be obtained by their subscribing WSPs.

$$\hat{V}_w(\hat{\boldsymbol{\chi}}_w, \hat{\boldsymbol{\pi}}) = (1-\gamma)\mathsf{E}\left[\sum_{t=1}^{\infty}(\gamma)^{t-1}\hat{F}_w\Big(\hat{\boldsymbol{\chi}}_w^t, \boldsymbol{\varphi}_w\Big(\hat{\boldsymbol{\pi}}^{(c)}(\hat{\boldsymbol{\chi}}^t)\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w^t)\Big)|\hat{\boldsymbol{\chi}}_w^1 = \hat{\boldsymbol{\chi}}_w\right], \tag{22}$$

where $\hat{\boldsymbol{\pi}}^{(c)}(\hat{\boldsymbol{\chi}}^t) = (\hat{\pi}_w^{(c)}(\hat{\boldsymbol{\chi}}_w^t) : w \in \mathcal{W})$, $\hat{\boldsymbol{\pi}}^{(c)} = (\hat{\pi}_w^{(c)} : w \in \mathcal{W})$ is the joint abstract channel auction policy, and $\hat{\boldsymbol{\chi}}^t = (\hat{\boldsymbol{\chi}}_w^t : w \in \mathcal{W})$ with each $\hat{\boldsymbol{\chi}}_w^t$ being the abstract network state of WSP w at slot t. Fig. 2 shows the similarity and difference between the games $\mathcal{G}$ and $\hat{\mathcal{G}}$.

For the purpose of theoretical analysis, we continue using l_w for WSP w as the hidden function mapping from $\mathcal{X}_{-w}$ to $\mathcal{S}_w$.

In the mapping from original stochastic game $\mathcal{G}$ to abstract stochastic game $\hat{\mathcal{G}}$, we denote $\varepsilon_w(s_w)$ as the length of the value interval that is associated with a state $s_w \in \mathcal{S}_w$ of each WSP $w \in \mathcal{W}$. We then have (23), where $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ are two matched control policies, respectively, in games $\mathcal{G}$ and $\hat{\mathcal{G}}$, and $\hat{\boldsymbol{\chi}} = (\hat{\boldsymbol{\chi}}_w : w \in \mathcal{W})$.

Definition 3. *Given the abstract joint control policy $\hat{\boldsymbol{\pi}}$ of all WSPs in the abstract game $\hat{\mathcal{G}}$, the matched joint control policy $\boldsymbol{\pi}$ in the original stochastic game $\mathcal{G}$ is a policy that satisfies*

$$\sum_{\hat{\boldsymbol{\chi}}'_w \in \hat{\mathcal{X}}_w} \mathsf{Pr}\Big\{\hat{\boldsymbol{\chi}}'_w | \hat{\boldsymbol{\chi}}_w, \boldsymbol{\varphi}_w\Big(\hat{\boldsymbol{\pi}}^{(c)}(\hat{\boldsymbol{\chi}})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big\} = \sum_{\boldsymbol{\chi}' \in \mathcal{X}^{|\mathcal{N}|}} \mathsf{Pr}\Big\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\varphi}\Big(\boldsymbol{\pi}^{(c)}(\boldsymbol{\chi})\Big), \boldsymbol{\pi}^{(p)}(\boldsymbol{\chi})\Big\}, \tag{24}$$

$\forall w \in \mathcal{W}$, where each $\hat{\boldsymbol{\chi}}_w = (\boldsymbol{\chi}_w, l_w(\boldsymbol{\chi}_{-w}))$.

Therefore, there exists an error bound of the payoffs for each WSP $w \in \mathcal{W}$, which can be given by

$$\varepsilon_w^{(max)} = \max_{s_w \in \mathcal{S}_w} \varepsilon_w(s_w). \tag{25}$$

If the value intervals are of equal length, then $\varepsilon_w(s_w) = \varepsilon_w^{(max)}$, $\forall s_w \in \mathcal{S}_w$. As indicated by the lemma below, the expected long-term payoff achieved by WSP w from a control policy $\hat{\boldsymbol{\pi}}_w$ in the abstract game $\hat{\mathcal{G}}$ is not far from that from the matched control policy $\boldsymbol{\pi}_w$ in the original game $\mathcal{G}$.

Lemma 1. *Given any two matched stationary control policies $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ in games $\mathcal{G}$ and $\hat{\mathcal{G}}$, we have*

$$\big|V_w(\boldsymbol{\chi}, \boldsymbol{\pi}) - \hat{V}_w(\hat{\boldsymbol{\chi}}_w, \hat{\boldsymbol{\pi}})\big| \le \varepsilon_w^{(max)}, \forall w \in \mathcal{W}, \tag{26}$$

$\forall \boldsymbol{\chi} \in \mathcal{X}^{|\mathcal{N}|}$, where $\hat{\boldsymbol{\chi}}_w = (\boldsymbol{\chi}_w, s_w)$ with $s_w = l_w(\boldsymbol{\chi}_{-w})$.

The proof of Lemma 1 is given in Appendix A, available in the online supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2017.2742949.

Based on the result in Lemma 1, Theorem 2 shows that the NE $\hat{\boldsymbol{\pi}}^* = (\hat{\boldsymbol{\pi}}_w^* : w \in \mathcal{W})$ in the abstract stochastic game $\hat{\mathcal{G}}$ leads to a bounded regret of playing the matched policy $\boldsymbol{\pi}^*$ in the original game $\mathcal{G}$. Here, $\hat{\boldsymbol{\pi}}_w^* = (\hat{\pi}_w^{(c),*}, \boldsymbol{\pi}_w^{(p),*})$ is the best-response abstract control policy for WSP w.

Theorem 2. *Let $\boldsymbol{\pi}_w$ be the control policy that is matched from an abstract control policy $\hat{\boldsymbol{\pi}}_w$, $\forall w \in \mathcal{W}$. $\forall \boldsymbol{\chi} \in \mathcal{X}^{|\mathcal{N}|}$, the corresponding policy $\boldsymbol{\pi}^*$ matched from $\hat{\boldsymbol{\pi}}^*$ satisfies*

$$V_w\big(\boldsymbol{\chi}, (\boldsymbol{\pi}_w, \boldsymbol{\pi}_{-w}^*)\big) \le \mathbb{V}_w(\boldsymbol{\chi}) + 2\varepsilon_w^{(max)}, \tag{27}$$

where $(\boldsymbol{\pi}_w, \boldsymbol{\pi}_{-w}^*)$ is the joint control policy in original stochastic game $\mathcal{G}$ that results from WSP w unilaterally deviating from $\boldsymbol{\pi}_w^*$ to $\boldsymbol{\pi}_w$.

The proof of Theorem 2 is given in Appendix B, available in the online supplemental material.

Hereinafter, we switch to focus on the abstract stochastic game $\hat{\mathcal{G}}$. Suppose all WSPs play the NE control policy $\hat{\boldsymbol{\pi}}^*$ in the abstract stochastic game $\hat{\mathcal{G}}$. Along with the discussions in previous sections, under the given expected long-term payoff functions $\hat{\mathbb{V}}_w(\hat{\boldsymbol{\chi}}_w) = \hat{V}_w(\hat{\boldsymbol{\chi}}_w, \hat{\boldsymbol{\pi}}^*)$, $\forall \hat{\boldsymbol{\chi}}_w \in \hat{\mathcal{X}}_w$ and $\forall w \in \mathcal{W}$, the best-response abstract control policy for a WSP w satisfies (28), $\forall \hat{\boldsymbol{\chi}}_w \in \hat{\mathcal{X}}_w$. From (28), WSP w is able to compute the optimal abstract control policy in the abstract game based on its local information.

Remark 4. Unfortunately, there are two main challenges involved in solving (28) for each WSP $w \in \mathcal{W}$: 1) a priori knowledge of abstract network state transition probability, which incorporates the statistics of channel state variations, the packet arrival distributions and the approximation of other competing WSPs' local network information (i.e., the statistics of $\mathcal{S}_w$), is not feasible; 2) given a classification of the payment/payoff values, the size of local abstract network state space $\hat{\mathcal{X}}_w$ grows exponentially as the number of subscribed MTs increases.

4.2 Decomposition of Abstract State-Value Function

From the facts underlying in (20): 1) the channel auction and the packet scheduling decisions, which are made sequentially, are independent across a WSP and the subscribed MTs; and 2) the weak interactions exist in the packet scheduling, the per-slot payoff function (8) of the WSP is separable into components of payment and per-MT utilities. The per-WSP MDP

$$\varepsilon_w(s_w) \ge \max_{\{\boldsymbol{\chi}, \hat{\boldsymbol{\pi}}_w(\hat{\boldsymbol{\chi}}_w): \boldsymbol{\chi}=(\boldsymbol{\chi}_w, \boldsymbol{\chi}_{-w}) \in \mathcal{X}^{|\mathcal{N}|}, \hat{\boldsymbol{\chi}}_w=(\boldsymbol{\chi}_w, s_w), s_w = l_w(\boldsymbol{\chi}_{-w})\}} \Big| F_w\Big(\boldsymbol{\chi}, \boldsymbol{\varphi}_w\Big(\boldsymbol{\pi}^{(c)}(\boldsymbol{\chi})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big) - \hat{F}_w\Big(\hat{\boldsymbol{\chi}}_w, \boldsymbol{\varphi}_w\Big(\hat{\boldsymbol{\pi}}^{(c)}(\hat{\boldsymbol{\chi}})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big)\Big| \tag{23}$$

$$\hat{\mathbb{V}}_w(\hat{\boldsymbol{\chi}}_w) = \max_{\hat{\boldsymbol{\pi}}_w(\hat{\boldsymbol{\chi}}_w)} \Bigg\{ (1-\gamma)\hat{F}_w\Big(\hat{\boldsymbol{\chi}}_w, \boldsymbol{\varphi}_w\Big(\hat{\boldsymbol{\pi}}_w^{(c)}(\hat{\boldsymbol{\chi}}_w), \hat{\boldsymbol{\pi}}_{-w}^{(c),*}(\hat{\boldsymbol{\chi}}_{-w})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big) + \gamma \sum_{\hat{\boldsymbol{\chi}}'_w \in \hat{\mathcal{X}}_w} \mathsf{Pr}\Big\{\hat{\boldsymbol{\chi}}'_w | \hat{\boldsymbol{\chi}}_w, \boldsymbol{\varphi}_w\Big(\hat{\boldsymbol{\pi}}_w^{(c)}(\hat{\boldsymbol{\chi}}_w), \hat{\boldsymbol{\pi}}_{-w}^{(c),*}(\hat{\boldsymbol{\chi}}_{-w})\Big), \boldsymbol{\pi}_w^{(p)}(\boldsymbol{\chi}_w)\Big\} \hat{\mathbb{V}}_w(\hat{\boldsymbol{\chi}}'_w) \Bigg\} \tag{28}$$

described by (28) can be hence decomposed into $|\mathcal{N}_w|+1$ independent single-agent MDPs. That is, $\forall w \in \mathcal{W}$, $\hat{\mathbb{V}}_w(\hat{\boldsymbol{\chi}}_w)$ can be computed as

$$\hat{\mathbb{V}}_w(\hat{\boldsymbol{\chi}}_w) = \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}_{k,w}} \alpha_{k,n}\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n}) - \mathbb{U}_w(s_w), \tag{29}$$

$\forall \hat{\boldsymbol{\chi}}_w \in \hat{\mathcal{X}}_w$, where the per-MT expected long-term utility $\mathbb{U}_{k,n}$ and the expected long-term payment $\mathbb{U}_w(s_w)$ of WSP w satisfy

$$\begin{aligned}\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n}) = \max_{\pi^{(p)}_{k,n}(\boldsymbol{\chi}_{k,n})} \Bigg\{ & (1-\gamma) U_{k,n}\Big(\boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi^{(p)}_{k,n}(\boldsymbol{\chi}_{k,n})\Big) \\ & + \gamma \sum_{\boldsymbol{\chi}'_{k,n}\in\mathcal{X}} \Pr\Big\{\boldsymbol{\chi}'_{k,n}|\boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi^{(p)}_{k,n}(\boldsymbol{\chi}_{k,n})\Big\} \mathbb{U}_{k,n}\Big(\boldsymbol{\chi}'_{k,n}\Big) \Bigg\},\end{aligned} \tag{30}$$

and

$$\begin{aligned}\mathbb{U}_w(s_w) = & (1-\gamma)\tau_w \\ & + \gamma \sum_{s'_w\in\mathcal{S}_w} \Pr\Big\{s'_w|s_w, \theta_w\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big)\Big\} \mathbb{U}_w(s'_w),\end{aligned} \tag{31}$$

respectively, with $\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}}) = (\hat{\pi}^{(c),*}_w(\hat{\boldsymbol{\chi}}_w) : w \in \mathcal{W})$. The winner determination and payment calculation during the centralized channel auction process at the CNC leads to the derivation of (31). Such a decomposition enables us to propose an online algorithm for learning the optimal abstract control policy $\hat{\boldsymbol{\pi}}^*$ in the following section. It is worth to note that the linear decomposition in (29) is different from the linear approximation technique applied in work [22], which ignores the coupling of channel auction among the participating agents.

Remark 5. The three key advantages of the linear decomposition approach in (29) are listed as follows.

(1) Low computational complexity: In order to deploy a control policy based on the abstract network state $\hat{\boldsymbol{\chi}}_w \in \hat{\mathcal{X}}_w$, each WSP $w \in \mathcal{W}$ has to store the abstract state value function with $|\mathcal{X}|^{|\mathcal{N}_w|} \cdot |\mathcal{S}_w|$ values. Using the linear decomposition, only $|\mathcal{N}_w| \cdot |\mathcal{X}| + |\mathcal{S}_w|$ values need to be stored. Moreover, the channel auction and the packet scheduling decision makings are simplified and depend on the limited feasible information at the WSP and the MTs, respectively.

(2) Low signalling overhead: The linear decomposition motivates the WSPs to let their MTs locally store the individual state value functions $\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n})$, $\forall k \in \mathcal{K}$, $\forall n \in \mathcal{N}_k$ and $\forall \boldsymbol{\chi}_{k,n} \in \mathcal{X}$, which alleviates the signalling of channel state and queue state information between the MTs and the WSPs. The values of $\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n})$ are then notified to the WSPs when deciding the true valuation during the channel auction. On the other hand, the WSPs only keep the expected long-term payment values $\mathbb{U}_w(s_w)$, $\forall w \in \mathcal{W}$ and $\forall s_w \in \mathcal{S}_w$.

(3) Near optimality: The solving of a complex Bellman's Equation (28) is broken into much simpler MDPs. The linear decomposition approach is a special case of the feature-based decomposition method [34], but provides an accuracy guarantee of the approximation of abstract state value function, the research of which has been extensively studied in the literature of reinforcement learning [23], [34, Theorem 2].

Algorithm 1. VCG Based Channel Scheduling at the CNC at the Beginning of Each Scheduling Slot t

1: At current scheduling slot t, each MT $n \in \mathcal{N}_{k,w}$ ($k \in \mathcal{K}$) forwards the private information of $[\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n}), i_{k,n}]$ to its WSP $w \in \mathcal{W}$, where $i_{k,n}$ is given by (34).
2: According to the best-response channel auction policy $\pi^{(c),*}_w$, WSP w submits its bidding vector $\boldsymbol{\beta}_w = [\nu_w, \boldsymbol{o}_w]$ to the CNC to report its true valuation on requesting channels, where ν_w is given by (33) and $\boldsymbol{o}_w = [o_{k,w} : k \in \mathcal{K}]$ with each $o_{k,w}$ given by (32).
3: After collecting bids announced by all WSPs, the CNC determines the auction winners $\boldsymbol{\theta}$ and the channel scheduling $\boldsymbol{\rho}_w = [\rho_{k,n} : k \in \mathcal{K}, n \in \mathcal{N}_{k,w}]$ according to (10) to maximize the sum valuations, calculates the payments τ_w according to (11) for each WSP w, and then feeds back $\{(\theta_w, \boldsymbol{\rho}_w, \tau_w) : w \in \mathcal{W}\}$ to all WSPs.

As addressed in Section 2, each WSP $w \in \mathcal{W}$ has a valuation on requesting a number of channels for its MTs at each scheduling slot, which is assumed to be strategic in the channel auction. By decomposing the abstract state value function (29), we can now define the number of requested channels in each cell $k \in \mathcal{K}$ as

$$o_{k,w} = \sum_{n\in\mathcal{N}_{k,w}} i_{k,n}, \tag{32}$$

and the true valuation of obtaining $\boldsymbol{o}_w = [o_{k,w} : k \in \mathcal{K}]$ as

$$\begin{aligned}\nu_w = & \frac{1}{1-\gamma} \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}_{k,w}} \alpha_{k,n}\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n}) \\ & - \frac{\gamma}{1-\gamma} \sum_{s'_w\in\mathcal{S}_w} \Pr\Big\{s'_w|s_w, \mathbb{1}_{\left\{\sum_{k\in\mathcal{K}} o_{k,w}>0\right\}}\Big\} \mathbb{U}_w(s'_w),\end{aligned} \tag{33}$$

which together constitute the bid $\boldsymbol{\beta}_w = [\nu_w, \boldsymbol{o}_w]$ of WSP w under the abstract network state $\hat{\boldsymbol{\chi}}_w \in \hat{\mathcal{X}}_w$ at current scheduling slot, where

$$\begin{aligned}i_{k,n} = \arg\max_{i\in\{0,1\}} \Bigg\{ & (1-\gamma) U_{k,n}\Big(\boldsymbol{\chi}_{k,n}, i, \pi^{(p),*}_{k,n}(\boldsymbol{\chi}_{k,n})\Big) \\ & + \gamma \sum_{\boldsymbol{\chi}'_{k,n}\in\mathcal{X}} \Pr\Big\{\boldsymbol{\chi}'_{k,n}|\boldsymbol{\chi}_{k,n}, i, \pi^{(p),*}_{k,n}(\boldsymbol{\chi}_{k,n})\Big\} \mathbb{U}_{k,n}\Big(\boldsymbol{\chi}'_{k,n}\Big) \Bigg\},\end{aligned} \tag{34}$$

indicates the preference of a MT $n \in \mathcal{N}_{k,w}$ ($k \in \mathcal{K}$) between obtaining one channel or not, and $\mathbb{1}_{\{\Omega\}}$ is an indicator function that equals 1 if the condition Ω is satisfied and 0 otherwise. Such a channel auction decision from WSP w cares about not only the immediate revenue from charging the MTs, but the payoff realizations from the future interactions with other competitors. Note that when determining the bid $\boldsymbol{\beta}_w$, the private information $[\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n}), i_{k,n}]$ at each MT n needs to be transferred to its subscribing WSP w. However, it's clear that the auction bids can be computed independently at the WSPs. With the definition of auction bids in (33) and (32), we briefly present in Algorithm 1 the design of the VCG auction based channel scheduling at the CNC at the beginning of each scheduling slot t.

4.3 Learning Optimal Abstract Control Policy

The optimal channel auction and packet scheduling decisions depend on both the expected long-term payments of WSPs and the per-MT expected long-term utilities.

Additionally, in the calculation of true valuation (33) for a WSP $w \in \mathcal{W}$ at each scheduling slot t, the state transition probability, which is used to forecast the value of expected future payments, is unknown. We propose that each WSP w maintains over the slots a three-dimensional table B_w^t with size $|\mathcal{S}_w| \cdot |\mathcal{S}_w| \cdot 2$. Each entry $b_{\ell,\ell',\iota}^t$ in the table B_w^t represents the number of transitions from $s_w^t = \ell$ to $s_w^{t+1} = \ell'$ when $\theta_w^t = \iota - 1$, where $\ell, \ell' \in \mathcal{S}_w$ and $\iota \in \{1, 2\}$. The update of B_w^t is simply based on the observations of the channel auction results. Then, the state transition probability at slot t can be estimated as[7]

$$\Pr\{s_w^{t+1} = \ell' | s_w^t = \ell, \theta_w^t = \iota - 1\} = \frac{b_{\ell,\ell',\iota}^t}{\sum_{\ell'' \in \mathcal{S}_w} b_{\ell'',\ell',\iota}^t}. \tag{35}$$

Using the union bound and the weak law of large numbers [35] in our considered stationary networking environment, we have $\forall \ell, \ell' \in \mathcal{S}_w$ and $\forall \iota \in \{1, 2\}$,

$$\begin{aligned} \lim_{t \to \infty} \Pr\Big\{ \Big| \Pr\{s_w^{t+2} = \ell' | s_w^{t+1} = \ell, \theta_w^{t+1} = \iota - 1\} \\ - \Pr\{s_w^{t+1} = \ell' | s_w^t = \ell, \theta_w^t = \iota - 1\} \Big| > \vartheta \Big\} = 0, \end{aligned} \tag{36}$$

for an arbitrarily small constant $\vartheta \in \mathrm{IR}_+$. And the values of $\mathbb{U}_w(s_w)$, $\forall s_w \in \mathcal{S}_w$, are updated according to the reinforcement learning rule,

$$\mathbb{U}_w^{t+1}(s_w) = \begin{cases} (1-\zeta^t)\mathbb{U}_w^t(s_w) + \zeta^t\big((1-\gamma)\tau_w^t \\ +\gamma \sum_{s_w^{t+1} \in \mathcal{S}_w} \Pr\{s_w^{t+1} | s_w, \theta_w^t\} \mathbb{U}_w^t\big(s_w^{t+1}\big)\big), & \text{if } s_w = s_w^t; \\ \mathbb{U}_w^t(s_w), & \text{otherwise}, \end{cases} \tag{37}$$

where $\zeta^t \in [0, 1)$ is the learning rate and the convergence of the learning rule is guaranteed by $\sum_{t=1}^{\infty} \zeta^t = \infty$ and $\sum_{t=1}^{\infty} (\zeta^t)^2 < \infty$ [23].

Given that all WSPs deploy the best-response channel auction policies, the well-known value iteration [23] can be used by the MTs to find the optimal state value functions (30). However, this method requires full knowledge of the local network state transition probabilities, which is challenging without a priori statistical information of channel state transitions and packet arrivals. To tackle this challenge, we define a post-decision state [36], [37] based on the observation that the new packet arrivals are independent of the channel auction and the packet scheduling decision makings. At current scheduling slot, the post-decision state of a MT $n \in \mathcal{N}_{k,w}$ ($\forall k \in \mathcal{K}$ and $\forall w \in \mathcal{W}$) is defined by $\tilde{\boldsymbol{\chi}}_{k,n} = (\tilde{H}_{k,n}, \tilde{Q}_{k,n}) \in \mathcal{X}$, where $\tilde{H}_{k,n} = H_{k,n}$ and $\tilde{Q}_{k,n} = Q_{k,n} - \varphi_{k,n}\,(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}}))\pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})$. By introducing a post-decision state, we ?A3B2 tlb?> are able to factor the utility function in (9) into two parts, which correspond to $\phi_{k,n}U_{k,n}^{(1)}(\cdot) + U_{k,n}^{(2)}(\cdot)$ and $U_{k,n}^{(3)}(\cdot)$. The probability of state transition from $\boldsymbol{\chi}_{k,n}$ to $\boldsymbol{\chi}'_{k,n}$ can be then expressed as

$$\begin{aligned} &\Pr\Big\{\boldsymbol{\chi}'_{k,n} | \boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})\Big\} \\ &= \Pr\Big\{\boldsymbol{\chi}'_{k,n} | \tilde{\boldsymbol{\chi}}_{k,n}\Big\} \Pr\Big\{\tilde{\boldsymbol{\chi}}_{k,n} | \boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})\Big\} \\ &= \Pr\Big\{H'_{k,n}\Big\} \Pr\Big\{Q'_{k,n} - \tilde{Q}_{k,n}\Big\}, \end{aligned} \tag{38}$$

where $\Pr\{\tilde{\boldsymbol{\chi}}_{k,n} | \boldsymbol{\chi}_{k,n}, \varphi_{k,n}(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})), \pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})\} = 1$. Let $\tilde{\mathbb{U}}_{k,n}(\tilde{\boldsymbol{\chi}}_{k,n})$ be the per-MT optimal post-decision state value function given by

$$\begin{aligned} \tilde{\mathbb{U}}_{k,n}(\tilde{\boldsymbol{\chi}}_{k,n}) = (1-\gamma) U_{k,n}^{(3)}\Big(\boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi_{k,n}^{(p),*}(\boldsymbol{\chi}_{k,n})\Big) \\ + \gamma \sum_{\boldsymbol{\chi}'_{k,n} \in \mathcal{X}} \Pr\Big\{\boldsymbol{\chi}'_{k,n} | \tilde{\boldsymbol{\chi}}_{k,n}\Big\} \mathbb{U}_{k,n}\Big(\boldsymbol{\chi}'_{k,n}\Big). \end{aligned} \tag{39}$$

Then (30) becomes,

$$\begin{aligned} &\mathbb{U}_{k,n}(\boldsymbol{\chi}_{k,n}) = \max_{\pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})} \\ &\Big\{(1-\gamma)\Big(\phi_{k,n} U_{k,n}^{(1)}\Big(\boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})\Big) \\ &+ U_{k,n}^{(2)}\Big(\boldsymbol{\chi}_{k,n}, \varphi_{k,n}\Big(\hat{\boldsymbol{\pi}}^{(c),*}(\hat{\boldsymbol{\chi}})\Big), \pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n})\Big)\Big) + \tilde{\mathbb{U}}_{k,n}(\tilde{\boldsymbol{\chi}}_{k,n})\Big\}. \end{aligned} \tag{40}$$

From (40), we find that the per-MT optimal state value function can be directly obtained from the per-MT optimal post-decision state value function by performing maximization over all feasible packet scheduling decisions.

As we know, the number of new packet arrivals in the end of a scheduling slot is unavailable beforehand and so is the number of packet drops at the slot. In this case, instead of directly computing the optimal post-decision state value function as in (39), we propose an online learning algorithm to approach the optimal post-decision state value function by exploring the conventional reinforcement learning techniques [23], [37]. Based on the observations of the local network state $\boldsymbol{\chi}_{k,n}^t$, the channel scheduling result $\boldsymbol{\rho}_{k,n}^t$ by the CNC, the number of packet departures $\varphi_{k,n}^t R_{k,n}^t$, the number of packet arrivals $A_{k,n}^t$, the number of packet drops $\max\{Q_{k,n}^t - \varphi_{k,n}^t R_{k,n}^t + A_{k,n}^t - Q^{(\max)}, 0\}$ at current scheduling slot t and the resulting local network state $\boldsymbol{\chi}_{k,n}^{t+1}$ at next slot $t+1$, each MT $n \in \mathcal{N}_{k,w}$ ($\forall k \in \mathcal{K}$ and $\forall w \in \mathcal{W}$) updates the post-decision state value function on the fly,

$$\begin{aligned} \tilde{\mathbb{U}}_{k,n}^{t+1}\Big(\tilde{\boldsymbol{\chi}}_{k,n}^t\Big) = \big(1-\zeta^t\big)\tilde{\mathbb{U}}_n^t\Big(\tilde{\boldsymbol{\chi}}_{k,n}^t\Big) \\ + \zeta^t\Big((1-\gamma) U_{k,n}^{(3)}\Big(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, R_{k,n}^t\Big) + \gamma \mathbb{U}_{k,n}^t\Big(\boldsymbol{\chi}_{k,n}^{t+1}\Big)\Big), \end{aligned} \tag{41}$$

where the number of packets to be delivered, $R_{k,n}^t$, during scheduling slot t is determined as

$$\begin{aligned} R_{k,n}^t = \arg\max_{\pi_{k,n}^{(p)}(\boldsymbol{\chi}_{k,n}^t)} \Big\{(1-\gamma)\Big(\phi_{k,n} U_{k,n}^{(1)}\Big(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, \pi_{k,n}^{(p)}\Big(\boldsymbol{\chi}_{k,n}^t\Big)\Big) \\ + U_{k,n}^{(2)}\Big(\boldsymbol{\chi}_{k,n}^t, \varphi_{k,n}^t, \pi_{k,n}^{(p)}\Big(\boldsymbol{\chi}_{k,n}^t\Big)\Big)\Big) + \tilde{\mathbb{U}}_{k,n}^t\Big(\tilde{\boldsymbol{\chi}}_{k,n}^t\Big)\Big\}, \end{aligned} \tag{42}$$

and the value of the local network state $\boldsymbol{\chi}_{k,n}^{t+1}$ at next slot $t+1$ is evaluated by (43). Therefore, the true valuation in (33) of a WSP w at the beginning of each slot t during the learning process can be replaced by

$$\begin{aligned} \nu_w^t = \frac{1}{1-\gamma} \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}_{k,w}} \alpha_{k,n} \mathbb{U}_{k,n}^t\Big(\boldsymbol{\chi}_{k,n}^t\Big) \\ - \frac{\gamma}{1-\gamma} \sum_{s_w^{t+1} \in \mathcal{S}_w} \Pr\Big\{s_w^{t+1} | s_w^t, \mathbb{1}_{\left\{\sum_{k \in \mathcal{K}} o_{k,w}^t > 0\right\}}\Big\} \mathbb{U}_w^t\big(s_w^{t+1}\big), \end{aligned} \tag{44}$$

7. To avoid division by zero, each entry in a table B_w^1, $\forall w \in \mathcal{W}$, needs to be initialized, for example, to 1 in simulations.

where $o^t_{k,w} = \sum_{n\in\mathcal{N}_{k,w}} i^t_{k,n}$ with

$$i^t_{k,n} = \arg\max_{i\in\{0,1\}} \left\{ \max_{\pi^{(p)}_{k,n}(\boldsymbol{\chi}^t_{k,n})} \left\{ (1-\gamma)\left(\phi_{k,n} U^{(1)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}, i, \pi^{(p)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}\right)\right)\right.\right.\right.$$
$$\left.\left.\left. + U^{(2)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}, i, \pi^{(p)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}\right)\right)\right) + \tilde{\mathbb{U}}^t_{k,n}\left(\tilde{\boldsymbol{\chi}}^t_{k,n}\right)\right\}\right\}. \quad (45)$$

The online localized algorithm for learning the optimal post-decision state value functions of a MT $n \in \mathcal{N}_{k,w}$ ($\forall k \in \mathcal{K}$ and $\forall w \in \mathcal{W}$) in the network is summarized in Algorithm 2. And Theorem 3 ensures that the online learning algorithm converges.

Algorithm 2. Online Algorithm for Learning Optimal Post-decision State Value Functions of a MT $n \in \mathcal{N}_{k,w}$ in a cell $k \in \mathcal{K}$ subscribed to a WSP $w \in \mathcal{W}$

1: **initialize** the post-decision state value functions for MT n, $\tilde{\mathbb{U}}^1_{k,n}(\tilde{\boldsymbol{\chi}}_{k,n}), \forall \tilde{\boldsymbol{\chi}}_{k,n} \in \mathcal{X}$.
2: **repeat**
3: At the beginning of a scheduling slot t, MT n observes the local network state $\boldsymbol{\chi}^t_{k,n}$, calculate $\mathbb{U}^t_{k,n}(\boldsymbol{\chi}^t_{k,n})$ according to (43) and $i^t_{k,n}$ according to (45), and sends the information of $[\mathbb{U}^t_{k,n}(\boldsymbol{\chi}^t_{k,n}), i^t_{k,n}]$ to the subscribing WSP w.
4: MT n and WSP w await, respectively, the channel allocation $\boldsymbol{\rho}^t_{k,n}$ and the payment τ^t_w, which are from Algorithm 1. Then WSP w updates the matrix B^t_w and the $\mathbb{U}^{t+1}_w(s^t_w)$ according to (37), and MT n makes packet scheduling decision $R^t_{k,n}$ according to (42).
5: After transmitting the scheduled packets, MT n observes the post-decision state $\tilde{\boldsymbol{\chi}}^t_{k,n} = (H^t_{k,n}, Q^t_{k,n} - \varphi^t_{k,n} R^t_{k,n})$, the realized utility $U^{(3)}_{k,n}(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n})$ regarding the packet drops at scheduling slot t, and the new local network state $\boldsymbol{\chi}^{t+1}_{k,n} = (\tilde{Q}^t_{k,n} + A^t_{k,n}, H^{t+1}_{k,n})$ at the following scheduling slot $t+1$.
6: According to (43) and (41), MT n calculates the state value function $\mathbb{U}^t_{k,n}(\boldsymbol{\chi}^{t+1}_{k,n})$ and updates the post-decision state value function $\tilde{\mathbb{U}}^{t+1}_{k,n}(\tilde{\boldsymbol{\chi}}^t_{k,n})$, respectively.
7: The scheduling slot index is updated by $t \leftarrow t+1$.
8: **until** A predefined stopping condition is satisfied.

Theorem 3. *For each MT $n \in \mathcal{N}_{k,w}$ subscribed to a WSP $w \in \mathcal{W}$ in a cell $k \in \mathcal{K}$, the sequence $\{\tilde{\mathbb{U}}^t_{k,n}(\tilde{\boldsymbol{\chi}}_{k,n}) : \forall t \in \mathbb{N}_+\}$ by Algorithm 2 converges to the optimal post-decision state value function $\tilde{\mathbb{U}}_{k,n}(\tilde{\boldsymbol{\chi}}_{k,n}), \forall \tilde{\boldsymbol{\chi}}_{k,n} \in \mathcal{X}$, if and only if the learning rate ζ^t satisfies: $\sum_{t=1}^{\infty} \zeta^t = \infty$ and $\sum_{t=1}^{\infty} (\zeta^t)^2 < \infty$.*

The proof of Theorem 3 is given in Appendix C, available in the online supplemental material.

5 Numerical Results

This section proceeds to quantify the performance of our proposed wireless resource scheduling scheme in a Matlab based simulation environment. When implementing the scheme, Algorithm 2 learns the optimal post-decision state value functions for the MTs across the time horizon, while the channel auction at the beginning of each scheduling slot is conducted according to Algorithm 1.

5.1 General Setup

We compare the achieved performance of MTs from the proposed scheme with the following baselines:

(1) Channel-aware control policy—A MT evaluates the need of having one channel for packet transmissions based on the channel state at each scheduling slot and does not take into account the queue status.
(2) Queue-aware control policy—A MT informs its subscribing WSP at each scheduling slot the preference between obtaining one channel or not by considering maximizing the expected long-term number of packets to be transmitted [21].
(3) Random control policy—This policy randomly generates the values of having one channel or not for a MT at each scheduling slot, and submits to the WSP the preference with the larger value, which means that the random policy does not consider any dynamics in the network.

The interactions among the competing WSPs can be reflected by the payments that are paid by the WSPs to the CNC at each scheduling slot. Hence, we formulate the abstract state space $\mathcal{S}_w$ for each WSP $w \in \mathcal{W}$ similarly as in [21] by classifying the value interval of the received payment τ^t_w. Specifically, we split the range of $[0, \Gamma_w]$ into $[\Gamma_{w,-1}, \Gamma_{w,0}]$, $(\Gamma_{w,0}, \Gamma_{w,1}]$, $(\Gamma_{w,1}, \Gamma_{w,2}], \ldots, (\Gamma_{w,|\mathcal{S}_w|-2}, \Gamma_{w,|\mathcal{S}_w|-1}]$, where $\Gamma_{w,|\mathcal{S}_w|-1} = \Gamma_w$ is the maximum value of payment for WSP w and we let $\Gamma_{w,-1} = \Gamma_{w,0} = 0$ for a special case in which WSP w wins the channel auction but pays nothing. As an example, if $\theta^t_w = 1$ and the payment τ^t_w of WSP w at slot t is within the range $(\Gamma_{w,\ell-2}, \Gamma_{w,\ell-1}]$, then $s^{t+1}_w = \ell$. Such a splitting can be locally done by each WSP w based on the channel auction results and does not rely on the aggregation function l_w. To optimize the transmit power, buffer delay and packet drops, the $U^{(1)}_{k,n}(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n})$, $U^{(2)}_{k,n}(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n})$ and $U^{(3)}_{k,n}(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n})$ in the positive utility function of a MT $n \in \mathcal{N}_{k,w}$ ($k \in \mathcal{K}$) are chosen to be the exponential functions, namely,

$$U^{(1)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n}\right) = e^{-P^t_{k,n}}, \quad (46)$$

$$U^{(2)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n}\right) = e^{-Q^t_{k,n}}, \quad (47)$$

$$U^{(3)}_{k,n}\left(\boldsymbol{\chi}^t_{k,n}, \varphi^t_{k,n}, R^t_{k,n}\right) = e^{-\eta^t_{k,n}}. \quad (48)$$

For simulation purpose, we assume Rayleigh channels for the MTs across the scheduling slots. At each slot t, the channel state over the spectrum band is an exponentially distributed random variable with a mean of $\overline{H}$ (dB). The value of $H^t_{k,n}$ is determined as in [40] with a common channel state space of discrete values, i.e., $\mathcal{H} = \{-18.82, -13.79,$

$$\mathbb{U}^t_{k,n}\left(\boldsymbol{\chi}^{t+1}_{k,n}\right) = \max_{\pi^{(p)}_{k,n}(\boldsymbol{\chi}^{t+1}_{k,n})} \left\{ (1-\gamma)\left(\phi_{k,n} U^{(1)}_{k,n}\left(\boldsymbol{\chi}^{t+1}_{k,n}, \varphi_{k,n}\left(\hat{\boldsymbol{\pi}}^{(c),*}\left(\hat{\boldsymbol{\chi}}^{t+1}\right)\right), \pi^{(p)}_{k,n}\left(\boldsymbol{\chi}^{t+1}_{k,n}\right)\right)\right.\right.$$
$$\left.\left. + U^{(2)}_{k,n}\left(\boldsymbol{\chi}^{t+1}_{k,n}, \varphi_{k,n}\left(\hat{\boldsymbol{\pi}}^{(c),*}\left(\hat{\boldsymbol{\chi}}^{t+1}\right)\right), \pi^{(p)}_{k,n}\left(\boldsymbol{\chi}^{t+1}_{k,n}\right)\right)\right) + \tilde{\mathbb{U}}^t_{k,n}\left(\tilde{\boldsymbol{\chi}}^{t+1}_{k,n}\right)\right\} \quad (43)$$

TABLE 2
Simulation Parameters

Parameter	Value
Set of WSPs $\mathcal{W}$	$\{1, 2, 3\}$
Set of cells $\mathcal{K}$	$\{1, 2, 3\}$
Number of MTs $\lvert\mathcal{N}_{k,w}\rvert$	$3, \forall k \in \mathcal{K}, \forall w \in \mathcal{W}$
Channel bandwidth M	500 KHz
Noise power spectral density ς^2	2×10^{-10} W/Hz
Scheduling slot duration δ	10^{-2} second
Discount factor γ	0.9
Utility price $\alpha_{k,n}$	$1, \forall n \in \mathcal{N}_k, \forall k \in \mathcal{K}$
Packet size μ	5,000 bits
Maximum transmit power $P^{(max)}$	3 Watts
Weight of transmit power $\phi_{k,n}$	$6, \forall n \in \mathcal{N}_k, \forall k \in \mathcal{K}$
Maximum queue length $Q^{(max)}$	10 packets

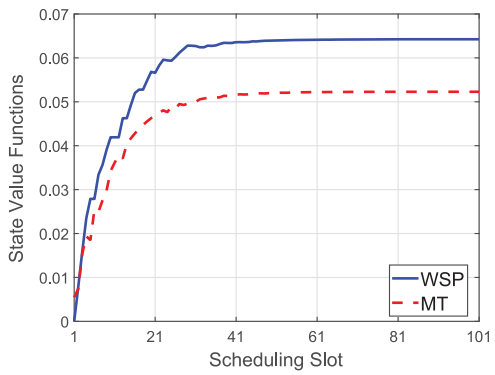

Fig. 3. Illustration for the convergence property of the proposed learning based scheme.

$-11.23, -9.37, -7.8, -6.3, -4.68, -2.08\}$ (dB). The number of arriving packets at a MT follows a Poisson arrival process, the average arrival rate of which is denoted by λ (packets per scheduling slot). Other general parameter values used in simulations are listed in Table 2.

5.2 Experiment Results

We perform experiments under different settings to validate the theoretical analysis carried out in this paper.

5.2.1 *Experiment 1—Convergence of the Proposed Wireless Scheduling Scheme*

In this experiment, our primary goal is to validate the convergence property of our proposed learning based wireless scheduling scheme. We assume that there are totally $J = 20$ channels shared by the MTs belonging to different WSPs in the network. At each scheduling slot, the packet arrivals for each MT are generated following a Poisson arrival process with average arrival rate $\lambda = 3$ (packets per scheduling slot). We fix the mean of channel states to be $\overline{H} = -2$ dB for all MTs. Without loss of the generality, we plot the simulated variations in state value functions, $\mathbb{U}_1(10)$ (in simulations, $|\mathcal{S}_w| = 55$, $\forall w \in \mathcal{W}$) and $\tilde{\mathbb{U}}_{1,1}((-11.23, 5))$ $((-11.23, 5) \in \mathcal{X})$, for WSP 1 and MT $1 \in \mathcal{N}_{1,1} = \{1, 2, 3\}$ versus the slots in Fig. 3, which reveals that the proposed scheme converges at a reasonably quick speed.

5.2.2 *Experiment 2—Performance with Changing Packet Arrival Rates*

This experiment tries to demonstrate the per-MT performance in terms of the average transmit power, the average queue length, the average packet drops and the average utility with different packet arrival rate settings. We choose for all MTs the mean of channel states as $\overline{H} = -6$ dB. The values of other simulation parameters are the same as in Experiment 1. The results are exhibited in Figs. 4, 5, 6 and 7. Fig. 4 illustrates the average transmit power per MT during the whole learning process, where the transmit power of each MT at each scheduling slot is defined in (7). Figs. 5 and 6 illustrate the average queue length and the average number of packet drops per MT, respectively. And Fig. 7 illustrates the achieved average utility per MT.

Each plot compares the performance of the proposed learning based scheme to the three baseline schemes. It can be observed from Fig. 7 that the proposed scheme achieves a significant gain in average utility. Similar observations can be made from the average transmit power per MT in

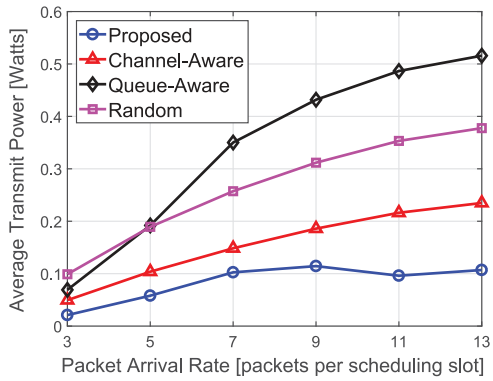

Fig. 4. Average transmit power per MT across the learning procedure versus average packet arrival rates.

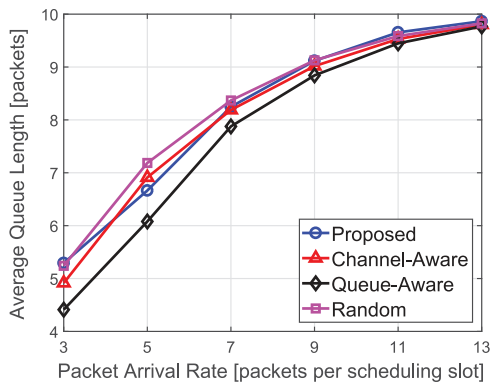

Fig. 5. Average queue length per MT across the learning procedure versus average packet arrival rates.

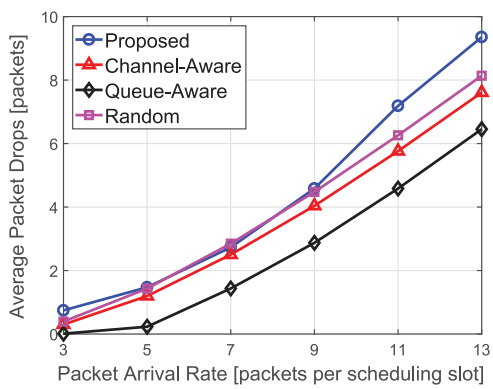

Fig. 6. Average packet drops per MT across the learning procedure versus average packet arrival rates.

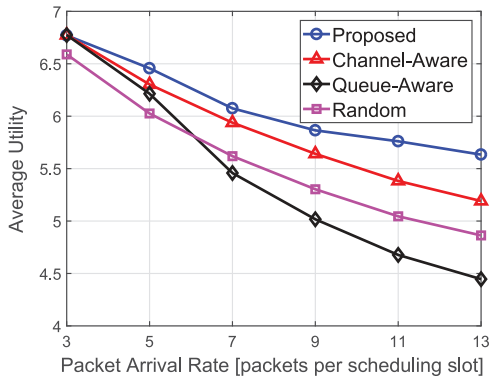

Fig. 7. Average utility per MT across the learning procedure versus average packet arrival rates.

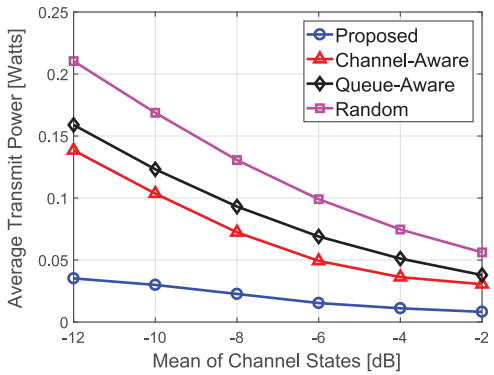

Fig. 8. Average transmit power per MT across the learning procedure versus means of channel states.

Fig. 4, though the average queue length and the average packet drops per MT with all four schemes are comparable. This can be explained by the reason that with the network settings by changing the average packet arrival rate in this experiment, the value of $U_{k,n}^{(1)}(\cdot)$ for each MT $n \in \mathcal{N}_{k,w}, \forall k \in \mathcal{K}$ and $\forall w \in \mathcal{W}$, dominates the utility function $U_{k,n}(\cdot)$. As the average packet arrival rate increases, the average utility performances decrease. Since there are not enough channels for all MTs (i.e., $J < |\mathcal{N}|$) during one scheduling slot, there might be the case that only some of the WSPs win the channel access opportunities from the channel auction, which gives rises to the number of packets remaining in the queue and the number of packets being dropped. Even if allocated a channel, a MT need to schedule more queued packets for transmission in order to reduce both the queue delay and the packet drops, but at a sacrifice of higher transmit power.

5.2.3 *Experiment 3—Performance under Various Means of Channel States*

We do this experiment to simulate the per-MT performance achieved from the proposed learning based scheme and other three baseline schemes versus the means of channel states. We configure the parameter values used in simulation as follows: $J = 30$ and $\lambda = 3$ (packets per scheduling slot). The average transmit power, the average queue length, the average packet drops and the average utility per MT across the entire learning period are depicted in Figs. 8, 9, 10 and 11. From the figures, we can clearly see that with a better channel condition on average, less transmit power on average is needed for packet transmissions. The reason behind this is straightforward, that is, since $J > |\mathcal{N}|$, all

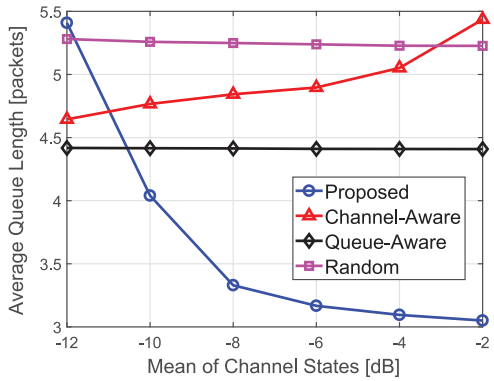

Fig. 9. Average queue length per MT across the learning procedure versus means of channel states.

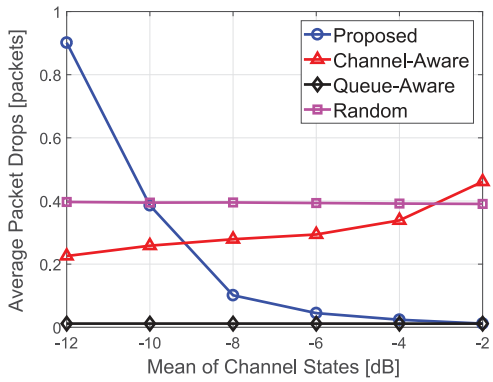

Fig. 10. Average packet drops per MT across the learning procedure versus means of channel states.

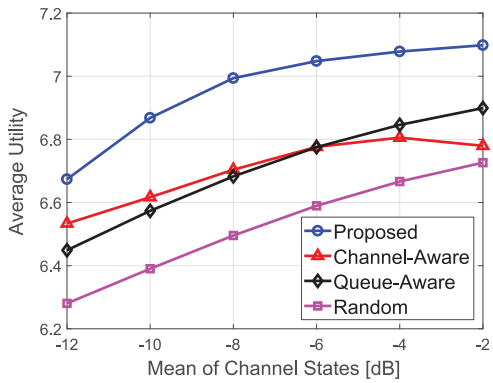

Fig. 11. Average utility per MT across the learning procedure versus means of channel states.

WSPs can be guaranteed the requested number of channels on behalf of their subscribed MTs from the CNC. However, Baseline 1 cares only the channel states and does not take into consideration the number of queued packets, hence it realizes worse performance in the per-MT average queue length, the per-MT average packet drops and even the per-MT average utility.

Overall, from Experiments 2 and 3, the proposed learning based scheme outperforms the three baseline schemes with respect to the per-MT average utility performance in all simulations. When making the channel auction and packet scheduling decisions following our proposed scheme, the WSPs and the MTs obviously strike a strategic tradeoff between the immediate payoff performance and the potential payoffs from future competitive interactions.

6 Conclusions

In this paper, we investigate the wireless resource scheduling problem under a SDN-enabled virtualized RAN scenario. More specifically, the WSPs bid for the access to the limited number of channels on behalf of their MTs for packet scheduling over scheduling slots with the objective of maximizing their respective expected long-term payoffs. The CNC applies a VCG pricing mechanism to regulate the auction of channels at each scheduling slot. The problem is modelled as a stochastic game. The channel auction and packet scheduling decisions of a WSP require complete information of both the global network state and the control policy from all the other WSPs in the network, which makes the problem solving extremely challenging in a competitive network. By approximating the interactions among the competing WSPs, the original stochastic game is transformed into an abstract stochastic game, in which each WSP is thus able to behave independently. Furthermore, we decompose the per-WSP MDP into multiple single-agent MDPs to tackle the high signalling overheads and the costly computational complexity faced by each WSP in the abstract stochastic game. Such a decomposition separates the channel auction from the packet scheduling. Following this logic, a WSP makes the channel auction decision at each scheduling slot based on the expected future payments and the valuations received from all its MTs, while each MT determines the number of packets for transmission once receiving the channel scheduling outcome from the CNC. The experiment results clearly show that significant performance gains can be achieved from our proposed work.

Acknowledgments

This research was supported in part by AKA grants 310786 and 289611, TEKES grants 2364/31/2014 and 2368/31/2014, US National Science Foundation grants 1717454, 1731424, 1704092, 1702850, 1646607, 1547201, 1434789, 1456921, 1443917, 1405121, 1457262 and 1461886, and the Program for Zhejiang Leading Team of Science and Technology Innovation under Grant 2013TD20.

References

[1] H. Taoka, "Views on 5G," Dusseldorf, Germany, Oct. 2011, http://www.wwrf.ch/files/wwrf/content/files/publications/libraries/WWRF_Library_27.pdf

[2] N. Bhushan, et al., "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[3] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, Jan.–Mar. 2015.

[4] R. Mahindra, M. A. Khojastepour, H. Zhang, and S. Rangarajan, "Radio access network sharing in cellular networks," in *Proc. IEEE Int. Conf. Netw. Protocols*, Oct. 2013, pp. 1–10.

[5] J. He and W. Song, "AppRAN: Application-oriented radio access network sharing in mobile networks," in *Proc. IEEE Int. Conf. Commu.*, Jun. 2015, pp. 3788–3794.

[6] M. M. Rahman, C. Despins, and S. Affes, "Design optimization of wireless access virtualization based on cost & QoS trade-off utility maximization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6146–6162, Sep. 2016.

[7] Active ran sharing could save operators $60 billion. [Online]. Available: http://www.cellular-news.com/story/36831.php. Accessed on: Sep. 13, 2016.

[8] X. Costa-Péerez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[9] L. Zhao, M. Li, Y. Zaki, A. Timm-Giel, and C. Görg, "LTE virtualization: From theoretical gain to practical solution," in *Proc. Int. Teletraffic Congr.*, Sep. 2011, pp. 71–78.

[10] Z. Liu, Y. Li, L. Su, D. Jin, and L. Zeng, "M2cloud: Software defined multi-site data center network control framework for multi-tenant," in *Proc. ACM SIGCOMM*, Aug. 2013, pp. 517–518.

[11] B. Liu and H. Tian, "A bankruptcy game-based resource allocation approach among virtual mobile operators," *IEEE Commun. Lett.*, vol. 17, no. 7, pp. 1420–1423, Jul. 2013.

[12] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[13] T. Chen, H. Zhang, X. Chen, and O. Tirkkonen, "SoftMobile: Control evolution for future heterogeneous mobile networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 70–78, Dec. 2014.

[14] T. Chen, M. Matinmikko, X. Chen, X. Zhou, and P. Ahokangas, "Software defined mobile networks: Concept, survey and research directions," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 126–133, Nov. 2015.

[15] D. Drutskoy, E. Keller, and J. Rexford, "Scalable network virtualization in software-defined networks," *IEEE Internet Comput.*, vol. 17, no. 2, pp. 20–27, Mar. 2013.

[16] L. E. Li, Z. M. Mao, and J. Rexford, "Toward software-defined cellular networks," in *Proc. Eur. Workshop Softw. Defined Netw.*, Oct. 2012, pp. 7–12.

[17] S. Bhaumik, et al., "CloudIQ: A framework for processing base stations in a data center," in *Proc. ACM Annu. Int. Conf. Mobile Comput. Netw.*, Aug. 2012, pp. 125–136.

[18] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *Proc. ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw.*, Aug. 2013, pp. 25–30.

[19] Z. Ji and K. J. R. Liu, "Dynamic spectrum sharing: A game theoretical overview," *IEEE Commun. Mag.*, vol. 45, no. 5, pp. 88–94, May 2007.

[20] X. Chen, Z. Han, H. Zhang, M. Bennis, and T. Chen, "Foresighted resource scheduling in software-defined radio access networks," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2015, pp. 128–132.

[21] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1904–1919, May 2009.

[22] Y. Cui and V. K. N. Lau, "Distributive stochastic learning for delay-optimal OFDMA power and subband allocation," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4848–4858, Sep. 2010.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[24] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[25] E. V. Belmega and S. Lasaulce, "Energy-efficient precoding for multiple-antenna terminals," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 329–340, Jan. 2011.

[26] D. P. Bertsekas and R. Gallager. *Data Networks*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 1992.

[27] J. Jia, Q. Zhang, Q. Zhang, and M. Liu, "Revenue generation for truthful spectrum auction in dynamic spectrum access," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, May 2009, pp. 3–12.

[28] M. N. Tehrani and M. Uysal, "Spectrum trading for non-identical channel allocation in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5100–5109, Oct. 2013.

[29] D. Adelman and A. J. Mersereau, "Relaxations of weakly coupled stochastic dynamic programs," *Oper. Res.*, vol. 56, no. 3, pp. 712–727, Jan. 2008.

[30] A. M. Fink, "Equilibrium in a stochastic n-person game," *J. Sci. Hiroshima Univ. Ser. A-I*, vol. 28, pp. 89–93, 1964.

[31] C. Kroer and T. Sandholm, "Imperfect-recall abstractions with bounds in games," in *Proc. ACM Conf. Economics Comput.*, Jul. 2016, pp. 459–476.

[32] D. Abel, D. Hershkowitz, and M. Littman, "Near optimal behavior via approximate state abstraction," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 2915–2923.

[33] S. Ganzfried, T. Sandholm, and K. Waugh, "Strategy purification and thresholding: Effective non-equilibrium approaches for playing large games," in *Proc. 11th Int. Conf. Auton. Agents Multiagent Syst.*, Jun. 2012, pp. 871–878.

[34] J. N. Tsitsiklis and B. van Roy, "Feature-based methods for large scale dynamic programming," *Mach. Learn.*, vol. 22, no. 1–3, pp. 59–94, Jan. 1996.

[35] M. Loève, *Probability Theory I*. Berlin, Germany: Springer-Verlag, 1977.

[36] N. Mastronarde and M. van der Schaar, "Joint physical-layer and system-level power management for delay-sensitive wireless communications," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 694–709, Apr. 2013.
[37] N. Salodkar, A. Bhorkar, A. Karandikar, and V. S. Borkar, "An online learning algorithm for energy efficient delay constrained scheduling over a fading channel," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 732–742, May 2008.
[38] K. Soumyanath and V. S. Borkar, "An analog scheme for fixed-point computation-part II: Applications," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 46, no. 4, pp. 442–451, Apr. 1999.
[39] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-Learning," *Mach. Learn.*, vol. 16, no. 3, pp. 185–202, Sep. 1994.
[40] H. Wang and N. B. Mandayam, "A simple packet-transmission scheme for wireless data over fading channels," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1055–1059, Jul. 2004.

Xianfu Chen received the PhD degree in signal and information processing from the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, in March 2012. He is currently a senior scientist with the VTT Technical Research Centre of Finland Ltd, Oulu, Finland. His research interests cover various aspects of wireless communications and networking, with emphasis on network virtualization, software-defined radio access networks, green communications, centralized and decentralized resource allocation, and the application of machine learning to cognitive radio networks. He is a member of the IEEE.

Zhu Han received the BS degree in electronic engineering from Tsinghua University, in 1997, and the MS and PhD degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a research associate with the University of Maryland. From 2006 to 2008, he was an assistant professor with Boise State University, Idaho. Currently, he is a professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department, University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (Best Paper Award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Currently, he is an IEEE Communications Society Distinguished lecturer. He is a fellow of the IEEE.

Honggang Zhang is a full professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He is an Honorary visiting professor with the University of York, York, United Kingdom. He was the International chair professor of Excellence for Université Européenne de Bretagne (UEB) and Supélec, France. He served as the chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is currently active in the research on green communications and was the leading guest editor of the *IEEE Communications Magazine* special issues on Green Communications. He was the co-author and an editor of two books with the titles of *Cognitive Communications Distributed Artificial Intelligence (DAI), Regulatory Policy and Economics, Implementation* (John Wiley & Sons) and *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press), respectively. He is a senior member of the IEEE.

Guoliang Xue received the PhD degree in computer science from the University of Minnesota, in 1991. He is a professor of computer science and engineering, Arizona State University. His research interests span the areas of Quality of Service provisioning, network security and privacy, crowdsourcing and network economics, RFID systems and Internet of Things, smart city, and smart grids. He has published more than 280 papers in these areas, many of which in top conferences such as INFOCOM, MOBICOM, NDSS and top journals such as the *IEEE/ACM Transactions on Networking*, the *IEEE Journal on Selected Areas in Communications*, and the *IEEE Transactions on Mobile Computing*. He was a keynote speaker at IEEE LCN'2011 and ICNC'2014. He was a TPC co-chair of IEEE INFOCOM'2010 and a general co-chair of IEEE CNS'2014. He has served on the TPC of many conferences, including ACM CCS, ACM MOBIHOC, IEEE ICNP, and IEEE INFOCOM. He served on the editorial board of the *IEEE/ACM Transactions on Networking*. He serves as the area editor of the *IEEE Transactions on Wireless Communications*, overseeing 13 editors in the wireless networking area. He is a fellow of the IEEE, and the VP-Conferences of the IEEE Communications Society.

Yong Xiao received the BS degree in electrical engineering from the China University of Geosciences, Wuhan, China, in 2002, the MSc degree in telecommunication from the Hong Kong University of Science and Technology, in 2006, and the PhD degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012. Currently, he is a research assistant professor with the Department of Electrical and Computer Engineering, University of Arizona. He is also the center manager of the NSF BWAC Center, University of Arizona. His research interests include machine learning, game theory and their applications in wireless networks. He is a senior member of the IEEE.

Mehdi Bennis received the MSc degree in electrical engineering jointly from the EPFL, Switzerland and the Eurecom Institute, France, in 2002. He received the PhD degree in December 2009, in spectrum sharing for future mobile cellular systems. From 2002 to 2004, he worked as a research engineer with IMRA-EUROPE investigating adaptive equalization algorithms for mobile digital TV. In 2004, he joined the Centre for Wireless Communications (CWC), University of Oulu, Finland, as a research scientist. In 2008, he was a visiting researcher in the Alcatel-Lucent chair on flexible radio, SUPELEC. Currently he is an adjunct professor with the University of Oulu and Academy of Finland research fellow. His main research interests include radio resource management, heterogeneous networks, game theory, and machine learning in 5G networks and beyond. He has co-authored one book and published more than 100 research papers in international conferences, journals, and book chapters. He was the recipient of the prestigious 2015 Fred W. Ellersick Prize from the IEEE Communications Society and the 2016 Best Tutorial Prize from the IEEE Communications Society. He serves as an editor for the *IEEE Transactions on Wireless Communication*. He is a senior member of the IEEE.