Stochastic Game for Wireless Network Virtualization

Fangwen Fu, *Associate Member, IEEE*, and Ulas C. Kozat, *Senior Member, IEEE*

***Abstract*—We propose a new framework for wireless network virtualization. In this framework, service providers (SPs) and the network operator (NO) are decoupled from each other: The NO is solely responsible for spectrum management, and SPs are responsible for quality-of-service (QoS) management for their own users. SPs compete for the shared wireless resources to satisfy their distinct service objectives and constraints. We model the dynamic interactions among SPs and the NO as a stochastic game. SPs bid for the resources via dynamically announcing their value functions. The game is regulated by the NO through: 1) sum-utility optimization under rate region constraints; 2) enforcement of Vickrey–Clarke–Groves (VCG) mechanism for pricing the instantaneous rate consumption; and 3) declaration of conjectured prices for future resource consumption. We prove that there exists one Nash equilibrium in the conjectural prices that is efficient, i.e., the sum-utility is maximized. Thus, the NO has the incentive to compute the equilibrium point and feedback to SPs. Given the conjectural prices and the VCG mechanism, we also show that SPs must reveal their truthful value functions at each step to maximize their long-term utilities. As another major contribution, we develop an online learning algorithm that allows the SPs to update the value functions and the NO to update the conjectural prices iteratively. Thus, the proposed framework can deal with unknown dynamics in traffic characteristics and channel conditions. We present simulation results to show the convergence to the Nash equilibrium prices under various dynamic traffic and channel conditions.**

***Index Terms*—Conjectural price, game theory, sequential auction, spectrum management, wireless network virtualization.**

I. Introduction

IN TODAY'S wireless networks (including 4G), service providers (SPs) have to pick one of few service classes network operators and have no control over the actual scheduling policy. Consider a case where a video streaming service has a pair of active customers in a given cell site contending for the channel resources with a storage service provider that also has a pair of customers in the same cell. Storage service provider might have some critical uploads from one user and noncritical updates from the other user. Video streaming service might have already buffered enough packets to one user for playback, while its other user might be depleting his playback buffer fast. It is a formidable task for a network operator to determine the true utilities of each packet, assess the available spectrum resources, and make all the scheduling decisions. Wireless network virtualization can facilitate a flexible and lower complexity solution for service customization, where service providers determine their own utilities according to their own objectives/constraints, and the network operator provides efficient access to the spectrum with respect to these true utilities. Our work provides such a virtualization framework that is incentive-compatible and spectrally efficient.

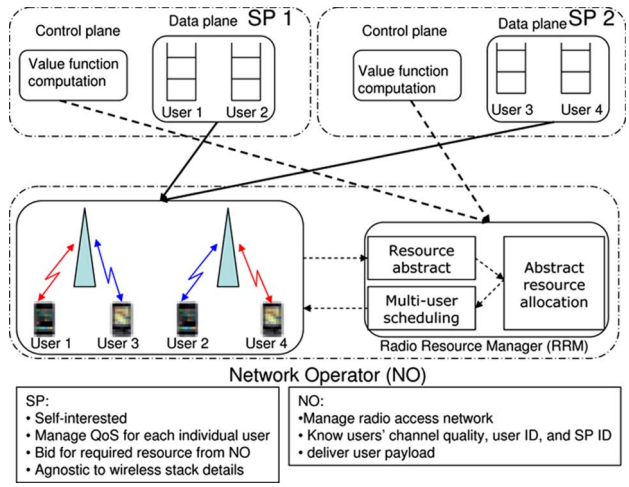

Fig. 1. Interface between NO and SPs.

In our framework, the wireless spectrum access is controlled by a single entity called network operator (NO) that focuses only on the efficient dynamic resource allocation by abstracting the underlying channel conditions via a time-varying feasible rate region. SPs on the other hand only focus on their own service objectives and constraints. In this virtualized architecture (see Fig. 1), SPs sequentially bid for the network resources on behalf of the end-users. Due to the stochastic nature of the channel and traffic dynamics, the problem becomes a stochastic game that has strong coupling across SPs and dependency on future actions. We combine VCG mechanism [13] and conjectural pricing as two key mechanisms to transform the game effectively into a sequence of single-shot games. To play the stochastic rate allocation game, SPs only need to choose the conjectural prices and to announce the value function (i.e., the preference on the rate) as required by the VCG mechanism. NO determines the rate allocation based on these bids and determines prices. As our major contributions, we prove that: 1) there exists one Nash equilibrium in the conjectural prices; 2) given the conjectural prices, the SPs have to truthfully reveal their own value function; and 3) this Nash equilibrium results in efficient rate allocation over the virtualization framework. Thus, NO has the incentive to compute the Nash equilibrium point in conjectural prices and feedback to SPs, while SPs have the incentives to follow NO's feedback.

Manuscript received May 08, 2010; revised December 16, 2010 and August 18, 2011; accepted March 02, 2012; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor V. Misra. Date of publication April 03, 2012; date of current version February 12, 2013. Parts of this work appeared in the IEEE International Conference on Computer Communications (INFOCOM), San Diego, CA, March 15–19, 2010.

F. Fu was with DOCOMO USA Labs, Palo Alto, CA 94304 USA. He is now with Intel, Folsom, CA 95630 USA (e-mail: fangwen.fu@intel.edu).

U. C. Kozat is with DOCOMO Innovations, Palo Alto, CA 94304 USA (e-mail: kozat@docomoinnovations.com).

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

Digital Object Identifier 10.1109/TNET.2012.2190419

TABLE I
BASIC NOTATION

i	SP index	M	number of SPs
k	user index	K	number of users
j	subchannel index	N	number of subchannels
t	time slot	$\mathcal{K}_i$	set of users subscribing service i
h_{kj}^t, h_{kj}	channel condition of user k at subchannel j	$\mathbf{h}_k^t, \mathbf{h}_k$	channel condition profile of user k
$\mathbf{H}_i^t, \mathbf{H}_i$	channel condition profile of users with SP i	$\mathbf{H}^t, \mathbf{H}$	channel condition profile of all users
f_{kj}	pdf of channel condition of user k at subchannel j	g_k^t, g_k	traffic state of user k at time t
$\mathbf{g}_i^t, \mathbf{g}_i$	traffic state of SP i	$\mathbf{g}^t, \mathbf{g}$	traffic state of all the users
s_k^t, s_k	state of user k	$\mathbf{s}_i^t, \mathbf{s}_i$	state of SP i
$\mathbf{s}^t, \mathbf{s}$	state of all the users	r_k^t, r_k	allocated rate to user k
ρ_k^t	allocated power to user k	w_k^t	allocated time to user k
u_k^t, u_k	immediate utility of user k	$\bar{u}_k$	average utility of user k
θ_i	value function of SP i	$\boldsymbol{\theta}$	value function profile of all SPs
λ_k	conjectural price of user k	$\boldsymbol{\lambda}_i$	conjectural price of SP i
τ_i	payment of SP i	V_i	payoff of SP i

Besides establishing an efficient separation among the NO and SPs, we also show how NOs can compute the conjectural prices and SPs compute the value functions efficiently in an iterative fashion without requiring the knowledge of time-varying channel and traffic dynamics. To this end, we use reinforcement learning methods [23]. Specifically, at each frame, the SPs update the value functions at the observed traffic states based on the advertised conjectural prices. Then, the SPs submit the updated value functions back to the NO. The NO computes the stochastic subgradient based on the submitted value functions, updates the conjectural price, and advertises it to all SPs. We prove that this iterative online learning algorithm converges to the optimal conjectural prices and the value functions. Via extensive simulations, we also observe that the algorithms converge fast and adapt well to various nonstationary channel and traffic dynamics.

The rest of the paper is organized as follows. Section II shows how the wireless network is virtualized and defines the interface to allocate the rate among the self-interested services. Section III formulates the dynamic rate allocation as a stochastic game. Section IV introduces the conjectural prices to represent the potential congestion level experienced by the end-users and remove the coupling between the SPs in the future game. Section V proves that one Nash equilibrium of conjectural prices exists that actually results in efficient rate allocation for the proposed virtualization framework. Section VI presents an online learning algorithm and proves its convergence. Section VII shows the simulation results. Section VIII summarizes the related work in the literature. Section IX draws the final conclusions. To make the paper more readable, we illustrate the major notations in Table I.

II. SYSTEM OVERVIEW

A. Wireless Network Virtualization

We consider a broadband wireless network that can support a diverse set of services over the same physical network. We divide the roles in the network into two: NO and SP. In the proposed system, there is only one NO, but in general there can be multiple SPs. SPs typically offer different services or have different business objectives reflecting distinct performance targets and constraints. Users can subscribe to one or more SPs separately.

NO dynamically manages the spectrum by collecting the channel state information (CSI) for each user, identifying rate regions and performing the wireless scheduling and wireless physical layer processing. Traffic flows have explicit identification that indicates the end-user as well as the SP. NO performs the wireless scheduling based on the most recent CSI information provided by the mobile users and the most recent value functions provided by each SP. NO does not have any explicit knowledge or control over the buffering performed above the radio link layer. NO is agnostic to the specifics (e.g., delay, jitter, long-term throughput, etc.) of quality-of-service (QoS) objectives of individual flow or SP.

On the other hand, SPs do not have a direct view of CSI information, how payloads are mapped onto the logical channels, and how physical transmission is realized. SPs are allocated isolated memory space and computation cycles so that they can implement their own buffering strategy (e.g., which user payloads to drop or which payload(s) to pass to the radio link layer) and carry out their own computation tasks (e.g., utility and value function computation) without interfering with each other. Thus, each SP has only the view about the traffic states of its own users. SPs dynamically bid on behalf of their users for the network resources to be allocated in the next scheduling interval.

In this virtualization framework, we let SPs to be self-interested, and hence the traffic information exchange may be strategic as it will be discussed in Section II-C. The NO is assumed to be an impartial entity applying a standard set of procedures for resource allocation and pricing. Since radio link layer and physical layer are controlled and implemented by the NO, the CSI exchange between the end-users and the NO is considered as nonstrategic, i.e., users cannot announce false CSI values to gain advantage over the others.

B. Channel Model: Network Operator's View

We consider a time-slotted system, in which the NO makes scheduling decisions every W seconds (referred to as *frame* or *scheduling interval* interchangeably hereon). The network operator has N orthogonal subchannels indexed by $j \in \{1, \ldots, N\}$.

In this network, there are in total K end-users indexed by $k \in \{1, \ldots, K\}$. During the transmission, we assume that end-users experience a block-fading channel. At frame t, the user k experiences channel gain h_{kj}^t at subchannel j, and the channel

gain is constant within the frame but changes across frames. The channel gain profile of user k over all subchannels is denoted by $\mathbf{h}_k^t = [h_{k1}^t, \ldots, h_{kN}^t]^T$, where $\mathbf{x}^T$ represents the transpose of a vector or matrix $\mathbf{x}$. We assume that the channel gain h_{kj}^t is i.i.d. across time for user k at subchannel j with the probability density function (pdf) of $f_{kj}(h)$.

Given the wireless network infrastructure, we assume that the channel gain profile of user k is truthfully known to both user k and the NO. For simplicity, we assume that the TDMA-like channel access is deployed and that any fraction of scheduling interval can be assigned to individual receivers. Accordingly, within frame t, the NO performs user scheduling and spectrum allocation by specifying the fraction of time w_{kj}^t for user k at subchannel j. Here, w_{kj}^t continuously takes values in $[0, W]$, which approximates the discrete time allocation in the real system. As another simplifying assumption, we normalize the total power as 1 and let power allocation ρ_{kj} be constant for user k at subchannel j during the whole transmission period. However, the proposed framework can be easily extended to the scenarios where the transmission power can be dynamically adapted. Since the resource allocation is performed by the NO, we are able to virtualize the wireless network and abstract the available wireless network resource as a rate region denoted by $\mathcal{R}^t$. The rate region is computed as the set of rates that can be achieved by any spectrum allocation policy under the current channel gain profile $\mathbf{H}^t = [\mathbf{h}_1^t, \ldots, \mathbf{h}_M^t]$. Specifically, we have

$$\mathcal{R}^t = \left\{ \mathbf{r}^t \in \mathbb{R}_+^K | \exists w_{kj}^t \geq 0,\ \forall k, j \right. \\ \left. r_k^t = \sum_{j=1}^{N} \frac{B \log(1+\rho_{kj} h_{kj}^t) w_{kj}^t}{2},\ \sum_{k=1}^{K} w_{kj}^t \leq W,\ \forall j \right\} \quad (1)$$

where r_k^t is the total transmission rate (e.g., information-theoretic rate[1]) for user k at frame t, and B the bandwidth of each subchannel. From (1), we note that the rate region $\mathcal{R}^t$ is determined by the channel condition profile $\mathbf{H}^t$, which is known by the NO. Hence, the wireless network at each frame can be represented by $\mathcal{R}^t = \mathcal{R}(\mathbf{H}^t)$. It is easy to verify that $\mathcal{R}^t$ is a convex region and changes across frames. Given the rate region $\mathcal{R}(\mathbf{H}^t)$, the resource competition between SPs becomes the rate allocation with the constraint of rate profile being in the feasible region.[2] In the following, we represent the wireless network at each frame t synonymously with state $\mathbf{s}^t$. This rate allocation separates the complicated spectrum sharing (i.e., user scheduling and spectrum allocation, etc.) from the services in the upper layer. In the next section, we discuss how the virtualized network resources (i.e., feasible rate region) should be allocated to the self-interested SPs.

C. Interface Between NO and SPs

Depending on the services that they subscribe, the end-users are divided into M groups, each of which corresponds to one type of service provided by the SP $i \in \{1, \ldots, M\}$. The set of end-users subscribed to service i is denoted by $\mathcal{K}_i$. Without any loss of generality, let us focus on the case where each end-user is subscribed to only one service in the network. Hence, $K = \sum_{i=1}^{M} |\mathcal{K}_i|$, where $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$. Also assume that each end-user at frame $t = 1, \ldots$ is able to be characterized by a state g_k^t representing the traffic state determined by the application user k runs. For example, the traffic state g_k^t can be the amount of packets that are available for transmission at the current frame or the amount of packets with different delay deadlines and distortion impacts [18]. Given the rate r_k^t, user k receives the immediate utility $u_k(g_k^t, r_k^t)$ at the traffic state g_k^t. We assume that the immediate utility $u_k(g_k^t, r_k^t)$ is a concave, increasing and differential function of the allocated rate r_k^t. The long-term average utility user k receives is computed as

$$\bar{u}_k = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} u_k^t. \quad (2)$$

For example, if the immediate utility of user k is the allocated rate r_k^t, the average utility is the average rate that user k receives. If the immediate utility is defined as $u_k(g_k^t, r_k^t) = -g_k^t$, where g_k^t is defined as the queueing length at frame t, the average utility becomes the negative average queue length, which is proportional to the average delay experienced by user k (then, maximizing the average utility is equivalent to minimizing the average delay). If the immediate utility is defined as the video distortion reduction of the transmitted video packets, the average utility is the average video quality user k obtains [18].

Given the transmission rate r_k^t, the transition of the traffic state g_k^t for each user k is denoted by $g_k^{t+1} = G_k(g_k^t, r_k^t) + a_k$, where a_k^t is the arriving data at the end of frame t and hence is available for transmission at next frame. $G_k(.)$ is a scheduling function that determines which packets will be transmitted given the rate allocation. For example, if g_k^t is the length of one queue in user k, the traffic state transition becomes $g_k^{t+1} = \max\{g_k^t - r_k^t, 0\} + a_k^t$ and $G_k(g_k^t, r_k^t) = \max\{g_k^t - r_k^t, 0\}$. If the traffic state is composed of packets with different delay deadlines and distortion impacts as in [18], the scheduling function is then determined based on the distortion impacts and delay deadlines. For simplicity, we assume that a_k^t is an i.i.d. random variable.

The role of SP i is to dynamically ask for the network resources (i.e., indirectly competing for the network resource with other SPs) for each of its subscribed users. The satisfaction function of SP i is denoted by $F_i(\bar{\mathbf{u}}_i)$, where $\bar{\mathbf{u}}_i = \{\bar{u}_k\}_{k \in \mathcal{K}_i}$. The satisfaction function $F_i(\bar{\mathbf{u}}_i)$ can also be interpreted as the willingness-to-pay (WTP) function of SP i, which is determined by the service level provided to the end-users in group i. We consider the case where the satisfaction functions of SPs are linear. Specifically, the utility function $F_i(\bar{\mathbf{u}}_i)$ for SP i has the following form:

$$F_i(\bar{\mathbf{u}}_i) = \sum_{k \in \mathcal{K}_i} \alpha_k \bar{u}_k \quad (3)$$

where $\alpha_k \in \mathbf{R}_+$ is the weight of the user k. Then, at frame t, SP i has the immediate utility $v_i^t = \sum_{k \in \mathcal{K}_i} \alpha_k u_k^t$ and $F_i = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} v_i^t$.

Due to the decentralized nature of the wireless network and self-interested service providers, we introduce a simple pricing

[1]The transmission rate can be closely approximated by considering different modulation and coding schemes.

[2]In essence, as long as a rate region can be specified given the channel state information of each receiver, the virtualization framework still applies.

mechanism named the Vickrey–Clarke–Groves (VCG) mechanism [13] into our framework. In this pricing mechanism, the SPs bid for the limited resources (i.e., the subchannels) on behalf of the end-users associated with them at each frame. Since only the NO knows the channel state $\mathbf{H}^t$, instead of directly bidding for the subchannels, SP i bids for the rate allocation.

At each frame t, SP i has the value over the potential allocated rate $\mathbf{r}_i^t$. This true value is denoted by $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$, where $\mathbf{g}_i^t = [\{g_k^t\}_{k\in\mathcal{K}_i}]$. Note that the value function $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$ may differ from the immediate utility function v_i^t, which will be illustrated in Section III.

Since the SPs are self-interested, they have incentives to announce a value function $\hat{\theta}_i(\mathbf{r}_i^t)$ different from $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$. In the VCG mechanism, receiving the announced value function $\hat{\theta}_i(\mathbf{r}_i^t)$, the NO performs the rate allocation within the feasible rate region $\mathcal{R}(\mathbf{H}^t)$ as follows:

$$\mathbf{r}^{t,*} = \arg \max_{\mathbf{r}\in\mathcal{R}(\mathbf{H}^t)} \sum_{i=1}^{M} \hat{\theta}_i(\mathbf{r}_i). \tag{4}$$

Note that $\mathbf{r}$ without subscript is the rate allocation for all the end-users, which is applied to other notation as well. Given the optimal rate allocation $\mathbf{r}^{t,*}$, the NO further computes the payment for SP i as follows:

$$\tau_i^t = \sum_{i'=1, i'\neq i}^{M} \hat{\theta}_{i'}(\mathbf{r}_{i'}^{t,*}) - \sum_{i'=1, i'\neq i}^{M} \hat{\theta}_{i'}(\mathbf{r}_{i',-i}^{t,*}) \tag{5}$$

where $r_{i',-i}^{t,*}$ is the optimal rate corresponding to the rate allocation rule in (4) when users $k \in \mathcal{K}_i$ are not included in the rate allocation. Notice that $\tau_i^t < 0$, which signifies the fact that SP i pays the amount of $|\tau_k^t|$ of money to the NO.

Properties of VCG mechanism [13] for one frame resource allocation are as follows:

- *Individual rationality:* The payoff of each SP, $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^{t,*}) + \tau_i^t$ at any frame t is not less than 0. In other words, participating the rate allocation game induced by the VCG mechanism at each frame is better than not participating it and having a zero payoff.
- *Incentive compatibility:* No matter what value function (truthful or not) other SPs announce to the NO, the truthful value function $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$ of SP i provides the best payoff. This implies that $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$ is the optimal value function SP i should announce to the NO, i.e., SPs have the incentive to anounce a value function $\hat{\theta}_i(\mathbf{r}_i^t)$ equal to their true value function $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$.
- *Efficiency:* When all SPs announce truthful value functions, the NO allocates the rate to maximize the sum of all the SPs' value function, which results in the efficient rate allocation.

Remark 1: The VCG mechanism is truth-revealing, incentive-compatible, individual-rational, and efficient only with respect to the value function $\theta_i(\mathbf{g}_i^t, \mathbf{r}_i^t)$ in one frame. However, in our context, the rate allocation is performed repeatedly with various channel conditions and end-users' traffic states. It is not clear if the above properties of the VCG are still true in the dynamic rate allocation, which will be discussed in the next section.

Remark 2: In the presented framework, we apply the VCG mechanism at each frame in order to capture the dynamics in the channel gains and traffic characteristics. When the channel gains change rapidly, it may require high computation cost and large signaling overhead to perform the VCG mechanism. However, to reduce the complexity, we notice that our proposed virtualization framework can be easily extended to the case in which the resource allocation as shown in (4) is performed every frame and the payment is computed in a larger period (multiple frames). In this way, the signaling about the value functions is executed only every multiple frames.

III. Stochastic Game Formulation

Although the VCG mechanism is efficient for the one frame resource allocation and has dominant strategy (i.e., announcing the truthful value function) [13] for each SP, it is still not clear how the VCG mechanism can be adapted to the stochastic environment in which the available resources are repeatedly allocated to the wireless users with time-varying states. In the following sections, we try to analyze the performance of the VCG mechanism in the stochastic environment by formulating the rate allocation problem as a stochastic game [15]. We assume that the NO performs the resource allocation based on the declared value functions and the underlying channel gains using the VCG mechanism. In other words, the VCG mechanism is fixed during each frame. The objective of SP i is to maximize the payoff (i.e., the achieved utility minus the payment), which is given by

$$\max_{\theta_i^t, t\geq 1} \{F_i(\bar{\mathbf{u}}_i) + \bar{\tau}_i\} \tag{6}$$

where $\bar{\tau}_i$ is the average payment to SP i that is computed as $\bar{\tau}_i = \lim_{T\rightarrow\infty} \frac{1}{T}\sum_{t=1}^{T} \tau_i^t$, and θ_i^t is the revealed value function. In order to maximize the payoff, SP i selects the value function $\theta_i^t \in \Theta_i$, which is viewed as the action to play the repeated rate allocation game. Here, Θ_i is the set of all possible value functions that SP i can take. For the repeated rate allocation among SPs, we can formulate it as a stochastic game as follows.

- There are M players, each of which corresponds to one SP and one network coordinator that is the NO.
- Each player has the state $\mathbf{g}_i^t$ at frame t.
- Each player has the action $\theta_i^t \in \Theta_i$, which represents the value function on the allocated rate at frame t.
- The state transition of each player has the form of

$$pr(\mathbf{g}_i^{t+1}|\mathbf{g}_i^t, \mathbf{r}_i^t) = \prod_{k\in\mathcal{K}_i} pr(g_k^{t+1}|g_k^t, r_k^t). \tag{7}$$

- Each player has the immediate payoff $v_i^t = \sum_{k\in\mathcal{K}_i} \alpha_k u_k^t + \tau_i^t$.
- The objective of each player is the same as in (6).
- The NO has the state $\mathbf{H}^t$.
- The state transition of the NO has the form of

$$pr(\mathbf{H}^{t+1}|\mathbf{H}^t) = pr(\mathbf{H}^{t+1}) = \prod_{k=1}^{K}\prod_{j=1}^{N} f_{jk}(h_{jk}^{t+1}). \tag{8}$$

- The resource allocation at each slot is performed by the NO via the VCG mechanism: $(\mathbf{r}^t, \boldsymbol{\tau}^t) = VCG(\boldsymbol{\theta}^t, \mathbf{H}^t)$.
- The state of the whole network is $\mathbf{s}^t = \{\mathbf{g}^t, \mathbf{H}^t\}$.

It is worth to note that the resource allocation performed by the NO is based on the declared value function $\boldsymbol{\theta}^t$ and the underlying channel conditions $\mathbf{H}^t$. The output of the stage game induced by the VCG mechanism (i.e., one frame resource allocation) is the allocated rate $\mathbf{r}_i^t$ and corresponding payment τ_i^t for each SP i. The state transition of SP i is only determined by the allocated rate $\mathbf{r}_i^t$. The channel state transition of the NO is independent of the resource allocation.

In this stochastic game, the policy $\boldsymbol{\pi}_i$ of SP i is a plan to play the game. Here, $\boldsymbol{\pi}_i = (\pi_i^1, \ldots, \pi_i^t, \ldots)$ is defined over the entire course of the game, where π_i^t is the decision rule at frame t mapping the history of the game up to time t to the action of selecting the value function: $\pi_i^t : \mathfrak{H}^t \mapsto \Theta_i$, where each element in $\mathfrak{H}^t$ is $\mathcal{H}^t = (\mathbf{s}^1, \boldsymbol{\theta}^1, \mathbf{r}^1, \tau^1, \ldots, \mathbf{s}^{t-1}, \boldsymbol{\theta}^{t-1}, \mathbf{r}^{t-1}, \tau^{t-1}, \mathbf{s}^t)$. $\boldsymbol{\pi}_i$ is called a *stationary* policy if $\pi_i^t = \pi_i$ for all t, and $\boldsymbol{\pi}_i$ is also called a *Markovian* policy if $\pi_i(\mathcal{H}^t) = \pi_i(\mathbf{s}^t)$ where $\mathcal{H}^t \in \mathfrak{H}^t$. In this paper, we focus on the stationary and Markovian policies for all the SPs although the nonstationary and non-Markovian policies may provide rich equilibria for the stochastic game. Instead of directly maximizing the long-term average payoff, i.e., $F_i(\bar{\mathbf{u}}_i) + \bar{\tau}_i = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} v_i^t$, we allow each SP to maximize the long-term discounted average payoff with discount factor $\beta \in [0, 1)$.[3] The long-term discounted average utility for SP i is expressed as

$$V_i^{\beta}(\mathbf{s}, \boldsymbol{\pi}) = (1 - \beta) \sum_{t=1}^{\infty} \beta^{t-1} v_i^t. \tag{9}$$

Note that the long-term discounted average payoff of SP i depends on the states and policies of all the SPs. It can be shown that the long-term undiscounted average payoff can be achieved when β approaches to 1 [17]. Hence, in the rest of the paper, we focus on the policies that maximize the discounted average payoff instead of the undiscounted average payoff.

The best response of SP i to the policy $\boldsymbol{\pi}_{-i}$ of other SPs is represented by

$$\pi_i^*(\boldsymbol{\pi}_{-i}) = \arg \max_{\pi_i \in \Pi_i} V_i^{\beta}(\mathbf{s}, \{\pi_i, \boldsymbol{\pi}_{-i}\}) \qquad \forall \mathbf{s}. \tag{10}$$

Based on the best response, we can define the Nash equilibrium in the stochastic game as follows.

Definition 1 (Nash Equilibrium): The Nash equilibrium of the stochastic game is a policy $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_M^*)$ such that for $\forall \mathbf{s}$ and $\forall i$, π_i^* is the best response against the other SP policies $\boldsymbol{\pi}_{-i}^*$.

It can be shown that, for the discounted stochastic game, there always exists a stationary and Markovian policy that is Nash Equilibrium [15].

However, it is notoriously hard to find the Nash equilibrium for the stochastic game. Actually, we note that in order to operate at Nash Equilibrium, each SP needs to know the global state $\mathbf{s}$, which is prohibited in our decentralized wireless network. In fact, during the resource allocation, each SP observes the partial history up to time t, $\mathcal{H}_i^t = \{\mathbf{g}_i^1, \boldsymbol{\theta}_i^1, \mathbf{r}_i^1, \tau_i^1, \ldots, \mathbf{g}_i^{t-1}, \boldsymbol{\theta}_i^{t-1}, \mathbf{r}_i^{t-1}, \tau_i^{t-1}, \mathbf{g}_i^t\}$ as shown in Fig. 2. In Section IV, we discuss how the SPs play this stochastic rate allocation game with the partially observed information.

[3]Although here we use the same discount factor for all SPs, in general each SP is allowed to choose a different discount factor.

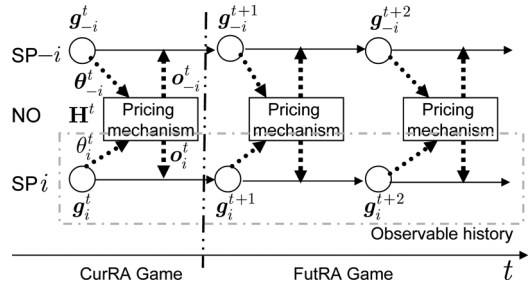

Fig. 2. Information structure of resource allocation game.

IV. Playing Stochastic Game via Conjectural Price

A. Information Structure

As shown in Fig. 2, in this stochastic resource allocation game, the interaction between SPs are through the VCG mechanism performed by the NO at each frame. At frame t, the output of the VCG mechanism (also called the allocation at frame t) is denoted by $\mathbf{o}^t = (o_1^t, \ldots, o_M^t)$ where $o_i^t = (\mathbf{r}_i^t, \tau_i^t)$.

Since the VCG mechanism is fixed during the whole course of the game, the allocation o_i^t is uniquely determined by the value function profile $\boldsymbol{\theta}^t$ and the channel profile $\mathbf{H}^t$ of all users. We explicitly express the allocation o_i^t as a function of the value function profile $\boldsymbol{\theta}^t$ and the channel profile $\mathbf{H}^t$, i.e., $o_i^t(\boldsymbol{\theta}^t, \mathbf{H}^t)$. In this stochastic game, SP i submits the value function θ_i^t to compete for the network resource, which affects the game in two folds.

- The announced value function θ_i^t affects SP i's long-term discounted average payoff through the allocation o_i^t. From Fig. 2, we know that the allocation o_i^t determines the immediate payoff $v_i^t(\mathbf{g}_i^t, \mathbf{r}_i^t)$ and the traffic state transition $pr(\mathbf{g}_i^{t+1} | \mathbf{g}_i^t, \mathbf{r}_i^t)$.
- The announced value function $\boldsymbol{\theta}_i^t$ also affects other SPs' long-term discounted average payoff through the allocation $\mathbf{o}_{-i}^t$ in a similar way.

Next, we will formally characterize these complex couplings and the idea of conjectural prices.

B. Conjectural Price

Since the one frame resource allocation game (i.e., stage game) is played repeatedly using the VCG mechanism with different states of the SPs at each frame, we can split the stochastic game into two phases as shown in Fig. 2: current resource allocation (CurRA) game (i.e., the stage game at the current frame) and future resource allocation (FutRA) game (which is also a stochastic game starting from different states of the SPs). As discussed in Section IV-A, the coupling between the CurRA game and FutRA game is that the output $\mathbf{o}^t$ of the CurRA game will affect the initial states of all SPs in the FutRA game. Assuming that in the FutRA game all SPs play the Nash Equilibrium policy $\boldsymbol{\pi}^*$, the corresponding discounted average utility is given by $V_i^{\beta}(\mathbf{s}, \boldsymbol{\pi}^*)$, $\forall i$. Then, given the Nash

equilibrium payoff $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi}^*),\ \forall i$, the best response of SP i for the CurRA game with state profile $\mathbf{s}$ can be expressed as

$$\theta_i(\mathbf{s},\boldsymbol{\theta}_{-i},\boldsymbol{\pi}^*)=\arg\max_{\theta_i\in\Theta_i} \underbrace{(1-\beta)\left(\sum_{k\in\mathcal{K}_i}\alpha_k u_k(g_k,r_k(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))+\tau_i(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H})\right)}_{\text{current reward } v_i^t} + \underbrace{\left\{\begin{array}{c}\beta\sum_{\mathbf{s}'}\prod_{k\in\mathcal{K}_i}\{pr(g_k'|g_k,r_k(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))\}\,pr(\mathbf{H}')\\ pr(\mathbf{g}_{-i}'|\mathbf{g}_{-i},\mathbf{r}_{-i}(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))V_i^{\beta}(\mathbf{s}',\boldsymbol{\pi}^*)\end{array}\right\}}_{\text{average future reward}}. \tag{11}$$

Note that $\mathbf{s}'=(\mathbf{g}_i',\mathbf{g}_{-i}',\mathbf{H}')$. It has been shown [15] that, corresponding to the Nash equilibrium payoff $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi}^*),\ \forall i$, there is one Nash equilibrium $\boldsymbol{\pi}^{\mathrm{CurRA}}(\mathbf{s})$ in the CurRA game. By the recursive nature of the stochastic game [15], the Nash equilibrium $\boldsymbol{\pi}^{\mathrm{CurRA}}(\mathbf{s})=\boldsymbol{\pi}^*(\mathbf{s})$. In other words, the Nash equilibrium policy $\boldsymbol{\pi}^*$ played in the FutRA game induces the Nash equilibrium $\boldsymbol{\pi}^{\mathrm{CurRA}}(\mathbf{s})=\boldsymbol{\pi}^*(\mathbf{s})$ played in the CurRA game.

Now consider the case where instead of playing the Nash equilibrium policy $\boldsymbol{\pi}^*$ in the FutRA game, the SPs play an arbitrary policy $\boldsymbol{\pi}$ that leads to the payoff $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi}),\ \forall i$. From (11), we know that the payoff $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi}),\ \forall i$ will induce a new CurRA game that is a one-stage game and has at least one (mixed) Nash equilibrium. The following lemma formally states the existence of the Nash equilibrium for the CurRA game and summarizes the discussion so far.

Lemma 3 (Existence of Nash Equilibrium): Any stationary policy $\boldsymbol{\pi}$ played by the SPs in the FutRA game can induce one Nash equilibrium policy $\boldsymbol{\pi}^{\mathrm{CurRA}}(\mathbf{s},\boldsymbol{\pi})$ played in the CurRA game with the state $\mathbf{s}$. It is clear that $\boldsymbol{\pi}^{\mathrm{CurRA}}(\mathbf{s},\boldsymbol{\pi}^*)=\boldsymbol{\pi}^*$. The payoff profile $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi})$ for each i induces the best response policy [as shown in (11)] played by SP i in the CurRA game. Hence, the policy of SP i to play the whole stochastic game can be interpreted as $(\pi_i^{\mathrm{CurRA}}(\mathbf{s},\boldsymbol{\pi}),\boldsymbol{\pi})$.

However, it is still difficult to find the Nash equilibrium $\boldsymbol{\pi}^*$ in the FutRA game. Even if the discounted average utility $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi}^*)$ at the Nash Equilibrium policy is known, SP i has to know the state transition $pr(\mathbf{g}_{-i}'|\mathbf{g}_{-i},\mathbf{r}_{-i}(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))$ of other SPs and the channel state distribution $pr(\mathbf{H})$ of the NO, which is impossible to be known in practice. Instead of directly finding the Nash equilibrium $\boldsymbol{\pi}^*$ in the FutRA game, we are interested in those policies that lead to decoupling in the payoff function, i.e., $V_i^{\beta}(\mathbf{s},\boldsymbol{\pi})=V_i^{\beta}(\mathbf{g}_i,\pi_i)$. The benefits of this decoupling will be clear later in this section.

The decoupling can be achieved by introducing a conjectural price $\boldsymbol{\lambda}_i=\{\lambda_k\}_{k\in\mathcal{K}_i}$ where $\lambda_k\in\mathbb{R}_+$. Via the conjectural price $\boldsymbol{\lambda}_i$, SP i no longer requires any information about other SPs and the NO, e.g., states, state transitions, etc. The conjectural price is defined as follows.

Definition 2 (Conjectural Price): The conjectural price $\boldsymbol{\lambda}_i$ is the belief of SP i on the per-unit cost (charged by the NO) on the allocated rate (by the NO) in the FutRA game.

The conjectural price $\boldsymbol{\lambda}_i$ represents the potential congestion level SP i expects in the future. Note that the conjectural price is not the true (average) price that SP i will be charged in the FutRA game. However, the conjectural price allows SPs to lump all the unknown dynamics of the problem into one price vector.

Given the conjectural price $\boldsymbol{\lambda}_i$, the FutRA game is decomposed into M independent Markov decision processes, each of which corresponds to the rate allocation for one SP and the discounted average utility (called "Conjectural State Value Function") of SP i starting from the traffic state $\mathbf{g}_i$ in the FutRA game is independently computed as

$$V_i^{\beta,cp}(\mathbf{g}_i,\boldsymbol{\lambda}_i)=\sum_{k\in\mathcal{K}_i}U_k^{\beta,cp}(g_k,\lambda_k) \tag{12}$$

where $U_k^{\beta,cp}(g_k,\lambda_k)$ is the solution to the following Bellman's equations:

$$U_k^{\beta,cp}(g_k,\lambda_k)=\max_{r_k\in\mathbb{R}_+}\Big\{(1-\beta)(\alpha_k u_k(g_k,r_k)-\lambda_k r_k) +\beta\sum_{g_k'}pr(g_k'|g_k,r_k)U_k^{\beta,cp}(g_k',\lambda_k)\Big\}\qquad\forall g_k. \tag{13}$$

SP i is now able to compute the conjectural state-value function as an approximation for the discounted average payoff of SP i achieved at the Nash equilibrium policy $\boldsymbol{\pi}^*$. The approximation enables us to simplify the best response given in (11) at the CurRA game as follows:

$$\theta_i(\mathbf{s},\boldsymbol{\theta}_{-i},\boldsymbol{\lambda}_i)=\arg\max_{\theta_i\in\Theta_i}\Big\{(1-\beta)\left(\sum_{k\in\mathcal{K}_i}\alpha_k u_k(g_k,r_k(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))+\tau_i(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H})\right) +\beta\sum_{k\in\mathcal{K}_i}\sum_{g_k'}pr(g_k'|g_k,r_k(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))U_k^{\beta,cp}(g_k',\lambda_k)\Big\}. \tag{14}$$

In this approximation, we ignore the states of other SPs and the channel states from next frame on.

Let us further explain the role of the conjectural price in the context of the stochastic game. SPs can unilaterally select their own conjectural prices $\boldsymbol{\lambda}_i,\ \forall i$ in the FutRA game and the output is $V_i^{\beta,cp}(\mathbf{g}_i',\boldsymbol{\lambda}_i),\ \forall i$. Hence, the policy of SP i to play this stochastic game becomes $(\pi_i^{\mathrm{CurRA}}(\mathbf{s},\boldsymbol{\lambda}_i),\boldsymbol{\lambda}_i)$ instead of $(\pi_i^{\mathrm{CurRA}}(\mathbf{s},\boldsymbol{\pi}),\boldsymbol{\pi})$, as shown in Fig. 3. Via the conjectural price, the payoff in the FutRA game is decomposed, significantly simplifying the selection of the value function θ_i in playing the CurRA game. Nonetheless, the answers for two critical questions are still pending.

- Given the conjectural prices, how can the SPs compute the value function θ_i?
- How can the SPs conjecture their own prices to maximize their long-term utilities?

In the next section, we first focus on the value function computation when the conjectural prices are given. We delay the discussion on the conjectural price selection to Section V.

C. Repeated CurRA Game With Fixed Conjectural Prices

In this section, we focus on the CurRA game when the conjectural prices of all the SPs are fixed. As discussed in Section II-C,

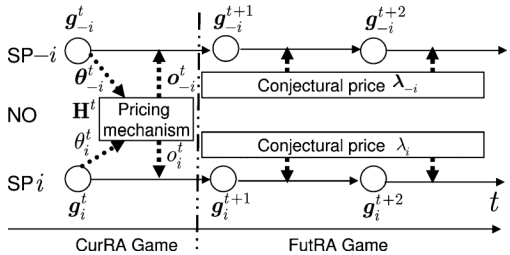

Fig. 3. Resource allocation game with conjectural price.

the resource allocation in the CurRA game is performed through the VCG mechanism. Rearranging(14), we obtain

$$\theta_i(\mathbf{s},\boldsymbol{\theta}_{-i},\boldsymbol{\lambda}_i)=(1-\beta)\cdot\arg\max_{\theta_i\in\Theta_i}\Bigg\{\sum_{k\in\mathcal{K}_i}[\alpha_k u_k(g_k,r_k(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))+\frac{\beta}{(1-\beta)}\sum_{g'_k}pr(g'_k|g_k,r_k(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H}))U_k^{\beta,cp}(g'_k,\lambda_k)]+\tau_i(\theta_i,\boldsymbol{\theta}_{-i},\mathbf{H})\Bigg\}. \quad (15)$$

Accordingly, the truthful value function of SP i in the CurRA game is defined as

$$\theta_i(\mathbf{g}_i,\mathbf{r}_i)=\sum_{k\in\mathcal{K}_i}\Bigg\{\alpha_k u_k(g_k,r_k)+\frac{\beta}{(1-\beta)}\sum_{g'_k}pr(g'_k|g_k,r_k)U_k^{\beta,cp}(g'_k,\lambda_k)\Bigg\}=\sum_{k\in\mathcal{K}_i}\theta_k(g_k,r_k). \quad (16)$$

In this value function, we note that SP i not only cares about its immediate utility, but also the future payoff through the state transition. The payoff of SP i in the VCG mechanism becomes $(1-\beta)(\theta_i(\mathbf{g}_i,\mathbf{r}_i)+\tau_i)$. From Section IV-B, we know that the payoff in the FutRA game affects the action selection in the CurRA game through the best response as shown in (11). Notice that the coupling in the payoff from the general policies played in the FutRA game prohibits the computation of the best response in the CurRA game. Conjectural prices remove this coupling. Given $\boldsymbol{\lambda}_i,\ \forall i$, the SPs have the fixed value function $\theta_i(\mathbf{g}_i,\mathbf{r}_i)$ in the CurRA game. Thus, the CurRA game becomes a one-shot game governed by the VCG mechanism. We know that in this one-shot game there exists one dominant strategy that is incentive-compatible and truth-revealing. However, the incentive-compatible and truth-revealing strategy is with respect to the conjectural prices, i.e., different conjectural price vectors would lead to different games with different strategies. We denote this dominant strategy by $\theta_i^*(\mathbf{g}_i,\boldsymbol{\lambda}_i)$. Going back to the stochastic rate allocation game, the selection of the conjectural price is analogous to the policy for playing the FutRA game. Once the conjectural prices are fixed, the curRA game is played independently of the FutRA game. Hence, the stochastic game is simplified into a repeated curRA game. In this repeated curRA game, we have the dominant strategy described as follows.

Proposition 4 (Dominant Strategy): In the stochastic game, if the SPs play the FutRA game by conjecturing the resource price $\lambda,\ \forall i$ i.e., selecting the policy $(\theta_i,\lambda_i),\ \forall i$ for playing the stochastic game, then for any conjectural price profile $\boldsymbol{\lambda}_i,\ \forall i$, $(\theta_i^*(\mathbf{g}_i,\boldsymbol{\lambda}_i),\boldsymbol{\lambda}_i),\ \forall i$ is a dominant strategy profile.

Proof: Given the conjectural prices $\boldsymbol{\lambda}_i,\ \forall i$, each CurRA game with any state $\mathbf{s}$ is a one-shot resource allocation game induced by the VCG mechanism, and $(\theta_i^*(\mathbf{g}_i,\boldsymbol{\lambda}_i),\boldsymbol{\lambda}_i)$ is the dominant strategy in this game as discussed in Section II-C. Hence, it is also the dominant strategy in the repeated CurRA game with the fixed conjectural prices. ■

Proposition 4 implies that there is an infinite number of dominant strategies in the repeated CurRA game since any conjectural price profile $\boldsymbol{\lambda}_i,\ \forall i$ induces one dominant equilibrium, similar to the Folk theorem in the repeated game [16]. The remaining problem is how to select an appropriate conjectural price profile to play the FutRA game.

V. Conjectural Price Selection

In this section, we discuss the selection of the conjectural prices to play the FutRA game such that the SPs maximize their own payoffs. Since within our virtualization framework, SPs only observe a partial history $\mathcal{H}_i^t=\{\mathbf{g}_i^1,\theta_i^1,\mathbf{r}_i^1,\tau_i^1,\ldots,\mathbf{g}_i^{t-1},\theta_i^{t-1},\mathbf{r}_i^{t-1},\tau_i^{t-1},\mathbf{g}_i^t\}$, it is often difficult to infer the congestion level (i.e., conjectural price) for the FutRA game from this partially observed history. However, the NO collects all the value functions (which represents the utility of the SPs) and then makes the rate allocation and payment computation. In other words, the NO has the global information about the whole network, and it is in a perfect position to advertise conjectural prices to SPs to guide their bidding decisions. For these advertised conjectural prices, we have to answer the following two questions.

- What conjectural prices should the NO advertise?
- Do SPs adopt these prices as their own conjectural prices?

To answer these two questions, we first look at the best performance (i.e., highest system utility) the NO can obtain using the conjectural prices in the cooperative and decentralized scenarios. Then, we analyze whether the conjectural prices corresponding to the best performance can be adopted by SPs.

A. Cooperative Solution Using Conjectural Prices

From the perspective of the NO, the efficient resource allocation is to cooperatively maximize the sum utility of all wireless users as given by

$$U^{\text{coop}}(\mathbf{s}^t)=\max_{\mathbf{r}^{t'}\in\mathcal{R}^{t'},\ \forall t'\geq t}(1-\beta)\sum_{t'=t}^{\infty}\beta^{t'-t}\sum_{k=1}^{K}\alpha_k u_k(g_k^{t'},r_k^{t'}).$$

Based on the conjectural price profile $\boldsymbol{\lambda}$, we relax the rate constraint $\mathbf{r}^t\in\mathcal{R}^t$ by introducing the cost of violating rate constraint at frame t, i.e., $\boldsymbol{\lambda}^T(\mathbf{r}^t-\hat{\mathbf{r}}^t(\boldsymbol{\lambda}))$, where $\hat{\mathbf{r}}^t(\boldsymbol{\lambda})$ is the optimal rate within the feasible rate region to the following optimization problem:

$$\hat{\mathbf{r}}^t(\boldsymbol{\lambda})=\arg\max_{\mathbf{r}\in\mathcal{R}^t}\boldsymbol{\lambda}^T\mathbf{r}. \quad (17)$$

The relaxation is a generalized Lagrangian relaxation for the convex constraint, e.g., $\mathbf{r}^t \in \mathcal{R}^t$ in this paper.

Then, we have

$$\begin{aligned}
&U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda}) = \max_{\mathbf{r}^{t'} \in \mathbb{R}_+^K, t' \geq t} (1-\beta) \cdot \\
&\sum_{t'=t}^{\infty} \beta^{t'-t} \left\{ \sum_{k=1}^{K} \alpha_k u_k(g_k^{t'}, r_k^{t'}) - \boldsymbol{\lambda}^T (\mathbf{r}^{t'} - \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda})) \right\} \\
&= \sum_{k=1}^{K} \max_{r_k^{t'} \in \mathbb{R}_+, t' \geq t} (1-\beta) \sum_{t'=t}^{\infty} \beta^{t'-t} \{ \alpha_k u_k(g_k^{t'}, r_k^{t'}) - \lambda_k r_k^{t'} \} \\
&\quad + (1-\beta) \boldsymbol{\lambda}^T \sum_{t'=t}^{\infty} \beta^{t'-t} \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}) \\
&= \sum_{k=1}^{K} U_k^{\text{coop}}(g_k^t, \lambda_k) + (1-\beta) \boldsymbol{\lambda}^T \sum_{t'=t}^{\infty} \beta^{t'-t} \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}). \qquad (18)
\end{aligned}$$

$\hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda})$ is determined based on the conjectural price $\boldsymbol{\lambda}$ and the rate region $\mathcal{R}^t$ (and hence, the channel condition $\mathbf{H}^t$) and is independent of the selection of the rate $\mathbf{r}^t$. Note also that $U_k^{\text{coop}}(g_k^t, \lambda_k) = U_k^{\beta,cp}(g_k^t, \lambda_k)$, and they can be computed by the corresponding SPs. Hence, $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda})$ is essentially composed of two terms that can be computed independently by the SPs (computing the first term) and the NO (computing the second term) using their own state transitions given $\boldsymbol{\lambda}$ and then combined together.

From [17], we know that $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda}) \geq U^{\text{coop}}(\mathbf{s}^t), \ \forall \mathbf{s}^t$. In other words, $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda})$ is the upper bound of $U^{\text{coop}}(\mathbf{s}^t)$ for any state $\mathbf{s}^t$. Then, the best conjectural price can be selected to minimize the gap between $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda})$ and $U^{\text{coop}}(\mathbf{s}^t)$, i.e.,

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda} \geq 0} \sum_{\mathbf{s}} \mu(\mathbf{s}) U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda}) \qquad (19)$$

where $\mu(\mathbf{s})$ is the stationary distribution of the network state. Using $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda}^*)$ as the approximated state-value function for the cooperative rate allocation, we are able to find an optimal feasible rate allocation $\mathbf{r}^{\boldsymbol{\lambda}^*}(\mathbf{s}^t) \in \mathcal{R}^t$ induced by $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda}^*)$, which is the solution to the following optimization problem:

$$\begin{aligned}
U^{\text{coop},\boldsymbol{\lambda}^*}(\mathbf{s}^t) = \max_{\mathbf{r}^t \in \mathcal{R}^t} &\left\{ (1-\beta) \sum_{k=1}^{K} \alpha_k u_k(g_k^t, r_k^t) - \lambda_k r_k^t \right. \\
&\left. + \beta \sum_{\mathbf{s}^{t+1}} pr(\mathbf{s}^{t+1} | \mathbf{s}^t, \mathbf{r}^t) U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda}^*) \right\} \\
= (1-\beta) \max_{\mathbf{r}^t \in \mathcal{R}^t} &\sum_{k=1}^{K} \left\{ \alpha_k u_k(g_k^t, r_k^t) - \lambda_k r_k^t \right. \\
&\left. + \frac{\beta}{1-\beta} \sum_{g_k^{t+1}} pr(g_k^{t+1} | g_k^t, r_k^t) U_k^{\text{coop}}(g_k^t, \lambda_k^*) \right\} \\
&+ (1-\beta)(\boldsymbol{\lambda}^*)^T \sum_{t'=t}^{\infty} \beta^{t'-t} \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}^*) \qquad (20)
\end{aligned}$$

where $R(\boldsymbol{\lambda}^*) = (1-\beta)(\boldsymbol{\lambda}^*)^T \sum_{t'=t}^{\infty} \beta^{t'-t} \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}^*)$ is computed by the NO and independent of the rate selection. From the monotonicity of the dynamic programming [19], we note that $U^{\text{coop}}(\mathbf{s}^t, \boldsymbol{\lambda}^*) \geq U^{\text{coop},\boldsymbol{\lambda}^*}(\mathbf{s}^t) \geq U^{\text{coop}}(\mathbf{s}^t), \ \forall \mathbf{s}^t$. Hence, the best conjectural price generates the feasible rate allocation policy as shown in (20), providing the optimal cooperative utility $U^{\text{coop},\boldsymbol{\lambda}^*}(\mathbf{s})$. We refer to the best conjectural price profile $\boldsymbol{\lambda}^*$ as the efficient price profile since it provides the efficient rate allocation in this distributed solution. Hence, the NO would like all the SPs to adopt this efficient price profile. With truthful revealing of the value functions by SPs, the NO is able to allocate the network resources efficiently. However, it is possible that the efficient price profile is not the preferable price for the SPs. We will check this next.

B. Nash Equilibrium of Efficient Price

From Section V-A, we know that $\boldsymbol{\lambda}^*$ provides the best cooperative utility in this decentralized resource allocation, i.e., it gives the efficient resource allocation. To enforce the SPs to adopt the conjectural prices advertised by the NO, we first compute the rate allocation based on the advertised prices, which is given as follows:

$$\mathbf{r}(\mathbf{s}, \boldsymbol{\lambda}^*) = \arg\max_{\mathbf{r} \geq \mathbf{0}} \sum_{k=1}^{K} \theta_k(g_k, r_k) - (\boldsymbol{\lambda}^*)^T \mathbf{r}. \qquad (21)$$

This rate can be computed by the NO since $\theta_k(g_k, r_k), \ \forall k$ are revealed by the SPs. Then, the following theorem shows that $\boldsymbol{\lambda}^*$ is the Nash equilibrium of the stochastic game played by the SPs as shown in Section IV-B.

Theorem 5 (Nash Equilibrium of Conjectural Price): $\boldsymbol{\lambda}^*$ results in the efficient rate allocation in the CurRA game and is the Nash equilibrium of the FutRA game in the stochastic game when SPs are charged with the additional payments $A \left\{ (1-\beta)(\boldsymbol{\lambda}^*)^T \sum_{t=1}^{\infty} \beta^{t-1} \mathbf{r}(\mathbf{s}^t, \boldsymbol{\lambda}^*) - R(\boldsymbol{\lambda}^*) \right\}^+$ with large enough $A \geq 0$.

Proof: From Proposition 4, given $\boldsymbol{\lambda}^*$, the SPs truthfully declare their value function, which is $\theta_i(\mathbf{g}_i, \mathbf{r}_i) = \sum_{k \in \mathcal{K}_i} \theta_k(g_k, r_k)$ as shown in (16). After receiving the value functions from the SPs, the NO performs the rate allocation as follows:

$$\begin{aligned}
\mathbf{r}^*(\mathbf{s}) &= \arg\max_{\mathbf{r} \in \mathcal{R}(\mathbf{H})} \sum_{k=1}^{K} \theta_k(g_k, r_k) - \lambda_k r_k \\
&= \arg\max_{\mathbf{r} \in \mathcal{R}(\mathbf{H})} \left\{ \sum_{k=1}^{K} \alpha_k u_k(g_k, r_k) - \lambda_k r_k \right. \\
&\qquad \left. + \frac{\beta}{(1-\beta)} \sum_{g_k'} pr(g_k' | g_k, r_k) U_k^{\beta,cp}(g_k', \lambda_k) \right\} \qquad (22)
\end{aligned}$$

where $\theta_k(g_k, r_k)$ is given as in (16). Since $U_k^{\text{coop}}(g_k^t, \lambda_k^*) = U_k^{\beta,cp}(g_k^t, \lambda_k^*)$, we can see the above optimization is equivalent to the optimization in (20). In other words, $\boldsymbol{\lambda}^*$ gives the efficient rate allocation in the CurRA game. $u_k(g_k, r_k)$ is a differential and concave function of r_k, thus it can be shown $\theta_k(g_k, r_k)$ is also a concave function for any conjectural price λ_k. Since $\boldsymbol{\lambda}^*$ is the efficient conjectural price, it implies that $(1-\beta)(\boldsymbol{\lambda}^*)^T \sum_{t=1}^{\infty} \beta^{t-1} \mathbf{r}(\mathbf{s}^t, \boldsymbol{\lambda}^*) - R(\boldsymbol{\lambda}^*) \leq 0$ when the SPs reveal their value functions computed with the conjectural prices $\boldsymbol{\lambda}^*$, which means that the rate allocation satisfies the long-term constraint as shown in [18]. When

the SPs announce the value functions with other conjectural prices $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^*$ that are not the solution to (19), we have $(1-\beta)(\boldsymbol{\lambda}^*)^T \sum_{t=1}^{\infty} \beta^{t-1} \mathbf{r}(\mathbf{s}^t, \boldsymbol{\lambda}^*) - R(\boldsymbol{\lambda}^*) \geq 0$. When A is large enough, the SPs do not have any incentive to select the conjectural prices other than $\boldsymbol{\lambda}^*$. ■

From Theorem 5, we know that when the SPs are enforced to take the conjectural prices to play the FutRA game, one Nash Equilibrium is the efficient price $\boldsymbol{\lambda}^*$. Furthermore, given the Nash equilibrium, the SPs play the CurRA game by truthfully revealing the value function that results in the efficient rate allocation. This truthful revelation actually leads to the dominant equilibrium in the CurRA game.

VI. Distributed Implementation and Online Learning

In Section V, we show that there exists an efficient conjectural price profile. However, to compute this price profile, the state transition of traffic states and the channel condition distribution must be known *a priori*. In this section, we discuss how to find such an efficient conjectural price profile when this knowledge is not available *a priori*, but learned over time.

A. Subgradient Method for Conjectural Price

From the duality theory [20], it is easy to verify that the optimization in (19) is a convex optimization with respect to the conjectural price profile $\boldsymbol{\lambda}$. This optimization can be solved by the NO using an iteration method as stated in the following proposition.

Proposition 6 (Subgradient Method): One of the subgradients for the optimization in (19) is given by

$$\Delta_{\boldsymbol{\lambda}} U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda}) = (1-\beta) \sum_{t=0}^{\infty} \beta^t \left\{ \mathbf{r}(\mathbf{s^t}, \boldsymbol{\lambda}) - \hat{\mathbf{r}}^t(\boldsymbol{\lambda}) \right\}. \quad (23)$$

The update of the conjectural price profile in the following converges to the efficient conjectural price:

$$\boldsymbol{\lambda}^{n+1} = \max\{\boldsymbol{\lambda}^n + \gamma^n \Delta_{\boldsymbol{\lambda}} U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda}^n), 0\} \quad (24)$$

where $\mathbf{r}(\mathbf{s^t}, \boldsymbol{\lambda})$ is computed as in (21) by replacing $\boldsymbol{\lambda}^*$ with $\boldsymbol{\lambda}$, and γ^n is a diminishing step size and satisfies $\sum_{n=1}^{\infty} \gamma^n = \infty$ and $\sum_{n=1}^{\infty} (\gamma^n)^2 \leq \infty$.

Proof: The computation of subgradient is directly from the definition of subgradient and similar to [18]. The convergence of the conjectural price update follows the subgradient method for nondifferential convex optimization [20]. ■

However, in order to compute the subgradient of $U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda})$, we have to wait for infinite time, which is not implementable in the real system. Instead, the subgradient is approximated by keeping $\Delta T \in \mathbb{N}$ items in (23) and truncating the remaining items. In other words, the subgradient is computed and the conjectural price profile is updated every ΔT frames.

B. Online Learning for Value Function

To play the resource allocation game, SP i has to submit the value function $\theta_k(g_k, r_k)$ for user $k \in \mathcal{K}_i$. However, we note that in order to compute $\theta_k(g_k, r_k)$, the SP must know the value function U_k^{cp} that is computed based on the Bellman's equation given in (13). To solve this Bellman's equation, the SPs has to know the state transition $pr(g_k'|g_k, r_k)$. Since we know that the data arrival process is independent of the transmission rate selection, the state transition can be rewritten as $pr(g_k'|g_k, r_k) = pr(a_k)\delta(g_k' - (G_k(g_k, r_k) + a_k))$, where $pr(a_k)$ is the distribution of the data arrival and G_k is a deterministic function. Similar to [21], we can define the traffic state just before the packet arrival as a post-decision traffic state $\tilde{g}_k = G_k(g_k, r_k)$, which is illustrated in Fig. 4. Then, similar to the Bellman's equation in (13), we have the Bellman's equation defined on the post traffic state, which is

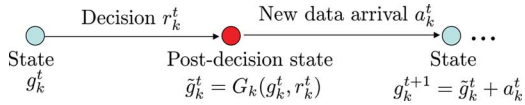

Fig. 4. Illustration of post-decision traffic state.

$$\begin{aligned}\tilde{U}_k^{\beta,cp}(\tilde{g}_k, \lambda_k) &= \sum_{a_k} pr(a_k) \max_{r_k \in \mathbb{R}_+} \{(1-\beta)(\alpha_k u_k(\tilde{g}_k + a_k, r_k) - \lambda_k r_k) \\ &\quad + \beta \tilde{U}_k^{\beta,cp}(G_k(\tilde{g}_k + a_k, r_k), \lambda_k)\} \qquad \forall \tilde{g}_k. \end{aligned} \quad (25)$$

The difference between the above post-state Bellman's equations and the original Bellman's equations in (13) is that the expectation on the data arrival is brought out of the maximization operator, which leads to simple online learning as described below. Note that $U_k^{cp}(g_k, \lambda_k) = \tilde{U}_k^{\beta,cp}(\tilde{g}_k + a_k, \lambda_k)$. Hence, SP i is able to compute the value function $\theta_k(g_k, r_k)$ once it knows $\tilde{U}_k^{\beta,cp}(\tilde{g}_k, \lambda_k)$.

Next, we propose an online learning algorithm to estimate $\tilde{U}_k^{\beta,cp}(\tilde{g}_k, \lambda_k)$. It is worth noting that, in (25), the maximization operator is performed with respect to the particular data arrival a_k, not over the the expectation. Hence, the expectation on the data arrival can be dropped by performing averaging in time. Specifically, we propose the following online learning algorithm to update $\tilde{U}_k^{\beta,cp}(\tilde{g}_k, \lambda_k)$ given the conjectural price λ_k. At each frame, the SP observes the post-decision traffic $\tilde{g}_k$ and the packet arrival a_k. Then, the update is performed as follows:

$$\begin{aligned}&\tilde{U}_k^{\beta,cp,t+1}(\tilde{g}_k^{t-1}, \lambda_k) \\ &\quad = (1-\kappa^t)\tilde{U}_k^{\beta,cp,t}(\tilde{g}_k^{t-1}, \lambda_k) \\ &\qquad + \kappa^t \max_{r_k \in \mathbb{R}_+} \{(1-\beta)(\alpha_k u_k(\tilde{g}_k^{t-1} + a_k^t, r_k) - \lambda_k r_k) \\ &\qquad\quad + \beta \tilde{U}_k^{\beta,cp,t}(G_k(\tilde{g}_k^{t-1} + a_k^t, r_k), \lambda_k)\} \\ &\tilde{U}_k^{\beta,cp,t+1}(\tilde{g}_k', \lambda_k) \\ &\quad = \tilde{U}_k^{\beta,cp,t}(\tilde{g}_k', \lambda_k) \qquad \forall \tilde{g}_k' \neq \tilde{g}_k^{t-1}. \end{aligned} \quad (26)$$

Above κ^t is a positive diminishing step size satisfying the conditions: $\sum_{t=1}^{\infty} \kappa^t = \infty$ and $\sum_{t=1}^{\infty} (\kappa^t)^2 \leq \infty$. The following proposition shows that the above update converges to the optimal post-decision state-value function with respect to the conjectural price λ_k.

Proposition 7 (Online Learning): $\lim_{t \to \infty} \tilde{U}_k^{\beta,cp,t}(\tilde{g}_k, \lambda_k) = \tilde{U}_k^{\beta,cp}(\tilde{g}_k, \lambda_k), \ \forall \tilde{g}_k$.

Proof: The convergence proof follows the theory of stochastic approximation [22] and is omitted here. ■

C. Primal Dual Update

In Sections VI-A and VI-B, we have discussed the conjectural price update and the state-value function evaluation, respectively. However, the conjectural price update assumes that the SPs submit the optimal value function with respect to the advertised price, and the state-value function evaluation assumes that the conjectural price is constant for the whole course of update. In the real system, the conjectural price update and state-value function evaluation should be performed simultaneously. To do this, we propose a primal dual update algorithm to update the conjectural price (corresponding to the dual update) and post-state-value function evaluation (corresponding to the primal update) on line. Specifically, the NO updates the conjectural price every $\Delta T \geq 1$ frames, and the SPs updates the post-state-value function for each associated user at every frame, which is formalized as follows:

NO:

$$\text{if } t = n\Delta T : \boldsymbol{\lambda}^{t+1} = \max\left\{\boldsymbol{\lambda}^t + \gamma^n(1-\beta) \sum_{t'=(n-1)\Delta T+1}^{n\Delta T} \beta^{t'}\left\{\mathbf{r}(\mathbf{s}^{\mathbf{t}'}, \boldsymbol{\lambda}) - \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}^t)\right\}, 0\right\}$$
$$\text{otherwise} : \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t. \tag{27}$$

SP i:

$$\begin{aligned}
&\tilde{U}_k^{\beta,cp,t+1}(\tilde{g}_k^{t-1}, \lambda_k^t) \\
&\quad = (1-\kappa^t)\tilde{U}_k^{\beta,cp,t}(\tilde{g}_k^{t-1}, \lambda_k^t) \\
&\qquad + \kappa^t \max_{r_k \in \mathbb{R}_+} \{(1-\beta)(\alpha_k u_k(\tilde{g}_k^{t-1} + a_k^t, r_k) - \lambda_k^t r_k) \\
&\qquad\qquad + \beta\tilde{U}_k^{\beta,cp,t}(G_k(\tilde{g}_k^{t-1} + a_k^t, r_k), \lambda_k^t)\} \\
&\tilde{U}_k^{\beta,cp,t+1}(\tilde{g}_k', \lambda_k) \\
&\quad = \tilde{U}_k^{\beta,cp,t}(\tilde{g}_k', \lambda_k^t),\ \tilde{g}_k' \neq \tilde{g}_k^{t-1} \qquad \forall k \in \mathcal{K}_i.
\end{aligned} \tag{28}$$

The following proposition ensures that the above update will converge to the optimal solution.

Proposition 8: $\boldsymbol{\lambda}^t \to \boldsymbol{\lambda}^*$ and $\tilde{U}_k^{\beta,cp,t}(\tilde{g}_k, \lambda_k^t) \to \tilde{U}_k^{\beta,cp}(\tilde{g}_k, \lambda_k^*)$ when the following two conditions hold.

- $\frac{\kappa^t}{\tilde{\Delta}T} \to 0$ as $t \to \infty$.
- $\tilde{\Delta}T$ is chosen large enough such that

$$(1-\beta)\sum_{t'=(n-1)\Delta T+1}^{n\Delta T} \beta^{t'}\left\{\mathbf{r}(\mathbf{s}^{\mathbf{t}'}, \boldsymbol{\lambda}) - \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}^t)\right\}$$

is the subgradient of $U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda})$.

Proof: The proof is similar to one presented in [23]. The main ideas are as follows. First, the value function $\tilde{U}_k^{\beta,cp,t}(\tilde{g}_k, \lambda_k)$ converges for any given conjectural price λ_k as shown in Proposition 7. Second, $U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda})$ is a convex and piecewise linear in $\boldsymbol{\lambda}$, which is shown in [17]. In other words, $U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda})$ is continuously differentiable except at finitely many points. Third, since $(1-\beta)\sum_{t'=(n-1)\Delta T+1}^{n\Delta T} \beta^{t'}\left\{\mathbf{r}(\mathbf{s}^{\mathbf{t}'}, \boldsymbol{\lambda}) - \hat{\mathbf{r}}^{t'}(\boldsymbol{\lambda}^t)\right\}$ is the subgradient of $U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda})$, the gradient update converges to the limit of the o.d.e. problem $\dot{\boldsymbol{\lambda}}^t = \nabla U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda}^t)$, which is one of the minimizers of $U^{\text{coop}}(\mathbf{s}, \boldsymbol{\lambda}^t)$, i.e., the efficient conjectural price. ■

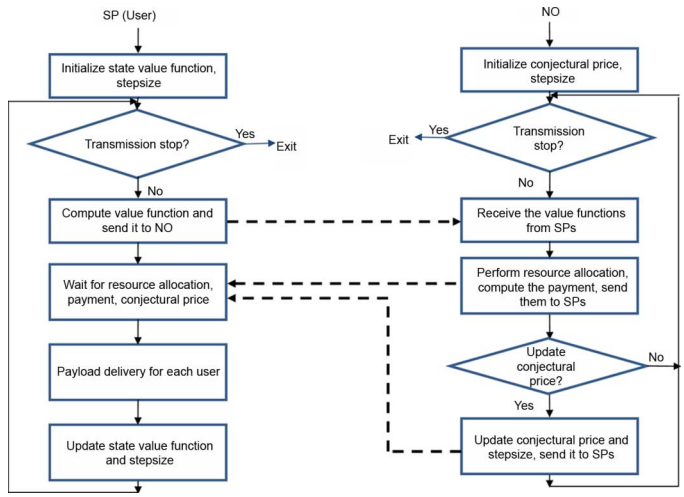

Fig. 5. Online learning and message exchange for the SPs and the NO.

The algorithms for updating the conjectural prices and state-value functions are given in Algorithms 1 and 2. The online learning algorithms and message exchange between the SPs and the NO are illustrated in Fig. 5 as well.

Algorithm 1: Resource Allocation and Conjectural Price Update Performed by NO

Initialize λ_k^1 for $k = 1 \ldots K$
$t = 0, \gamma = 1$
repeat
  $t \Leftarrow t + 1$
  Receive the value function $\theta_k(g_k^t, r_k^t)$ from all the users
  Perform the resource allocation given in (4)
  Compute the payment for each SP as in (5)
  if $t == n\Delta T$ **then**
    Update $\lambda_k^{n+1} \forall k$ as in (27)
    Send λ_k^{n+1} for user $k \forall k$ to SP i s.t. $k \in \mathcal{K}_i$
    Set $\gamma = \frac{1}{n+1}$
  end if
until 0

Algorithm 2: Online Learning for Value Function Estimation for User k Performed by SP i s.t. $k \in \mathcal{K}_i$

Initialize state value $V_k^1(g_k)\ \forall g_k$
$t = 0, \kappa = 1$
repeat
  $t \Leftarrow t + 1$
  Send the value function $\theta_k(g_k^t, r_k^t)$ to the NO
  Wait for resource allocation
  Receive the resource allocation r_k^*, payment τ_i and conjectural price λ_k^n if it is sent by the NO.
  Perform data transmission based on the resource allocation
  Update the value function as in **(28)**.
  Set $\kappa = \frac{1}{t+1}$
until 0

TABLE II
TRAFFIC PARAMETERS FOR EACH USER

Users	1	2	3	4
α_k	2	3	2.5	4
Arrival Rate (packets/msec)	2	2	3	3
Packet Length (kbits)	1.5	1.5	1.5	1.5

VII. SIMULATION RESULTS

We performed various simulations to verify our analytical results as well as to observe behaviors under more heterogeneous conditions. We simulated the wireless network as a time-slotted system with the frame length of 20 ms. The wireless channels for each wireless user are modeled as i.i.d. Rayleigh fading channels. For a Rayleigh fading channel, the channel gain has an exponentially distribution with pdf $f(h) = \frac{1}{\rho}e^{-\frac{h}{\rho}}$, where ρ is the average channel gain. The number of the simulated channels is 3. In this simulation, we use constant power allocation and consider the *QAM* modulation with the constellation size of 2^{ω}, $\omega = 1, \ldots, \Omega$, where Ω determines the highest modulation allowed and the discrete rate adaption is adopted. The set of possible SNR (i.e., the nonnegative real line) is partitioned into $\Omega + 1$ disjoint regions by the boundary points $b_0, b_1, \ldots, b_{\Omega+1}$. Each region corresponds to one channel state. When the SNR falls into the range of $[b_\omega, b_{\omega+1}]$ (i.e., the channel state is ω), then ω bits are loaded for transmission (i.e., the transmission rate is ω bits/s/Hz). As shown in [26], the SNR regions with the average SNR of 20 dB and $\Omega = 8$ is given as follows: $b_0 = -\infty$, $b_1 = 7.3350$ dB, $b_2 = 10.7856$ dB, $b_3 = 14.4654$ dB, $b_4 = 17.7753$ dB, $b_5 = 20.9280$ dB, $b_6 = 24.0078$ dB, $b_7 = 27.0525$ dB, $b_8 = \infty$.

A. Convergence of the Distributed Algorithm

In this section, we consider that, in the network, there are two SPs, each of which has two wireless users. Specifically, SP 1 has Users 1 and 2, and SP 2 has Users 3 and 4. All four users experience the wireless channels with average SNR of 20 dB. The SPs aim to minimize the average delay across their users. In other words, the immediate utility for each user k is $u_k(g_k^t, r_k^t) = -\alpha_k g_k^t$, where g_k^t is the queue length of user k at time t. Poisson process is used for the arrival process. The parameters of each user are shown in Table II.

We demonstrate the convergence of our proposed distributed algorithm in Section VI for both estimating the value function $\theta_k(g_k, r_k)$ and updating the conjectural prices $\boldsymbol{\lambda}$. We compare the convergence results of three scenarios as shown in Table III. The conjectural prices of Users 1 and 2 are illustrated in Figs. 6 and 7. From these figures, we notice that in the three scenarios shown in Table III, the conjectural prices converge to the same price. We have the same observations for the conjectural prices of Users 3 and 4. This verifies that our proposed distributed algorithm is able to optimally learn the value function and conjectural price on line.

B. Manipulation of Conjectural Price

In this section, we show how the SP is penalized when it manipulates the conjectural price. The simulation settings are the same as in Section VII-A. From the simulation results in Section VII-A, we know that the optimal conjectural prices are $[5.3578, 4.8244, 5.3411, 4.7810]$ for Users 1, 2, 3 and 4, respectively.[4] However, in this simulation, we assume that SP 1 does not follow the conjectural price for User 1 advised by the NO. Instead, SP 1 chooses the price arbitrarily from 2 to 8. The utility is computed with $A = 10$. From Fig. 8, we notice that when the conjectural price for User 1 deviates from the advertised one by the NO, the utility of SP 1 is decreased. The optimal conjectural price for User 1 is the advertised one, which shows that SP 1 does not have incentives to deviate from the advertised price. This demonstrates the advertised price is the Nash Equilibrium for the stochastic game.

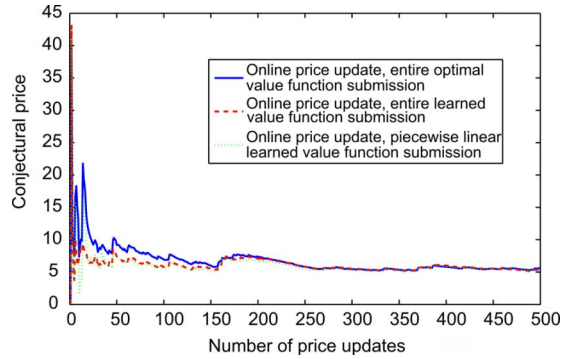

Fig. 6. Conjectural price of user 1.

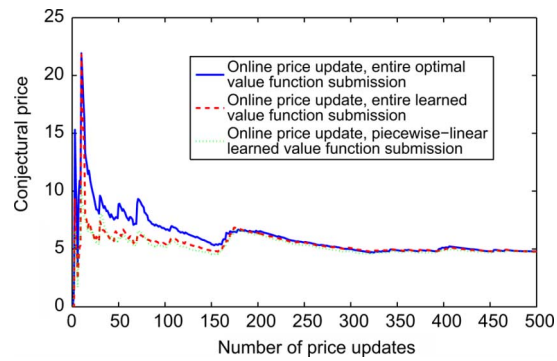

Fig. 7. Conjectural price of user 2.

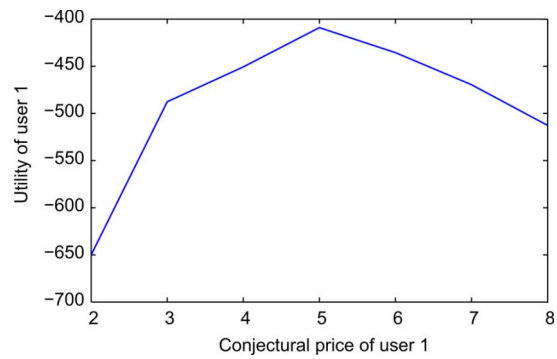

Fig. 8. Manipulating the conjectural price of User 1.

C. Wireless Channel With Nonstationary Dynamics

In this section, we first verify how the conjectural prices vary when the underlying wireless channels are nonstationary (i.e.,

[4]Note that although the channel statistics are the same, the arrival rates and priorities differ across users. Thus, the optimal conjectural prices for distinct users are different in this scenario.

TABLE III
SIMULATION SCENARIOS

Scenarios	conjectural price	value function	bid function
1	online update	optimal value function obtained by solving Bellman's equations	entire value function
2	online update	online estimate	entire value function
3	online update	online estimate	piecewise linear value function

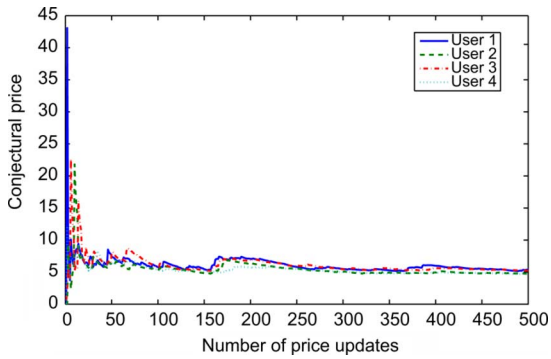

Fig. 9. Conjectural prices of all the users when they experience symmetric channels with average SNR of 20 dB.

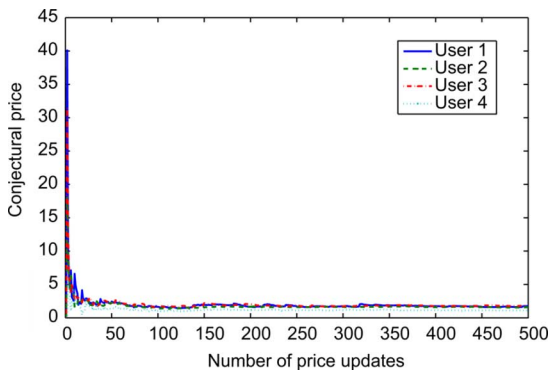

Fig. 10. Conjectural prices of all the users when they experience asymmetric channels with mean of 20 dB for Users 1–3 and mean of 28 dB for User 4.

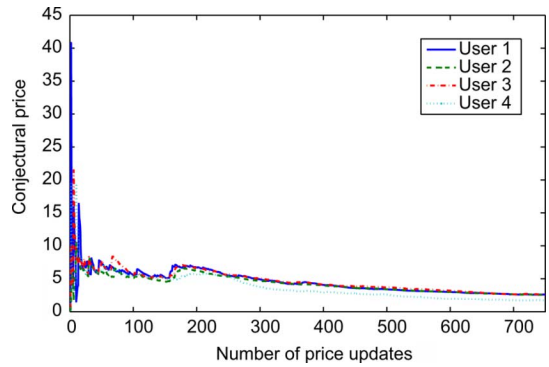

Fig. 11. Conjectural prices of all the users when they experience the channels with time-varying mean of User 4 from 20 to 28 dB.

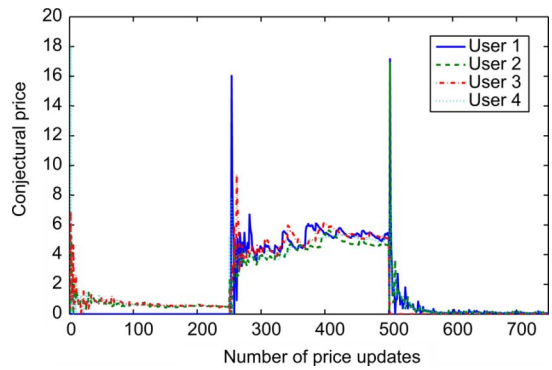

Fig. 12. Conjectural prices of all the users when some user enters into or departs from the network.

the average SNR of the underlying channels are time-varying). We then verify how the conjectural prices are changed when the number of users are changed (i.e., users enter or leave the network.)

We first consider three different channel dynamics: 1) all the four users experience the wireless channels with the average SNR of 20 dB; 2) Users 1–3 experience the channels with the average SNR of 20 dB, but User 4 experiences the channels with the average SNR of 28 dB (i.e., the channel condition of User 4 is improved); 3) Users 1–3 experience the channels with the average SNR of 20 dB, but User 4 experiences the channels with the average SNR varying from 20 to 28 dB (i.e., the channel condition of User 4 is gradually improved). Figs. 9–11 show the changes of the conjectural prices of all the users. Comparing Fig. 9 to Fig. 10, we note that when the channel condition of User 4 is improved, the conjecture prices are reduced from $[5.3578, 4.8244, 5.3411, 4.7810]$ to $[1.6464, 1.5705, 1.8159, 1.1040]$ for all the users. This demonstrates that the improvement in the channel condition of User 4 reduces the congestion level of the whole network, leading to the lower conjectural prices for all users. That is, as expected, it benefits all users in the network. When the channel state of User 4 is nonstationary (in this case, its average SNR is gradually improving), the NO should adapt the conjectural prices in order to capture this change. To accommodate this, we allow the NO to update the conjectural prices using a small constant step size γ instead of the diminishing one. In our simulations, we set $\gamma = 0.01$. The results shown in Fig. 11 capture the change in the conjectural prices: They improve as the channel state of User 4 improves gradually.

We further simulate a scenario for 10 min. We equally divide the time into three periods. In the first period, all the users except User 1 are in the network. In the second period, all the users are in the network (i.e., User 1 enters the network at the beginning of this period). In the third period, User 3 leaves the network. All users experience an average SNR of 20 dB. User join and leave events serve as triggers to reset parameters of Algorithms 1 and 2 as $n = 0$, $t = 0$ and $\gamma = 1$, $\kappa = 1$. Fig. 12 shows how such system dynamics impact the conjectural prices of users in the system. In the first period, the optimum conjectural prices are $[0, 0.5380, 0.5357, 0.5467]$, where zero price for User 1 means that User 1 is not in the

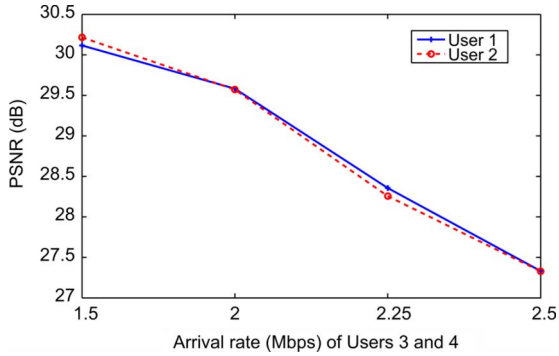

Fig. 13. PSNRs of Users 1 and 2 when Users 3 and 4 have arrival rates of 1.5, 2, 2.25, and 2.5 Mb/s.

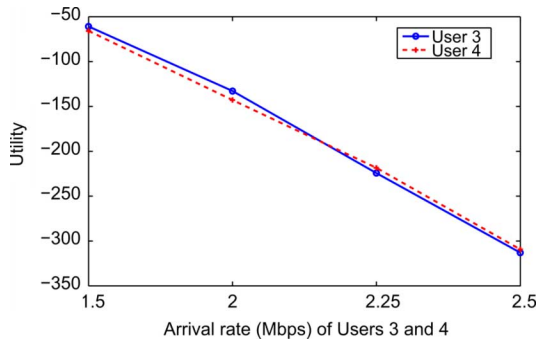

Fig. 14. Reward of Users 3 and 4 when Users 3 and 4 have arrival rates of 1.5, 2, 2.25, and 2.5 Mb/s.

network. In the second period, the conjectural prices are $[5.3021, 4.7193, 5.2071, 4.9867]$. In the third period, the conjectural prices are $[0.0695, 0.0772, 0, 0.0831]$ where User 3 is not in the network. Our conclusion after playing with various system dynamics is that the learning algorithms can very quickly adapt against unpredictable sudden changes in the network as long as the system behaves in a quasi-stationary fashion.

D. SPs With Different Objectives

In this section, we present our results to cover the case where SPs have different objectives. Specifically, we evaluate a scenario where the first SP provides a video streaming service and the other SP provides mobile data service with the same objectives as in Section VII-A. The channel conditions of all the users are symmetric with an average SNR of 20 dB. Both Users 1 and 2 stream the video sequence "Mobile" (CIF resolution, 30 Hz). To compress the video data, we used a scalable video coding scheme [24], which is attractive for wireless streaming applications because it provides on-the-fly application adaptation to channel conditions, support for a variety of wireless receivers with different resource capabilities and power constraints, and easy prioritization of various coding layers and video packets. In this simulation, the average rate of video data for both Users 1 and 2 is 1 Mb/s. The arrival rate for both Users 3 and 4 is the same and varies from 1.5, 2, 2.25, to 2.5 Mb/s. Fig. 13 shows the peak SNR (PSNR) of Users 1 and 2 representing the reconstructed video quality. From this figure, we note that the reconstructed video quality is gradually degraded when the arrival rate of Users 3 and 4 is increased (because the congestion level is increased). We further illustrate the average rewards (i.e., $-\alpha_k g_k$) of Users 3 and 4 in Fig. 14, which also shows that the average rewards of Users 3 and 4 are gradually degraded when the arrival rate is increased. This demonstrates that our proposed algorithm is scalable (gradually degrading the performance) with respect to the service utility functions when the network becomes congested.

VIII. Related Work

Network virtualization is an emerging architectural choice to support concurrent heterogeneous services with various QoS requirements [1], [2]. Wireless network virtualization is a special case at its infancy as a research area [3]–[6]. The resource abstraction for the underlying physical networks and the efficient utilization of these resources are the fundamental problems in constructing the wireless network virtualization. Existing solutions [3], [4] in wireless network virtualization directly divide the available network resources (e.g., wireless spectrum) orthogonally into network slices and mimic the existing spectrum access over shared but virtualized hardware. This requires that the service providers explicitly understand the wireless access protocols. In contrast, our proposed virtualization separates the wireless resource management performed at the NO level from the quality-of-service control performed at the SP level. Furthermore, existing solutions often statically assign the virtualized network resources to the supported services with limited reconfiguration [4], [7] without considering the heterogeneous services and underlying time-varying wireless features (e.g., channel conditions, available spectrum resources, etc.).

Another strongly related body of work for network virtualization looks at the incentives of the service providers over shared physical networks [9] since one SP's decisions about the resource requirement can greatly affect the other SPs'. For the self-interested SPs, auctioning is a promising tool to bring the incentives for the users to efficiently utilize the limited resources [10]. The multiple SP interactions are often analyzed in the context of Internet, where each SP is associated with a predetermined utility function and the equilibrium of the interaction is proved as in [25]. The implicit assumption is that the service and network environment is static, which is not true in our considered wireless networks. The auctions for dynamic wireless resources (e.g., spectrum, transmission time) have been investigated in [11] and [12], in which the network resources are auctioned without considering the heterogeneous services and the dynamics in both the traffic characteristics and the time-varying wireless channel conditions. VCG mechanism, which we use in our framework to ensure revealing of true value functions, is heavily used and investigated in economics and other disciplines [27]. Its problems in bid preparation and costs, revenue deficiency, weak equilibria, information privacy, complexity of winner determination, budget constraints, and strategy proofness in sequences of auctions do not apply to our overall framework due to the following: Value functions for different services are readily available (e.g., we used one for Video and another for average delay in our experiments) and can be efficiently communicated in our case via piecewise linear approximation; NO

is a trusted party and only interested in efficient spectrum allocation rather than in revenue maximization; SPs do not see each other's bids; conjectural prices are introduced and it has provable equilibrium properties; convex optimization renders the computation of equilibrium, pricing, and the winner determination fast; SPs typically have long-term budgets at timescales much larger than the spectrum bidding cycles.

IX. Conclusion

We have presented a new virtualization framework for wireless networks to support multiple heterogeneous self-interested services over the same physical network. Our framework combines VCG auction and conjectural prices to show that spectral efficiency can be achieved while the complex spectrum management and QoS provisioning decisions can be decoupled from each other. We have also developed iterative learning algorithms executed by SPs and the NO that converge to this spectrally efficient point of operation. Our results over various stationary and nonstationary scenarios demonstrate that the convergence is fast and we can handle various dynamics such as variations in average channel qualities and user join/departure events.

For illustration purposes, in this paper, we simplified the channel access strategy where a scheduling interval could be shared at arbitrary fractions of the available channel capacity. Real systems divide the time-frequency resources into resource blocks and make assignment at a finite granularity. Furthermore, only a finite number of rates can be supported. These render the capacity region a nonconvex one in general. Extending the established framework to nonconvex system models remains as a future work.

References

[1] G. Schaffrath, C. Werle, P. Papadimitriou, A. Feldmann, R. Bless, A. Greenhalgh, A. Wundsam, M. Kind, O. Maennel, and L. Mathy, "Network virtualization architecture: Proposal and initial prototype," *Proc. ACM VISA*, pp. 63–72, Aug. 2009.

[2] J. Carapinha and J. Jiménez, "Network virtualization: A view from the bottom," *Proc. ACM VISA*, pp. 73–80, Aug. 2009.

[3] S. Singhal, G. Hadjichristofi, I. Seskar, and D. Raychaudhri, "Evaluation of UML based wireless network virtualization," in *Proc. NGI*, Krakow, Poland, Apr. 2008, pp. 223–230.

[4] R. Mahindra, G. D. Bhanage, G. Hadjichristofi, I. Seskar, D. Raychaudhuri, and Y. Y. Zhang, "Space versus time separation for wireless virtualization on an indoor grid," in *Proc. NGI*, Krakow, Poland, Apr. 2008, pp. 215–222.

[5] R. Mahindra, G. Bhanage, G. Hadjichristofi, S. Ganu, P. Kamat, I. Seskar, and D. Raychaudhuri, "Integration of heterogeneous testbeds," in *Proc. ACM TridentCom*, Mar. 2008, Article no. 27.

[6] G. Smith, A. Chaturvedi, A. Mishra, and S. Banerjee, "Wireless virtualization on commodity 802.11 hardware," in *Proc. ACM WinTECH Workshop*, Montreal, ON, Canada, Sep. 2007, pp. 75–82.

[7] A. Haider, R. Potter, and A. Nakao, "Challenges in resource allocation in network virtualization," in *Proc. 20th ITC Specialist Seminar*, May 2009, pp. 1–9.

[8] A. Chandra, W. Gong, and P. Shenoy, "Dynamic resource allocation for shared data center using online measurements," *Proc. ACM SIGMETRICS*, pp. 300–301, 2003.

[9] C. Courcoubetis and R. Weber, "Economic issues in shared infrastructures," *Proc. ACM VISA*, pp. 89–96, Aug. 2009.

[10] R. Myerson, "Optimal auction design," *Math. Oper. Res.*, vol. 6, pp. 58–73, 1981.

[11] J. Bae, E. Beigman, R. Berry, M. Honig, and R. Vohra, "Sequential bandwidth and power auctions for distributed spectrum sharing," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 7, pp. 1193–1203, Sep. 2008.

[12] X. Zhou, S. Gandi, S. Suri, and H. Zheng, "eBay in the sky: Strategy-proof wireless spectrum auctions," in *Proc. ACM MobiCom*, Sep. 2008, pp. 2–13.

[13] M. O. Jackson, "Mechanism theory," in *The Encyclopedia of Life Support Systems*. Oxford, U.K.: EOLSS, 2003.

[14] P. Maille and B. Tuffin, "Multibid auctions for bandwidth allocation in communication networks," in *Proc. IEEE INFOCOM*, Mar. 2004, vol. 1, pp. 54–65.

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

[16] R. Myerson and B. Roger, *Game Theory, Analysis of Conflict*. Cambridge, MA: Harvard Univ. Press, 1991.

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

[18] F. Fu and M. van der Schaar, "A systematic framework for dynamically optimizing multi-user video transmission," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 308–320, Apr. 2010.

[19] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA: Athena Scientific, 2005.

[20] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[21] N. Salodkar, A. Bhorkar, N. Karandikar, and V. Borkar, "An on-line 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.

[22] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–407, 1951.

[23] V. S. Vorkar, "An actor-critic algorithm for constrained Markov decision processes," *Syst. Control Lett.*, vol. 54, pp. 207–213, 2005.

[24] J. R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 559–571, Sep. 1994.

[25] S. Shakkottai and R. Srikant, "Economics of network pricing with multiple ISPs," *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1233–1245, Dec. 2006.

[26] M. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Wireless Person. Commun.*, vol. 13, pp. 119–143, May 2000.

[27] M. H. Rothkopf, "Thirteen reasons why the Vickrey–Clarke–Groves process is not practical," *Oper. Res.*, vol. 55, no. 2, Mar. 2007.

Fangwen Fu (A'11) received the Bachelor's and Master's degrees in electrical and electronics engineering from Tsinghua University, Beijing, China, in 2002 and 2005, respectively, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 2010.

He currently works with Intel, Folsom, CA, as a Media Architect. He worked as an Intern with IBM T. J. Watson Research Center, Yorktown Heights, NY, and with DOCOMO USA Labs, Palo Alto, CA. He was selected by IBM Research as one of the 12 top Ph.D. students to participate in the 2008 Watson Emerging Leaders in Multimedia Workshop in 2008. His research interests include wireless multimedia streaming, resource management for networks and systems, stochastic optimization, applied game theory, video coding, processing, and analysis.

Dr. Fu received the Dimitris Chorafas Foundation Award in 2009.

Ulas C. Kozat (S'97–M'04–SM'10) received the B.S. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 1997, the M.S. degree in electrical engineering from the George Washington University, Washington, DC, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, in 2004.

He has conducted research under the Institute for Systems Research (ISR) and Center for Hybrid and Satellite Networks (CSHCN), University of Maryland. He worked at HRL Laboratories and Telcordia Technologies Applied Research as a Research Intern. He currently works with DOCOMO Innovations (formerly DOCOMO USA Labs), Palo Alto, CA, as a Principle Researcher and Project Manager. He has been conducting research in the broad areas of computer/communication networks.