Resource Virtualization for Customized Delay-Bounded QoS Provisioning in Uplink VMIMO-SC-FDMA Systems

Xiaofeng Lu, Qiang Ni, Danping Zhao, Wenchi Cheng, and Hailin Zhang

***Abstract*—Wireless Network Virtualization (WNV), which decouples the physical supply process and the service provisioning process, can abstract, isolate and share the physical infrastructure network equipment. This paper studies the resource virtualization in virtual multiple-input multiple-output single-carrier frequency-division-multiple-access (VMIMO-SC-FDMA) uplink systems, where resources are abstracted to hide the complex details of the fading channel and the link rates are virtualized using the statistical method. Furthermore, the virtual link rates are scheduled and instantiated to different slices with customized delay-bounded quality of service (QoS) provisioning. In this scheme, physical mobile network operator (PMNO) is in charge of the network resource at the physical layer while virtual mobile network operators (VMNOs) are responsible for the traffic admission and the slice management at the MAC layer. Furthermore, we build up the resource virtualization problem as a cross-layer Stackelberg game, which has the interactive dual processes based on the QoS exponent: top-to-down sub-game of leaders at the MAC layer and down-to-top sub-game of follower at the physical layer. Using the newly designed functions for PMNO and VMNOs, we develop an effective dynamic algorithm with iterative dual update to meet the optimization targets of PMNO and VMNOs. Simulation results verify the superiority and stability of delay-bounded QoS guaranteed wireless resource virtualization algorithm developed in this paper in terms of convergence, access rate, and delay-outage probability.**

***Index Terms*—Wireless resource virtualization, Stackelberg game, effective bandwidth, effective capacity, resource allocation, admission control.**

I. Introduction

WNV is regarded as a key networking paradigm for diverse network services over a shared wireless network infrastructure [1]. A representative wireless virtual network is composed of PMNOs and VMNOs. PMNO is in charge of network infrastructures such as frequency spectrum, bandwidth resource and so on, while the underlying resources owned by a PMNO are abstracted and isolated into multiple virtual resource (VR) slices. VMNO, which has no physical substrate, rents the VR slices and corresponding functionalities supplied by PMNO, and then operates the allocated VR slices to provide more flexible and customized QoS to their end users independently.

For next generation wireless systems, VMIMO techniques are expected to increase capacity by spreading the spatial resources among multiple users, especially for uplink systems. Since a base station (BS) usually knows downlink data information, downlink resource is easy to be allocated. Then in this paper, VMIMO-SC-FDMA uplink systems are studied for the WNV, where a BS serves multiple users within its coverage. From a PMNO's perspective, its network node such as a BS is partitioned into slices, and each of them represents a virtual mobile network (VMN).

The following aspects need to be considered in the design of WNV:

- Slice isolation: As multiple slices with different requirements coexist in the system, the first problem is the isolation of these slices. In other words, any change in one virtual slice should not introduce any interference to other slices. In addition, there are many levels of isolation, such as the lowest level on hardware, flow level on time-slot, application level on traffic, etc. In this paper, slice isolation is implemented at the MAC layer with virtual link rate, and the customized delay-bounded service of each slice is independent.
- QoS and delay: The purpose of VMNs is to provide flexible and customized services to end users, and there may be many services with different traffic characteristics in VMNs.
- Slice resource requirement and allocation efficiency: Another issue is how to meet the different requirements of all slices simultaneously and allocate the resources efficiently.

The technical aim of the paper is to propose an efficient resource reservation algorithm for WNV with delay-bounded QoS guarantee. There are three goals to be achieved in this paper: virtual network isolation, efficient resource allocation and per-flow delay-bounded QoS guarantee. Compared with the existing literatures on WNV, this paper has the following major contributions:

- To propose a mechanism of resource abstraction and link rate virtualization and instantiation for required delay-bounded QoS provisioning. Unlike wired networks, radio links are variable because of random fading and additive noise of wireless channels. Then, we estimate the channel statistical information for one slice scheduling period

This work was supported by the National Natural Science Foundation of China (61371127 & 61572389 & 61671347 & U1705263), the Key Research and Development Plan of Shaanxi Province (2018ZDCXL-GY-04-06), the EU FP7 CROWN Project under Grant Number PIRSES-GA-2013-610524 and the Royal Society project IEC170324.

Xiaofeng Lu(e-mail: luxf@xidian.edu.cn), Danping Zhao, Wenchi Cheng and Hailin Zhang are with State Key Laboratory of Integrated Services Networks, Xidian University, China. e-mail:(luxf@xidian.edu.cn).

Qiang Ni is with School of Computing and Communications, Lancaster University, LA1 4WA, U.K. (e-mail: q.ni@lancaster.ac.uk).

under the condition of given initial channel state. Based on the finite-state Markov channel (FSMC) model and QoS exponent method, we abstract the resources reserved on the link by using average rate to hide the random characteristics of the fading channel so that effective bandwidth (EB) can be used for virtual rate splitting and sharing. Correspondingly, to instantiate the virtual link service rate, effective capacity (EC) is used to give the actual service capability with specific QoS exponent.

- To formulate the resource virtualization problem as a multi-leader single-follower Stackelberg game, where the VMNOs are considered as the leaders while the PMNO as the follower. The follower sub-game performs resource reservation to meet the data service rate requirements from VMNs, and the leader sub-game maximizes the individual utilities of different VMNOs with the special delay-bounded QoS constraints. During the resource reservation process, the spatial domain of VMIMO and time-frequency domain of adjacent time-frequency resource blocks (RBs) are virtualized and instantiated.
- To develop an iterative cross-layer dual update algorithm to the solution of the virtualization problem, and to address the rate gap between virtualization and instantiation, we design the convex utility functions of PMNO and VMNOs, and we propose an iterative algorithm to dynamically adjust the traffic admission and conjectural service rate until they converge to the Stackelberg Equilibrium.

The rest of this paper is organized as follows. We give a review of the related research work in Section II. Section III describes the system model in terms of VMIMO-SC-FDMA systems and further clarifies the problem solved in this paper. Section IV formulates the above delay-constrained resource virtualization as a Stackelberg game. Section V presents the algorithmic solutions to the formulated problems and the existence of Nash Equilibrium. Section VI presents the simulation results. The conclusions are stated in Section VII.

Here are some notations to be used in this paper :

- $\mathrm{P}(\cdot)$: probability operation.
- $\mathbf{I}_m$, $\mathbf{1}_{m\times n}$: $m\times m$ identity matrix and $m\times n$ matrix with all 1's, respectively.
- $\otimes$: Kronecker product.
- $(\bullet)^T$,$(\bullet)^H$: Transposition, Hermitian, respectively.
- $[\bullet]_{u,u}$: elements at the u-th row and the u-th column of a matrix.
- $\mathrm{E}[\bullet]$: expectation operation.

Throughout this paper, all matrices and sets are denoted by capital letters in boldface, and vectors are denoted by lowercase letters in boldface.

II. RELATED WORK

A. Wireless Resource Virtualization

With the gradual ossification of the Internet, network virtualization has emerged as an important potential solution, and enables deploying customized services on a shared infrastructure. The authors in [2], [3] provide brief surveys on some existing works about wireless virtualization, and also mention the challenges and future directions. Further, in [4], J. van de Belt *et al*. first revisit several key concepts about wireless network virtualization and clarify the difference between abstraction and representation. Then, they develop a theory of virtualization to discuss virtualization in a coherent and structured manner. In existing architectures, wireless network infrastructures, especially in radio access network (RAN) [5-12], are sliced to create wireless virtual resources, which can offer customized services to VMNs by different schedulers in a secure and isolated manner. M. Yang *et al*. only propose OpenRAN, an architecture for software-defined RAN via virtualization [5]. L. Zhao *et al*. propose an air interface virtualization scheme in LTE system, and further study the time-frequency PRB allocation problem for different traffic models [6], [7]. In [8], R. Kokku *et al*. propose and implement a network virtualization substrate for effective virtualization of wireless resources in WiMAX cellular networks. The design provides flow-level virtualization on time-slot to foster a broad set of deployment scenarios and meets the requirements of isolation, customization and resource utilization. Though the above research works have been carried out on wireless resource virtualization, channel status is neglected in network virtualization widely. Further, the authors in [9], [10] study the resource measurement and allocation problems on a flat fading channel. Specifically, in [9], X. Zhang *et al*. propose an information-centric wireless network virtualization technique for the time sensitive multimedia data transmission problem, where the diverse delay-bounded QoS is measured by the EC theory. In [10], F. Fu *et al*. present a new wireless network virtualization framework to support multiple heterogeneous self-interested services over the same physical network. In detail, they model the dynamic interactions among service providers (SPs) and the network operator (NO) as a stochastic game, where the NO focuses on the efficient dynamic resource allocation by abstracting the underlying channel conditions via a time-varying feasible rate region, while SPs only focus on their own service objectives and constraints. The algorithms proposed in [9], [10] are for flat fading channels, but cannot be directly applied to the actual broadband wireless communication system, where channels are with frequency-selective fading channels. Thus, to address this issue, we have done some research works on resource virtualization in orthogonal frequency division multiplexing access (OFDMA) systems under a frequency-selective wireless channel [11], [12], where the virtual resource slices are implemented on subcarrier at the physical layer. However, the delay-bounded QoS provisioning and the virtualization of multi-dimensional resources, such as spatial and time-frequency resources, are not considered. In this paper, we provide a multi-dimensional resource virtualization mechanism. To describe simply, we focus on the resource virtualization in uplink VMIMO-SC-FDMA systems, which can be easily extended to other systems such as MIMO-OFDMA systems.

B. EC and EB

With the gradual diversification of business, more and more new services with strict QoS requirements in terms

of low delays have emerged, such as popular multimedia applications. To characterize the effect of delay on the system, two important concepts, EB [13] and EC [16], have been proposed for matching the source traffic arrival process and the network service process respectively. In [13]-[15], EB has been developed to model the statistical behaviour of traffic. In particular, the theory shows that the queuing constraints are imposed on buffer violation probabilities and specified by the QoS exponent, which indicates the exponential decay rate of the QoS violation probability. Inspired by EB theory, Wu *et al*. define the concept of EC, which provides the maximum constant arrival rate that can be supported by a given channel service process while satisfying a statistical QoS requirement specified by the QoS exponent [16]. The analysis and application of EC in various settings have attracted much interest recently [17]-[19]. Since EB and EC mentioned above facilitate capturing the delay-bounded constraint of wireless link without going into complex queuing analysis, we use these dual concepts to characterize the delay-bounded constraints in our cross-layer resource virtualization scheme.

C. Wireless Channel Model and VMIMO-SC-FDMA

The FSMC model has been widely used to model the wireless fading channel mathematically, which is analytically tractable and can provide closed-form results. Specifically, a useful FSMC model is designed to represent the Rayleigh fading channel according to the received SNR [20]. In [21], M. Hassan *et al*. propose a new partitioning approach that results in a FSMC model with tractable queuing performance. The authors in [22] introduce the Gauss-Markov model to describe the MIMO channel matrix and derive the bounds of the ergodic capacity in closed-form. Similarly, S. H. Ting *et al*. [23] propose a Markov-kronecker model for analysis of time varying channel in MIMO systems.

VMIMO, also called multiuser MIMO, refers to a communication system where a BS with multiple antennas serves two or more single antenna users on the same frequency band and time slot. Compared with the conventional MIMO system, VMIMO can obtain additional multiuser diversity gain by grouping users using well-designed strategies [24]-[26].

In order to obtain both multiuser diversity gain and frequency selective gain, VMIMO is usually applied to uplink SC-FDMA systems [24]-[26]. In [24] and [25], joint resource allocation algorithms are proposed for uplink VMIMO-SC-FDMA systems with fixed 2-user pairing. In [26], dynamic user grouping and joint resource allocation algorithm is proposed for multi-cell uplink VMIMO-SC-FDMA systems.

D. Stackelberg Game

The Stackelberg model, also known as "leader-follower model", is a first-mover advantage model in which the leader first takes advantage of the competition. To date, Stackelberg game has been considered as a powerful tool to analyze interactive decision-making processes in the resource allocation problems, and has been extensively studied and adopted in various fields. Specifically, in [27], K. Zhu *et al*. formulate the power control hierarchical competition between the macro BSs and small cell BSs as a distribution-free multi-leader multi-follower robust Stackelberg game, where the MBSs are the players of the sub-game and the SBSs are the players of the follower sub-game. In [28], [29], H. zhang *et al*. model the resource allocation and pricing problem in the unlicensed spectrum as a multi-leader multi-follower Stackelberg game. S.Ji *et al*. propose a dual power allocation algorithm based on Stackelberg game to maximize the utilities of users and networks, where the network plays a role as a leader, while users as followers [30]. The authors in [31] employ game theoretic approaches to model the problem of minimizing energy consumption as a Stackelberg game. In this paper, VMNOs first make traffic flow access strategy. Then, according to the accessed traffic flows of the VMNOs, the PMNO makes the optimal resource allocation strategy. Thus, it is well suited to be modeled as a Stackelberg game model, where the VMNOs are considered as leaders while the PMNO as the follower.

III. SYSTEM MODEL AND RESOURCE MEASUREMENT

In this section, we introduce the model of WNV, the multi-cell VMIMO-SC-FDMA uplink systems and the resources in spatial and time-frequency domains. Then, we derive the resource measurement of the physical resources based on the FSMC model.

A. Model of WNV and Basics of VMIMO-SC-FDMA Uplink Systems

In this paper, we mainly study a multi-cell uplink VMIMO system including N_c cells, where each cell consists of one BS equipped with N_r receiving antennas and N_u users with single transmitting antenna. Besides, a centralized controller is required to determine the user grouping, resource allocation, and multi-cell information combining.

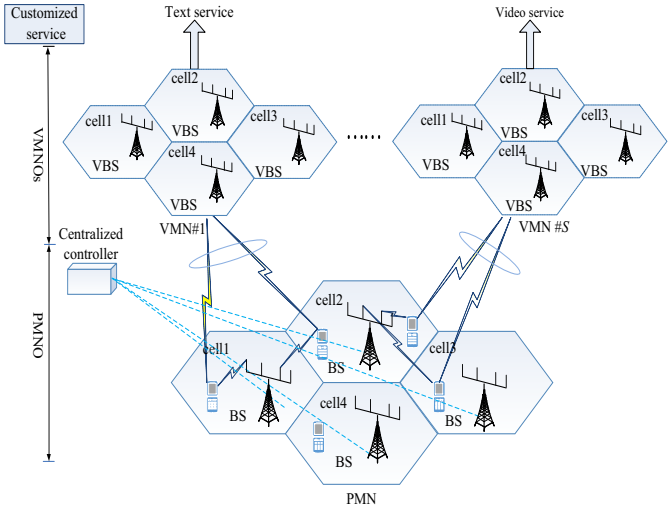


Fig. 1: Model of the WNV.

Further, the user schedulers in centralized controller selects N_t users from $K = N_c N_u$ users to form a VMIMO system,

and the multi-cell joint processing is adopted at BSs side. Then, multiple user equipment in one group transmit the signal to all cells on the same RB, and the centralized controller combines the information received by different cells to further improve the system performance. The channel gain from the u-th user of the j-th cell to the f-th antenna of the BS in the l-th cell is given by $h_{f,l,u,j} = \sqrt{\beta_{f,l,u,j}}\gamma_{f,l,u,j}$, where $\gamma_{f,l,u,j}$ is the small scale fading factor which is independent and identically distributed zero mean, circularly-symmetric complex Gaussian CN(0,1) random variables, and $\beta_{f,l,u,j}$ is the large scale fading coefficient which models the geometric attenuation and shadow fading that are assumed to be constant over a coherence time and known a priori.

In the WNV model, the network node such as a BS is partitioned into slices, and each of them represents a VMN respectively. As depicted in Figure 1, a physical mobile network (PMN) is split into S VMNs which support different users with different delay-bounded requirements.

As shown in Figure 2, let $\mathbf{\Omega} = \{\Omega_1, \cdots, \Omega_g, \cdots, \Omega_{|\mathbf{\Omega}|}\}$ denote the set of user groups and γ_{Ω_g} denote the set of users in user group Ω_g. Then $\left|\gamma_{\Omega_g}\right|$ corresponds to N_t mentioned above. Further, assuming user group Ω_g is scheduled in M_g consecutive subcarriers with the first index p_g, we write the received signal vector of user group Ω_g before MIMO detector as

$$\mathbf{Y}_{p,g} = \mathbf{H}_{p,g}\mathbf{X}_{p,g} + \mathbf{n}_{p,g}, \tag{1}$$

for $p = p_g, p_g + 1, ..., p_g + M_g - 1$, $M_g = N_{RB}^g N_{sc}^{RB}$, where N_{RB}^g and N_{sc}^{RB} denote the number of RBs occupied by user group Ω_g and the number of subcarriers in one RB respectively, $\mathbf{H}_{p,g}$ is the $N_c N_r \times N_t$ virtual MIMO channel matrix, $\mathbf{X}_{p,g}$ is the $N_t \times 1$ transmitting signal vector, $\mathbf{n}_{p,g}$ is the $N_c N_r \times 1$ zero-mean additive white Gaussian noise (AWGN) vector with covariance matrix $\mathrm{E}\left\{\mathbf{n}_{p,g}{\mathbf{n}_{p,g}}^H\right\} = \sigma^2 \mathbf{I}_{N_c N_r}$. Perfect power control over user groups is assumed in this paper. Thus, the total transmitting power of each user group signal vector $\mathbf{X}_{p,g}$ is constrained to E_s, and the transmitting power of each user signal is normalized to $\frac{E_s}{N_t}$.

With the minimum mean square error (MMSE) frequency-domain equalization, the transmitting symbol vector can be estimated by

$$\hat{\mathbf{X}}_{p,g} = (\sigma^2 \mathbf{I}_{N_c N_r} + {\mathbf{H}_{p,g}}^H \mathbf{H}_{p,g})^{-1} {\mathbf{H}_{p,g}}^H \mathbf{Y}_{p,g}, \tag{2}$$

where σ^2 denotes the spectral density power of noise. Then, we get the post-processing SINR of user-u after MIMO equalization as [26]

$$\mathrm{SINR}_{p,u} = \frac{E_s}{\sigma^2 \left[\left({\mathbf{H}_{p,g}}^H \mathbf{H}_{p,g} + \frac{\sigma^2}{E_s}\mathbf{I}_{N_t}\right)^{-1}\right]_{u,u}} - 1. \tag{3}$$

For uplink SC-FDMA systems, adjacent time-frequency RBs should be assigned to one user group. Let N_{RB} denote the number of RBs, when the RB pattern contains only one RB, that is, only one RB is assigned to one user group, the number of RB pattern is N_{RB}. When the RB pattern contains two RBs, that is, two RBs are assigned to one user group, the number of RB pattern is $N_{RB}-1$. In this way, when the RB pattern contains N_{RB} RBs, that is, N_{RB} RBs are assigned to one user group, the number of RB pattern is 1. Therefore, the total number W of RB allocation patterns can be computed by $W = N_{RB} + (N_{RB}-1) + (N_{RB}-2) + \cdots + 2 + 1 = \frac{N_{RB}(N_{RB}+1)}{2}$, and the resource pattern matrix $\mathbf{T}$ can be designed as follows:

$$\mathbf{T}_{N_{RB}\times W} = \begin{array}{c} \text{pattern} \\ \\ RB_1 \\ RB_2 \\ \vdots \\ RB_{N_{RB}} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & W \end{array} \\ \begin{bmatrix} 1 & 0 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}, \tag{4}$$

where RB_n denotes the n-th RB, each element indicates whether the RB is involved in the RB pattern (1) or not (0).

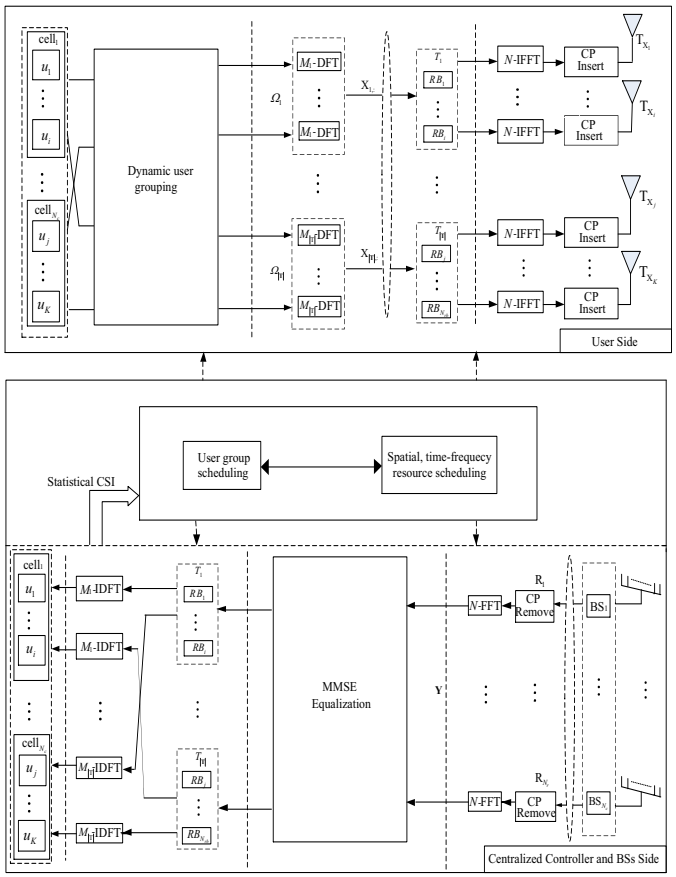


Fig. 2: Block diagram structure of VMIMO-SC-FDMA uplink system.

The available spatial resources come from the user grouping process. For the user group $\Omega_g = (u_1, u_2, \cdots, u_m)$, we can write the group index g of Ω_g as [25], [26]

$$g = \begin{cases} \sum_{j=1}^{m-1} \Delta_j + (u_m - u_{m-1}), m > 1 \\ u_1 \qquad\qquad\qquad\qquad\quad , m = 1 \end{cases}, \tag{5}$$

where $\Delta_j = \begin{pmatrix} K - u_{j-1} \\ m - j + 1 \end{pmatrix} - \begin{pmatrix} K - (u_j - 1) \\ m - j + 1 \end{pmatrix}$, $\{u_1, u_2, \cdots, u_m\}$ in group Ω_g indicates the user index from user set $\mathcal{K} = \{1, \cdots, k, \cdots, K\}$. Further, we describe the user

grouping matrix as follows:

$$\mathbf{B}=\left[\mathbf{B}^{(1)},\mathbf{B}^{(2)},\cdots,\mathbf{B}^{(K)}\right], \tag{6}$$

where $\mathbf{B}^{(k)}$ denotes the fixed-k user grouping matrix. Then, we take $\mathbf{B}^{(2)}$ as an example, and it is designed as follows:

$$\begin{array}{c} \text{group index} \\ \\ \mathbf{B}^{(2)}= \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & \cdots & \left|\mathbf{\Omega}^{(2)}\right|-1 & \left|\mathbf{\Omega}^{(2)}\right| & \text{user index} \end{array} \\ \left[\begin{array}{ccccc} 1 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ K-1 \\ K \end{array}. \tag{7}$$

B. The EB and QoS Exponent

Considering a queueing system with stationary ergodic arrival and service processes, the probability that the queue length exceeds a certain threshold C satisfies the following formula [17]:

$$-\lim_{C\to\infty}\frac{\log\left(\mathrm{P}\left\{Q\left(\infty\right)\geq C\right\}\right)}{C}=\theta, \tag{8}$$

where $\mathrm{P}\left\{a\geq b\right\}$ is the probability that $a\geq b$ holds, $Q\left(\infty\right)$ represents the steady-state queue length, θ is QoS exponent, which indicates the exponential decay rate of the QoS violation probability. A looser QoS requirement can be implied by a smaller θ, while a more stringent QoS requirement can be implied by a larger θ. $\theta\to 0$ means that there is no delay-bounded constraint, which represents that the system can tolerate unlimited long delay.

Based on the above concept of QoS exponent, the EB function, for a stationary ergodic traffic data arrival process $\{A\left(t\right),t\geq 0\}$, can be described as follows:

$$EB\left(\theta\right)=\frac{1}{\theta}\lim_{t\to\infty}\frac{1}{t}\log\mathrm{E}\left[e^{\theta A(t)}\right], \tag{9}$$

where $A\left(t\right)$ represents the amount of arrival traffic data over the time interval $[0,t)$.

In this paper, we make research in different scheduling periods. Thus, optimization condition $t\to\infty$ can be transformed into $MT_s\to\infty$, i.e, $M\to\infty$. Then, the EB function can be accumulated by the amount of arrival traffic data in each scheduling period, and the equation can be written as follows:

$$EB\left(\theta\right)=\frac{1}{\theta}\lim_{M\to\infty}\frac{1}{MT_s}\log\mathrm{E}\left[e^{\theta[A_1(T_s)+A_2(T_s)+\cdots+A_M(T_s)]}\right], \tag{10}$$

where $A_m\left(T_s\right)$ denotes the amount of traffic data in the m-th scheduling period. Because the traffic flows in different scheduling periods are independent, the EB function can be simplified as:

$$\begin{aligned} EB(\theta) &= \tfrac{1}{\theta}\lim_{M\to\infty}\tfrac{1}{MT_s}\log\left\{\mathrm{E}\left[e^{\theta A_1(T_s)}\right]\cdots\mathrm{E}\left[e^{\theta A_M(T_s)}\right]\right\} \\ &= \lim_{M\to\infty}\tfrac{1}{M}\sum_{m=1}^{M}\widetilde{EB}_m\left(\theta\right) \end{aligned}, \tag{11}$$

where $\widetilde{EB}_m\left(\theta\right)=\frac{1}{\theta T_s}\log\mathrm{E}\left[e^{\theta A_m(T_s)}\right]$.

C. The EC of Sub-channel Based on FSMC Model

In this paper, we consider a RB as a sub-channel, which is comprised of consecutive N_{sc}^{RB} subcarriers. Due to the time-frequency correlation, we assume all subcarriers have the same channel state information (CSI) in one RB, which can be obtained by taking the average of the CSIs of the subcarriers within the RB. In addition, the received signal undergoes the flat Rayleigh fading in a typical sub-channel. To keep procedures simple, we drop the subscripts of subcarrier/RBs and groups in the description of this section.

Assuming that the channel state transfers L times in one scheduling period, as shown in Figure 3, one scheduling period T_s consists of L frame slots T_p, that is, $T_s{=}LT_p$.

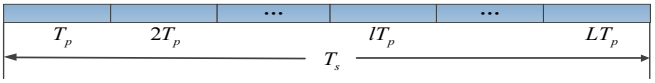


Fig. 3: The frame slots in one scheduling period.

For each sub-channel, VMIMO with N_t transmitting and N_cN_r receiving antennas is used. We assume that the fading gains between all antenna pairs are independent and identically distributed (i.i.d.) Rayleigh fading, and vary according to a Gauss-Markov model, which is widely adopted to describe the channel variation. Using this model, at time instance l, the channel matrix can be written as [22]

$$vec\left(\mathbf{H}\left(l\right)\right)=\sqrt{\alpha}vec\left(\mathbf{H}\left(l-1\right)\right)+\sqrt{1-\alpha}\mathbf{u}\left(l\right), \tag{12}$$

where $0\leq\alpha\leq 1$, $\mathbf{H}\left(l\right)$ denotes the channel matrix of the l-th frame slot in one scheduling period, α is the channel de-correlation coefficient, which can be determined by $\alpha^{T_c/(2T_s)}=r_{hh}\left(T_c\right)$, where T_c is the channel coherence time, $r_{hh}\left(t\right)$ denotes the time autocorrelation function. In addition, $\mathbf{u}\left(l\right)\in C^{N_tN_cN_r\times 1}$ is independent with $\mathbf{H}\left(l-1\right)$. $\mathbf{u}\left(l\right)$ and $\mathbf{H}\left(l\right)$ have i.i.d. complex Gaussian $CN\left(0,1\right)$ entries respectively.

In addition, let $\mathbf{\Gamma}=\left[\Gamma_1,\Gamma_2,\ldots,\Gamma_{Q+1}\right]^T$ represent the thresholds of the element h of the matrix $\mathbf{H}$ in increasing order with $\Gamma_1=0$ and $\Gamma_{Q+1}=\infty$. h is in state q if the value of h is between Γ_q and Γ_{q+1}. The value of h in state q is denoted as h_q, and $h_q=\Gamma_q$. $\mathbf{H}$ can be designed as follows:

$$\mathbf{H}=\left[\begin{array}{cccc} h_{1,1} & h_{1,2} & \cdots & h_{1,N_t} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N_t} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N_cN_r,1} & h_{N_cN_r,2} & \cdots & h_{N_cN_r,N_t} \end{array}\right], \tag{13}$$

where $h_{i,j}$ denotes the channel gain between the i-th receiving antenna and the j-th transmitting antenna, and takes values from the set $\{h_q|q=1,2,\cdots,Q\}$. Thus, we can get the state space of matrix $\mathbf{H}$, that is $\mathbf{H}\in\left\{\mathbf{H}_d|d=1,2,\cdots,Q^{N_tN_cN_r}\right\}$, where $\mathbf{H}_d$ denotes the d-th state of the matrix $\mathbf{H}$. Specifically, $\mathbf{H}_d$ is described as follows:

$$\mathbf{H}_d=\left[\begin{array}{cccc} h_{d_1} & h_{d_{N_cN_r+1}} & \cdots & h_{d_{N_cN_r(N_t-1)+1}} \\ h_{d_2} & h_{d_{N_cN_r+2}} & \cdots & h_{d_{N_cN_r(N_t-1)+2}} \\ \vdots & \vdots & \ddots & \vdots \\ h_{d_{N_cN_r}} & h_{d_{2N_cN_r}} & \cdots & h_{d_{N_tN_cN_r}} \end{array}\right], \tag{14}$$

where $\left\{h_{d_\gamma}|\gamma=1,2,\cdots,N_tN_cN_r\right\}$ takes values from $\{h_q|q=1,2,\cdots,Q\}$.

When the initial state information in one scheduling period $\mathbf{H}(1)$ is given, and all elements of $\mathbf{H}(1)$ are h_1, we can compute the probability distribution of $\mathbf{H}(2)$ according to Eq. (11). The specific equation can be designed as follows:

$$\mathrm{P}(\mathbf{H}(2)=\mathbf{H}_d)=\int_{h_{d_1}}^{h_{d_1+1}}f_h(h_1)dh\int_{h_{d_2}}^{h_{d_2+1}}f_h(h_1)dh\cdots\int_{h_{d_{N_tN_cN_r}}}^{h_{d_{N_tN_cN_r+1}}}f_h(h_1)dh, \tag{15}$$

where $\int_a^b f(h)dh$ denotes the integral operation of the function $f(h)$ in the interval $[a,b]$, $f_h(h_q)$ denotes the probability density function of h that follows $CN(\sqrt{\alpha}h_q,1-\alpha)$.

In this way, we can get the probability distribution of $\mathbf{H}$ in all frame slots in one scheduling period $\{\mathbf{H}(l)\mid l=1,2,\cdots,L\}$.

Then, according to the Eq. (3), the capacity in the l-th frame slot using post-processing SNR can be expressed as [26]

$$r_l=\sum_{u\in\Omega_g}\sum_{d=1}^{Q^{N_tN_cN_r}}\log_2\left(\frac{E_s}{\sigma^2\left[\left(\mathbf{H}_d{}^H\mathbf{H}_d+\frac{\sigma^2}{E_s}\mathbf{I}_{N_t}\right)^{-1}\right]_{u,u}}\right)\mathrm{P}(\mathbf{H}(l)=\mathbf{H}_d). \tag{16}$$

Further, the sub-channel capacity in one scheduling period can be written as follows:

$$\begin{aligned}R&=\frac{(r_1+\cdots+r_l\cdots+r_L)T_p}{T_s}\\&=(r_1+\cdots+r_l\cdots+r_L)/L\end{aligned}. \tag{17}$$

Under the condition of given $\mathbf{H}(1)$, we can get the sub-channel capacity in the 1st frame in one scheduling period r_1. Further, according to the nature of the Markov process, the probability distribution of the sub-channel capacity can be written as follows:

$$\begin{aligned}&\mathrm{P}\left(R=R^j\,|r_1\right)=\mathrm{P}\left(r_1+\cdots+r_L=LR^j\,|r_1\right)\\&=\sum_{\xi=1}^{|\xi|}\mathrm{P}\left(r_2=r_2^\xi,\cdots,r_L=LR^j-\sum_{l=2}^{L-1}r_l^\xi-r_1\,|r_1\right)\\&=\sum_{\xi=1}^{|\xi|}\mathrm{P}\left(r_2=r_2^\xi\,|r_1\right)\cdots\mathrm{P}\left(r_L=LR^j-\sum_{l=1}^{L-1}r_l^\xi-r_1\,\middle|\,r_{L-1}=r_{L-1}^\xi\right)\end{aligned}, \tag{18}$$

where ξ denotes the number index of state combinations, $|\xi|$ denotes the total number of the combinations, r_l^ξ denotes the sub-channel capacity of $\xi-th$ combination in the $l-th$ frame in one scheduling period and takes value from the capacity corresponding to all states. Therefore, under the condition of given initial channel state information, the average sub-carrier channel capacity in one scheduling period can be obtained as follows:

$$\mathrm{E}[R]=\sum_{j=1}^{J}\mathrm{P}\left(R=R^j\,|r_1\right)R^j, \tag{19}$$

where J denotes the number of channel capacity in the scheduling period.

As the dual concept of the EB source model, the EC is defined as the maximum constant arrival rate that a given channel service process can support in order to guarantee the QoS requirement specified by QoS exponent θ [16][32]. For a stationary ergodic service process, $S(t)$ denotes the amount of data that the channel service counter can provide over the time interval $[0,t)$. Then, $S(t)=\tilde{R}t$, and the EC function can be designed as follows:

$$\begin{aligned}EC(\theta)&=-\tfrac{1}{\theta}\lim_{t\to\infty}\tfrac{1}{t}\log\mathrm{E}\left[e^{-\theta S(t)}\right]\\&=-\tfrac{1}{\theta}\lim_{t\to\infty}\tfrac{1}{t}\log\mathrm{E}\left[e^{-\theta\tilde{R}t}\right]\end{aligned}, \tag{20}$$

where $\tilde{R}$ denotes the channel service rate over the time interval $[0,t)$.

Similar to the derivation process of EB function, the EC function can be rewritten as:

$$EC(\theta)=-\frac{1}{\theta}\lim_{M\to\infty}\frac{1}{MT_s}\log\mathrm{E}\left[e^{-\theta(R_1T_s+R_2T_s+\cdots+R_MT_s)}\right], \tag{21}$$

where R_m is the channel service rate in the m-th scheduling period. Because the service rates in different scheduling periods are independent, the EC function can be simplified as:

$$\begin{aligned}EC(\theta)&=-\tfrac{1}{\theta}\lim_{M\to\infty}\tfrac{1}{MT_s}\left\{\log\left[\mathrm{E}\left(e^{-\theta R_1T_s}\right)\cdots\mathrm{E}\left(e^{-\theta R_MT_s}\right)\right]\right\}\\&=-\lim_{M\to\infty}\tfrac{1}{M}\sum_{m=1}^{M}\widetilde{EC}_m(\theta)\end{aligned}, \tag{22}$$

where $\widetilde{EC}_m(\theta)=-\frac{1}{\theta T_s}\log\mathrm{E}\left(e^{-\theta R_mT_s}\right)$, and $\mathrm{E}\left(e^{-\theta R_mT_s}\right)=\sum_{j=1}^{J}\mathrm{P}\left(R=R^j\,|r_1\right)e^{-\theta R^jT_s}$.

Because we make research in each scheduling period, we drop the subscripts of scheduling period to keep description simple in the following.

D. Link Rate Virtualization and Instantiation

3GPP 5G network architecture will involve the integration of several cross-domain networks, and the 5G systems will be built to enable logical network slices across multiple domains. The proposal of this manuscript is implemented in radio access network (RAN) part and focuses on RAN slice. Advanced orchestration and management are required to release the configuration burden from users and enable an integrated end-to-end network slicing. The RAN part of 5G is significantly different from the core network (CN) and transport domains. It is difficult to virtualize RAN due to the diversity of wireless access technologies which are adapted to the random fading of wireless channels. The proposal provides a scheme of resource abstraction to hide the complex details of the fading channel. Thus, the main SDN principle of separating the control and user planes can also be adopted for the RAN. Moreover, the resource virtualization and instantiation algorithm proposed in this manuscript can be considered as the potential control plane function operated by the PMNO and VMNOs.

Therefore, in this paper, we take the mean value of the link rates as its logic representation, that is, the virtual rate service of resources on the link. Correspondingly, EC, which is the actual rate service provided by the resources on the link, is considered as virtualization instance of the link rate.

Figure 4 shows the link rate virtualization and instantiation for the m-th scheduling period. During the virtualization process, on the one hand, we abstract the resources using average rate to hide the complex details or concrete realities of the fading channel, and then, the rate service can be considered as constant and EB theory can be applied for the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2886337, IEEE Transactions on Communications

resource reservation of VMNs with the delay-bounded QoS guarantee. On the other hand, to satisfy the delay-bounded QoS provisioning of traffic flows in each VMN with actual rate service, EC is used to provide the real service capability of the fading channel. Though these two processes may have a small rate gap, we can dynamically adjust the traffic flow access and the resource reservation to address the gap. The details are described below.

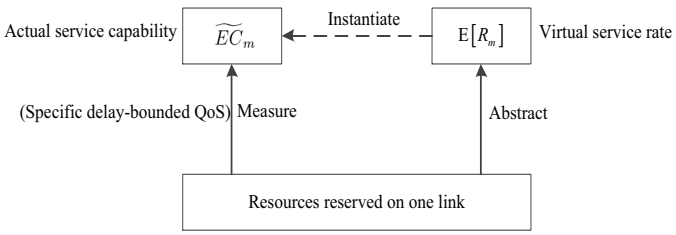


Fig. 4: Link rate virtualization and instantiation for the m-th scheduling period.

E. Resource Abstraction Based on Gaussian-like Fitting

Since the fading process that the signal undergoes is independent in each sub-channel, considering the complexity of analysing the probability distribution of the channel capacity, we give a Gaussian-like fitting method in the following.

Assuming that the Gaussian-like fitting curve $\widehat{R}$ obeys $N(A,\mu,\sigma^2)$, which means that the probability density function of $\widehat{R}$ is $Ae^{-\frac{(\hat{R}-\mu)^2}{2\sigma^2}}$, we give the optimization problem as follows:

$$\min \left|\widehat{EC}-\widetilde{EC}\right| \tag{23}$$

subject to

$$\sum_{j=1}^{J}\mathrm{P}(\widehat{R}=R^j)=1, \tag{23a}$$

where $\mathrm{P}(\widehat{R}=R^j)=\int_{R^j}^{R^{j+1}}e^{-\theta\hat{R}T_s}Ae^{-\frac{(\hat{R}-\mu)^2}{2\sigma^2}}d\hat{R}$ denotes the probability that Gaussian-like fitting capacity $\widehat{R}$ equals R^j, and $\widehat{EC}=-\frac{1}{\theta T_s}\log\left[\sum_{j=1}^{J}\int_{R^j}^{R^{j+1}}e^{-\theta\hat{R}T_s}Ae^{-\frac{(\hat{R}-\mu)^2}{2\sigma^2}}d\hat{R}\right]$ denotes the corresponding EC after the Gaussian-like fitting.

Further, we solve the above problem by the least squares fitting method, that is, we reformulate the optimization problem (23) as $\min\sum_{j=1}^{J}\left[\int_{R^j}^{R^{j+1}}e^{-\theta\hat{R}T_s}Ae^{-\frac{(\hat{R}-\mu)^2}{2\sigma^2}}d\hat{R}-e^{-\theta R^jT_s}\mathrm{P}(R=R^j)\right]^2$. Then, we obtain the optimal A, μ and σ^2.

Finally, after the Gaussian-like fitting process, $\widehat{EC}$ is close to $\widetilde{EC}$, and it is computed by:

$$\begin{aligned}\widehat{EC}&=-\tfrac{1}{\theta T_s}\log\mathrm{E}\left(e^{-\theta\hat{R}T_s}\right)\\&=-\tfrac{1}{\theta T_s}\log\left(\int_{-\infty}^{+\infty}e^{-\theta\hat{R}T_s}Ae^{-\frac{(\hat{R}-\mu)^2}{2\sigma^2}}d\hat{R}\right)\\&=\tfrac{1}{\ln 2}\mu-\tfrac{\theta T_s\sigma^2}{2\ln 2}-\tfrac{\log(2\pi)+2\log(A\sigma)}{2\theta T_s}\end{aligned}. \tag{24}$$

From the above formula, it can be known that $\widehat{EC}$ is a linear function of μ. In the resource virtualization phase, we compute the virtual service rate $\mathrm{E}[R]$ with Gaussian-like fitting $\widehat{R}$, that is, $\mathrm{E}[R]$ is equal to μ, thus $\widehat{EC}$ is a linear function of $\mathrm{E}[R]$. Besides, because $\widehat{EC}$ is close to $\widetilde{EC}$, $\widehat{EC}$ is nearly a linear function of $\mathrm{E}[R]$.

IV. PROBLEM FORMULATION

In this section, we consider the resource virtualization problem in the multi-cell wireless network virtualization environment, where PMNO owns physical underlying resources and VMNOs provide flexible and customized services to end users by renting resources from PMNO. And we formulate the problem as a cross-layer Stackelberg game[27]-[31], where VMNOs are the players of the leader sub-game at the MAC layer and PMNO is the player of the follower sub-game at the physical layer respectively.

A. Optimization Scheme Using Stackelberg Game

In our model, $\{\boldsymbol{\delta},\mathbf{Y}\}$ are the critical parameters in Stackelberg game between PMNO and VMNOs, where $\boldsymbol{\delta}$ is RB pattern allocation vector, and $\mathbf{Y}=[\mathbf{Y}_1,\cdots,\mathbf{Y}_s,\cdots,\mathbf{Y}_S]$ is traffic flow access matrix, where $\mathbf{Y}_s$ is the traffic flow access vector in the s-th VMN. Based on the duality and convex optimization theory, an iterative method is designed to obtain the optimal RB pattern allocation vector $\boldsymbol{\delta}^*=\mathrm{argmax}\,U^{\mathrm{PHY}}\big|_{\mathbf{Y}}$ and the optimal traffic flow access vector in the s-th VMN $\mathbf{Y}_s^*=\mathrm{argmax}\,U_s^{\mathrm{MAC}}\big|_{\boldsymbol{\delta}}$, where U^{PHY} is the utility function at the physical layer and U_s^{MAC} is the utility function of the s-th VMN at the MAC layer.

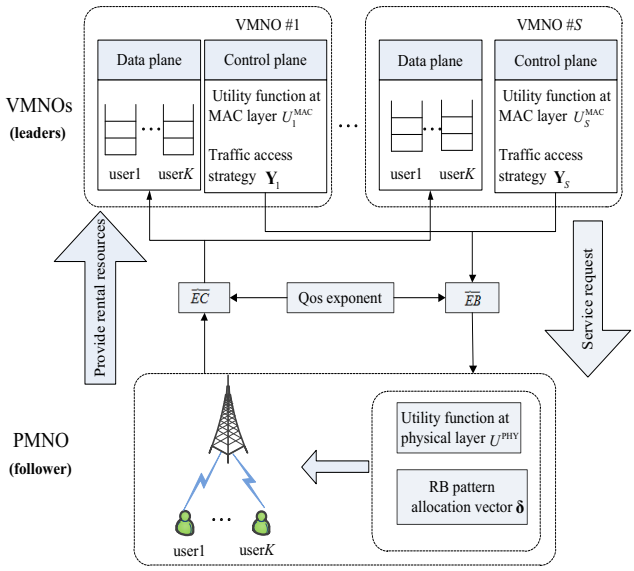


Fig. 5: The interaction diagram between PMNO and VMNOs.

Figure 5 shows the leaders-follower dynamic interaction game between PMNO and VMNOs in one scheduling period. First, at the MAC layer, VMNOs set the traffic flow access strategy of each VMN $\{\mathbf{Y}_1,\cdots,\mathbf{Y}_s,\cdots,\mathbf{Y}_S\}$, and present the $\widetilde{EB}$ requirement of each user to the physical layer. Second,

$$\begin{array}{l} \text{user index} \quad 1 \quad 2 \quad \cdots \quad u_1 \quad \cdots \quad u_m \quad \cdots \quad K \ \text{pattern index} \\ \mathbf{E}\left[\mathbf{R}\right]_{W\times K}^{g} = \begin{bmatrix} 0 & 0 & \cdots & \mathrm{E}\left[R_{1,g,u_1}\right] & \cdots & \mathrm{E}\left[R_{1,g,u_m}\right] & \cdots & 0 \\ 0 & 0 & \cdots & \mathrm{E}\left[R_{2,g,u_1}\right] & \cdots & \mathrm{E}\left[R_{2,g,u_m}\right] & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathrm{E}\left[R_{W,g,u_1}\right] & \cdots & \mathrm{E}\left[R_{W,g,u_m}\right] & \cdots & 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ W \end{matrix} \end{array} . \tag{25}$$

PMNO allocates the physical infrastructure resources to maximize utility function U^{PHY} at the physical layer according to the principle of on-demand allocation (ODA). Then, we can get the RB pattern allocation vector $\boldsymbol{\delta}$, the capacity of each link, and the corresponding $\widetilde{EC}$ constraint of each traffic flow. Finally, VMNOs compete for these data rates of all links, which are from PMNO, adjust the traffic flow access strategy $\mathbf{Y}$ to maximize their utility function U_s^{MAC}, update the $\widetilde{EB}$ requirements, and transfer the $\widetilde{EB}$ requirements to the physical layer again. The above processes are repeated until VMNOs and PMNO achieve the balance of their own optimal strategy through the dynamic interaction. In this way, we can study an infinite number of scheduling periods, and get the EB requirement and EC constraint by taking the average of the $\widetilde{EB}$ and $\widetilde{EC}$ in all scheduling periods.

B. Follower Side: The Resource Allocation Problem Model at The Physical Layer

In this section, we assume that the pattern w contains v continuous RBs, which is denoted as $\{RB_{w_n}|n=1,2,\cdots,v\}$. Then, when the capacity that the w-th RB pattern is allocated to user-k in the g-th group in one scheduling period, the actual channel capacity can be denoted as $R_{w,g,k}=R_{w_1,g,k}+\cdots+R_{w_v,g,k}$, where $R_{w_v,\mathrm{g},k}$ denotes the capacity of the w_v-th RB that is occupied by user-k in the g-th group.

However, obviously, it is very difficult to obtain and allocate $R_{w,g,k}$ in one scheduling period. This is because $R_{w,g,k}$ has a certain probability distribution as Eq. (18) rather than a certain value. Thus, in the following, we virtualize $R_{w,g,k}$ to $\mathrm{E}\left[R_{w,g,k}\right]$, which is considered as the virtual channel resource. Further, according to the user grouping rule as Eq.(6), we can know that the g-th group includes user set $\{u_1,u_2,\cdots,u_m\}$. Therefore, the virtual resource metric matrix in the g-th group can be expressed in (25), as shown at the top of this page.

Therefore, the total virtual resource metric matrix can be written as follows:

$$\mathbf{E}[\mathbf{R}]_{W\times|\mathbf{\Omega}|\times K}=\begin{bmatrix} \mathbf{E}\left[\mathbf{R}\right]_{W\times K}^{1} & \mathbf{E}\left[\mathbf{R}\right]_{W\times K}^{2} & \cdots \\ & \mathbf{E}\left[\mathbf{R}\right]_{W\times K}^{g} & \cdots \quad \mathbf{E}\left[\mathbf{R}\right]_{W\times K}^{|\mathbf{\Omega}|} \end{bmatrix}. \tag{26}$$

Define $W\left|\mathbf{\Omega}\right|K\times 1$ resource allocation vector $\boldsymbol{\delta}$ as

$$\boldsymbol{\delta}=\begin{bmatrix}\delta_{1,1,1},\cdots,\delta_{1,1,K},\cdots,\delta_{1,|\mathbf{\Omega}|,1},\cdots,\delta_{1,|\mathbf{\Omega}|,K},\\ \cdots,\delta_{W,|\mathbf{\Omega}|,1},\cdots,\delta_{W,|\mathbf{\Omega}|,K}\end{bmatrix}, \tag{27}$$

where $\delta_{w,g,k}$ is the assignment index which indicates whether user-k in the g-th user group occupies the w-th RB pattern or not.

To provide service for more traffic flows, the resource allocation at the physical layer should meet the users' traffic requirements as far as possible. In addition, considering the proportional fairness among different users, we can design the utility function with ODA strategy at the physical layer as follows:

$$\text{ODA}: U_{\mathrm{ODA}}^{\mathrm{PHY}}\left(\boldsymbol{\delta}\right)=\sum_{k=1}^{K}\log\left(\frac{\sum_{w=1}^{W}\sum_{g=1}^{|\mathbf{\Omega}|}\mathrm{E}\left[R_{w,g,k}\right]\delta_{w,g,k}}{\widetilde{EB}_k}\right), \tag{28}$$

where $\widetilde{EB}_k$ is $\widetilde{EB}$ requirement of user-k, which can be obtained from the MAC layer, and $\delta_{w,g,k}$ is resource allocation variable. Then, the resource allocation optimization problem at the physical layer can be designed as follows:

$$\arg\max_{\boldsymbol{\delta}} \ U_{\mathrm{ODA}}^{\mathrm{PHY}}\left(\boldsymbol{\delta}\right) \tag{29}$$

subject to

$$\begin{cases} \text{AC1:} \quad \mathbf{A_1}\boldsymbol{\delta}=1_{N_{RB}\times 1}; & \text{(29a)} \\ \text{AC2:} \quad \sum_{w=1}^{W}\sum_{k=1}^{K}\delta_{w,g,k}\geq 1 \quad \forall g=1,\cdots,|\mathbf{\Omega}|; & \text{(29b)} \\ \text{AC3:} \quad \delta_{w,g,k}\in\{0\ ,\ 1\} \quad \forall w=1,\cdots,W; \\ \qquad g=1,\cdots,|\mathbf{\Omega}|;k=1,\cdots,K. & \text{(29c)} \end{cases}$$

where $\mathbf{A_1}=\mathbf{T}_{N_{RB}\times W}\otimes\mathbf{1}_{1\times K|\mathbf{\Omega}|}$. The constraint AC1 is to ensure that each RB can only be allocated to one user group, AC2 is to ensure each user group can obtain one RB pattern, AC3 indicates whether the w-th RB allocation pattern is assigned to user-k in the g-th group (value 1) or not (value 0).

By solving problem (29), we can get the optimal RB pattern allocation result $\delta_{w,g,k}{}^{*}$, which means that the virtual resource, that is the average channel capacity, has been allocated optimally. Then, we instantiate the virtual resource to $\widetilde{EC}$ and derive the $\widetilde{EC}$ corresponding to traffic flow $f_{s,k}$ as follows:

$$\widetilde{EC}_{f_{s,k}}=-\frac{1}{\theta_{f_{s,k}}T_s}\log\mathrm{E}\left[e^{-\theta_{f_{s,k}}X_{f_{s,k}}R_kT_s}\right], \tag{30}$$

where $f_{s,k}$ is the requested traffic flows of the k-th user in the s-th VMN, $X_{f_{s,k}}$ is link bandwidth splitting variable of $f_{s,k}$, $\theta_{f_{s,k}}$ is the QoS exponent of $f_{s,k}$, R_k is the capacity on the k-th link, and $R_k=\sum_{w=1}^{W}\sum_{g=1}^{|\Omega|}R_{w,g,k}\delta_{w,g,k}{}^{*}$.

In order to verify the strong correlation between the ODA strategy of the physical layer and the objective of the MAC layer, two other strategies of resource allocation at the physical layer have been designed and implemented. One is to obtain the maximum capacity of the system in each scheduling slot, denoted as MCA. The other is to assign the resources to the physical links equally and sequentially, denoted as ACA. The corresponding utility functions can be designed as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2886337, IEEE Transactions on Communications

$$\text{MCA}: \quad U_{\text{MCA}}^{\text{PHY}}(\boldsymbol{\delta}) = \sum_{k=1}^{K}\sum_{w=1}^{W}\sum_{g=1}^{|\boldsymbol{\Omega}|} \text{E}\left[R_{w,g,k}\right]\delta_{w,g,k}, \quad (31)$$

$$\text{ACA}: \quad U_{\text{ACA}}^{\text{PHY}}(\boldsymbol{\delta}) = \sum_{i=1}^{K/2}\sum_{k\in\Omega_{\bar{g}_i}} \text{E}\left[R_{\bar{w}_i,\bar{g}_i,k}\right], \quad (32)$$

where the user group set corresponding to the index $\left\{\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_{K/2}\right\}$ is $\{(u_1,u_2),(u_3,u_4),\cdots,(u_K,u_{K+1})\}$ and the RB pattern set corresponding to the index $\left\{\bar{w}_1, \bar{w}_2, \cdots, \bar{w}_{K/2}\right\}$ is $\left\{\left(RB_1,\cdots,RB_{2N_{RB}/K}\right),\left(RB_{2N_{RB}/K+1},\cdots,RB_{4N_{RB}/K}\right),\cdots,\left(RB_{N_{RB}-2N_{RB}/K+1},\cdots,RB_{N_{RB}}\right)\right\}$

C. Leader Side: The Traffic Flow Access Problem Model at The MAC Layer

In this section, the traffic flow set $F=\{f_{s,k},\ s=1,\cdots,S,\ k=1,\cdots,K\}$ is given firstly. Then, for ease of analysis, we assume that the arrival of traffic flow follows Poisson process. The QoS character of the traffic flow $f_{s,k}$ can be described by $\left\{\lambda_{f_{s,k}}, D_s^{\max}, P_{delay}^{out}\right\}$, where $\lambda_{f_{s,k}}$ is average arrival rate of $f_{s,k}$, $D_s^{\max}$ denotes the maximum delay-bounded constraint of traffic flows in the s-th VMN, P_{delay}^{out} is the delay-outage probability of $f_{s,k}$, which shows the probability that the delay exceeds a maximum delay bound. Then, based on the concept of P_{delay}^{out} in [17], we can get

$$P_{delay}^{out}=\text{P}\{Delay\geq D_s^{\max}\}\approx\frac{\lambda_{f_{s,k}}Y_{f_{s,k}}}{\text{E}[R_k]X_{f_{s,k}}}e^{-\theta_{f_{s,k}}\lambda_{f_{s,k}}Y_{f_{s,k}}Ds^{\max}}, \quad (33)$$

where $Y_{f_{s,k}}$ is the access variable of the traffic flow $f_{s,k}$, $\text{E}[R_k]$ is the average capacity of user-k, $\theta_{f_{s,k}}$ can be used for the calculation of EB and EC, and the specific expression can be written as follows:

$$\theta_{f_{s,k}} \approx -\frac{1}{\lambda_{f_{s,k}}Y_{f_{s,k}}D_s^{\max}}\ln\left[\frac{P_{delay}^{out}\text{E}[R_k]X_{f_{s,k}}}{\lambda_{f_{s,k}}Y_{f_{s,k}}}\right]. \quad (34)$$

Further, for the given virtual link rate $\text{E}[R_k]$, which has been determined by $\boldsymbol{\delta}^*$ at the physical layer, different VMNs compete with each other but have a consistent goal to maximum the overall utility. Besides, considering the relevant contents in economics [33], we can design the overall utility function at the MAC layer as

$$\begin{aligned} U^{\text{MAC}} &= \sum_{s=1}^{S} U_s^{\text{MAC}}(\mathbf{Y}_s,\mathbf{Y}_{-s}) \\ &= \sum_{s=1}^{S}\left[\left(\alpha-\beta\sum_{k=1}^{K}\widetilde{EB}_{f_{s,k}}\right)\sum_{k=1}^{K}Y_{f_{s,k}}\lambda_{f_{s,k}}\right], \end{aligned} \quad (35)$$

where $U_s^{\text{MAC}}(\mathbf{Y}_s,\mathbf{Y}_{-s})$ is the utility function of the s-th VMN, $\mathbf{Y}_{-s}$ is the set of traffic flow access vectors of all VMNs, except the s-th one, α is the income of VMN, β is the positive adjustment factor, $\widetilde{EB}_{f_{s,k}}$ is $\widetilde{EB}$ requirement of $f_{s,k}$, $\beta\sum_{k=1}^{K}\widetilde{EB}_{f_{s,k}}$ denotes the cost of the s-th VMN, $\alpha-\beta\sum_{k=1}^{K}\widetilde{EB}_{f_{s,k}}$ denotes the unit profit in the s-th VMN [33]. Considering that a higher EB requirement delivers a better service, we set $\alpha=\rho f\left(\theta_{f_{s,k}}\right)$, where ρ is the positive income coefficient and $f\left(\theta_{f_{s,k}}\right)$ is the function positively related to $\theta_{f_{s,k}}$.

Therefore, the traffic flow access and link bandwidth splitting optimization problem at the MAC layer can be described as follows:

$$\arg\max_{\mathbf{Y}} \ U^{\text{MAC}} \quad (36)$$

subject to

$$\begin{cases} \text{BC1: } Y_{f_{s,k}}\lambda_{f_{s,k}}\leq\widetilde{EC}_{f_{s,k}} & \forall k=1,\cdots,K; s=1,\cdots,S; \ (36a)\\ \text{BC2: } \sum_{s=1}^{S}X_{f_{s,k}}\leq 1 & \forall k=1,\cdots,K; \ (36b)\\ \text{BC3: } 0\leq X_{f_{s,k}}\leq 1 & \forall k=1,\cdots,K; s=1,\cdots,S; \ (36c)\\ \text{BC4: } 0\leq Y_{f_{s,k}}\leq 1 & \forall k=1,\cdots,K; s=1,\cdots,S. \ (36d) \end{cases}$$

The objective of the problem (36) is to maximize the utility function in the s-th VMN. BC1 guarantees the traffic flows that are admitted into VMNs are no more than the corresponding $\widetilde{EC}$ constraints, BC2 ensures the resources of all VMNs rented from each link must be no more than the total resource on the link, BC3 and BC4 are to ensure both $X_{f_{s,k}}$ and $Y_{f_{s,k}}$ value from 0 to 1.

Further, by solving problem (36), we can get the traffic flow access variable ${Y_{f_{s,k}}}^*$. Then, the $\widetilde{EB}$ requirement of user-k at the physical layer can be written as follows:

$$\widetilde{EB}_k=\sum_{s=1}^{S}\widetilde{EB}_{f_{s,k}}, \quad (37)$$

where $\widetilde{EB}_{f_{s,k}}$ is the $\widetilde{EB}$ requirement of traffic flow $f_{s,k}$, and it has been assumed that the arrival of $f_{s,k}$ follows Poisson process. Thus, the specific expression of $\widetilde{EB}_{f_{s,k}}$ can be computed as follows:

$$\widetilde{EB}_{f_{s,k}}=\frac{{Y_{f_{s,k}}}^*\lambda_{f_{s,k}}\left(e^{\theta_{f_{s,k}}}-1\right)}{\theta_{f_{s,k}}}. \quad (38)$$

V. DYNAMIC ALGORITHM FOR RESOURCE VIRTUALIZATION PROBLEM

According to the different aims of PMNO and VMNOs, the wireless resource virtualization problem is decoupled into two sub-games, whose utility functions are both convex. Then, it can converge to the NE through multiple iterations. Further, a dynamic algorithm is developed, which includes the solution algorithms to the two sub-games and the message exchange between these dual processes.

A. Proposed Algorithm

To access and serve more traffic flows, we initialize $Y_{f_{s,k}}$=1 and $\text{E}[R_k]=\frac{1}{N_{RB}}\sum_{n=1}^{N_{RB}}\sum_{g=1}^{|\boldsymbol{\Omega}|}\text{E}[R_{n,g,k}]$. Then, according to the interaction process between PMNO and VMNOs described in section IV, the overall dynamic interaction algorithm and the message exchange are presented in Algorithm 1, which is described as follows:

Algorithm 1 Dynamic Algorithm for Resource Virtualization

Step 1:

Initialize $\left\{\lambda_{f_{s,k}}, D_s^{\max}, P_{delay}^{out}\right\}$, $\mathbf{Y}^{(1)}=\mathbf{1}$, $X_{f_{s,k}}{}^{(1)} = \frac{\lambda_{f_{s,k}}}{\sum_{s=1}^{S}\lambda_{f_{s,k}}}$, $\mathrm{E}[R_k]^{(1)}=\frac{1}{N_{RB}}\sum_{n=1}^{N_{RB}}\sum_{g=1}^{|\mathbf{\Omega}|}\mathrm{E}\left[R_{n,g,k}\right]$ and $i=0$;

Step 2:

Repeat

$i \Leftarrow i+1$;

Compute $\theta_{f_{s,k}}^{(i)}$ by Eq. (34);

Compute $\widetilde{EB}_k{}^{(i)}$ by Eq. (37);

Run Algorithm 2 for problem (29) to obtain $\boldsymbol{\delta}^{(i)}$ and $\mathrm{E}[R_k]^{(i+1)}$;

Update $\theta_{f_{s,k}}^{(i+1)}$, $\widetilde{EB}_{f_{s,k}}{}^{(i+1)}$ and $\widetilde{EB}_k{}^{(i+1)}$ by Eq. (34), Eq. (38) and Eq. (37) respectively;

Run Algorithm 3 for problem (36) to obtain $\mathbf{Y}^{(i+1)}$;

Until $\left|\mathbf{Y}^{(i+1)}-\mathbf{Y}^{(i)}\right| < \varepsilon$

Step 3:

Obtain the optimal $\mathbf{Y}^* = \mathbf{Y}^{(i+1)}$.

B. Solutions to the Sub-games

1) Resource Allocation Performed by PMNO

Obviously, the optimization problem (29) is a typical binary integer programming problem, and it is suitable to be solved by BNB algorithm [34]. However, BNB algorithm is too complex for a practical implementation especially when the number of users and RBs becomes large. Thus, inspired by the fast unfolding algorithm (FUA) [35] and the iterative Hungarian algorithm (IHA) [26], we propose an efficient algorithm including two parts. Specifically, we describe the detailed process as follows:

Part I: adopting the FUA to decompose the bipartite graph consisting of users and RBs into multiple sub-graphs based on the principle of maximizing network modularity

First, we introduce the concept of complete set, which consists of many complete subsets. In one complete subset, the intersection of all the elements must be empty, and the union of all the elements is the complete subset itself.

Second, because adjacent RBs must be assigned to one user group, we construct the complete RB pattern set by putting 0 or 1 on the underline of $\{RB_1, \_, RB_2, \_, \cdots, \_, RB_{N_{RB}}\}$, where putting 0 on the underline between RB_i and RB_{i+1} means that RB_i and RB_{i+1} are assigned to the same RB pattern and vice versa. Then, it can be denoted as $\mathbf{T}^{RB} = \left\{\mathbf{T}_1^{RB}, \cdots, \mathbf{T}_t^{RB} \cdots, \mathbf{T}_{|\mathbf{T}^{RB}|}^{RB}\right\}$, where $\left|\mathbf{T}^{RB}\right|$ is the number of all subsets, and $\left|\mathbf{T}^{RB}\right| = 2^{N_{RB}-1}$.

Third, for all complete RB pattern subsets, we perform the same operation in the following. Specifically, we consider all possible matches between all users and RB patterns in the each subset as a bipartite graph. Then, after running the FUA, the bipartite graph is divided into multiple sub-graphs. Finally, we find out the subset that makes the modularity the most, the schematic is shown in Figure 6.

Part II: adopting the IHA to achieve the best match between user groups and RB patterns in each sub-graph

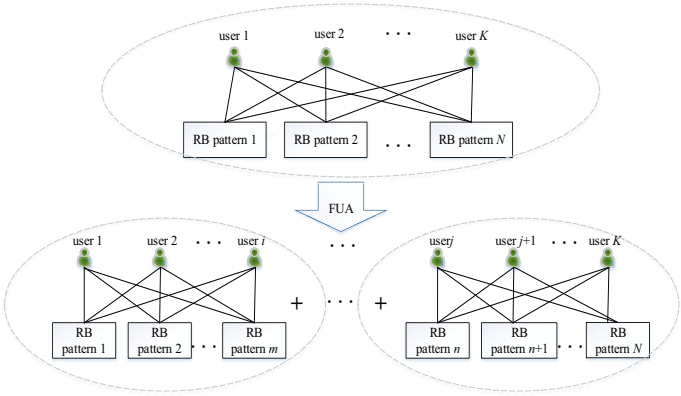


Fig. 6: The schematic of the FUA.

Assuming that there are V sub-graphs, and one of sub-graphs contains M RB patterns and N users, similar to Eq.(6), we can generate all possible user groups as $\hat{\mathbf{\Omega}} = \left\{\hat{\mathbf{\Omega}}^{(1)}, \cdots, \hat{\mathbf{\Omega}}^{(n)}, \cdots, \hat{\mathbf{\Omega}}^{(N)}\right\}$, where $\hat{\mathbf{\Omega}}^{(n)}$ denotes the user group set including n users, $\left|\hat{\mathbf{\Omega}}^{(n)}\right| = C_N^n$, and $\left|\hat{\mathbf{\Omega}}\right| = \sum_{i=1}^{N} C_N^i$. In addition, it is obvious that the index m of M RB patterns can correspond to the index w in RB pattern matrix $\mathbf{T}$. Thus, according to Eq.(26), we can compute the metric matrix $\mathbf{TP}_v$ as

$$\mathbf{TP}_v = \begin{array}{c} \text{group index } \begin{matrix} 1 & 2 & \cdots & \left|\hat{\mathbf{\Omega}}\right| \end{matrix} \text{ RB pattern index} \\ \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,|\hat{\mathbf{\Omega}}|} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,|\hat{\mathbf{\Omega}}|} \\ \vdots & \vdots & \ddots & \vdots \\ U_{M,1} & U_{M,2} & \cdots & U_{M,|\hat{\mathbf{\Omega}}|} \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ M \end{matrix} \end{array}, \tag{39}$$

where $U_{m,g} = \sum_{k\in\Omega_g} \log\left(\frac{\mathrm{E}[R_{m,g,k}]}{\widetilde{EB}_k}\right)$, $\mathrm{E}\left[R_{m,g,k}\right]$ can correspond to $\mathrm{E}\left[R_{w,g,k}\right]$ in Eq.(26).

The specific steps can be described as follows:

Algorithm 2 Resource reservation algorithm based on fast unfolding and iterative Hungarian method

Step 1:

Initialize $t=0$ and $v=0$;

Step 2:

Generate the complete RB pattern set by putting 0 or 1 on the underline of $\{RB_1, \_, RB_2, \_, \cdots, \_, RB_{N_{RB}}\}$ as mentioned above;

Repeat

$t \Leftarrow t+1$;

Consider all possible matches between all users and RB patterns in the t-th subset as a bipartite graph;

Run the FUA to decompose the bipartite graph into multiple sub-graphs;

Compute the corresponding modularity;

Until $t = 2^{N_{RB}-1}$

Find out the subset that makes the modularity the most;

Step 3:
Repeat
$v \Leftarrow v+1$;
Repeat
Compute the metric matrix $\mathbf{TP}_v$ by Eq. (39);
Select the maximum value $U^* = U_{m^*,g^*}$ from matrix $\mathbf{TP}_v$, and $(m^*, g^*) = \arg\max_{m,g} \{ U_{m,g} | \ m \in [1, M], \ g \in [1, \left|\hat{\mathbf{\Omega}}\right|]\}$;
Delete the m^*-th row and the g^*-th column in matrix $\mathbf{TP}_v$;
Record the row labels of elements whose values are 1 in the g^*-th column of matrix B;
Traverse all the elements of the row, record the column labels of elements whose values are 1;
Delete the corresponding columns in matrix $\mathbf{TP}_v$;
Until $\mathbf{TP}_v = \phi$
Until $v = V$
Record the corresponding $\boldsymbol{\delta}$.

2) Traffic Flow Access Control Performed by VMNOs

The optimization problem (36) is a nonlinear programming problem with high complexity, which involves the traffic flow access variable $Y_{f_{s,k}}$ and link bandwidth splitting variable $X_{f_{s,k}}$. To make it tractable, we propose a dynamic iteration method including two steps. The first step is to compute $X_{f_{s,k}}$ based on $Y_{f_{s,k}}$ obtained by the last iteration, and the second step is to get the optimal $Y_{f_{s,k}}$ based on the results of step1. The specific update process is designed as follows:

First, we introduce the Lagrangian factor $\omega_{f_{s,k}} \geq 0$ to the optimization problem (36). Then, the problem (36) is reformulated in Lagrangian form as:

$$L^{\mathrm{MAC}} = \sum_{s=1}^{S} \left[\left(\alpha - \beta \sum_{k=1}^{K} \widetilde{EB}_{f_{s,k}} \right) \sum_{k=1}^{K} Y_{f_{s,k}} \lambda_{f_{s,k}} \right] + \sum_{k=1}^{K} \omega_{f_{s,k}} \left(\widetilde{EC}_{f_{s,k}} - Y_{f_{s,k}} \lambda_{f_{s,k}} \right) . \quad (40)$$

It is noticed that the following conditions about $\omega_{f_{s,k}}$ must be satisfied:

$$0 \leq \omega_{f_{s,k}} \perp \left(\widetilde{EC}_{f_{s,k}} - Y_{f_{s,k}} \lambda_{f_{s,k}} \right) \geq 0, \quad (41)$$

where the notation $0 \leq a \perp b \geq 0$ means $a \cdot b = 0$, $a \geq 0$ and $b \geq 0$. The iterative formula of $\omega_{f_{s,k}}$ is described as follows:

$$\omega_{f_{s,k}}(t+1) = \left[\omega_{f_{s,k}}(t) + \nabla\omega_{f_{s,k}} \left(Y_{f_{s,k}} \lambda_{f_{s,k}} - \widetilde{EC}_{f_{s,k}} \right) \right]^{+}, \quad (42)$$

where $(\bullet)^+ = \max(\bullet, 0)$, t is the iteration index, $\nabla\omega_{f_{s,k}}$ is the positive iteration step.

Second, for given $Y_{f_{s,k}}$, we analyze the characteristics of Eq. (40). Take the derivative of Eq. (40) with respect to $X_{f_{s,k}}$ as follows:

$$\frac{\partial L^{\mathrm{MAC}}}{\partial X_{f_{s,k}}} = \omega_{f_{s,k}} \frac{\mathrm{E}\left[R_k e^{-X_{f_{s,k}} \theta_{f_{s,k}} T_s R_k} \right]}{\mathrm{E}\left[e^{-X_{f_{s,k}} \theta_{f_{s,k}} T_s R_k} \right]}. \quad (43)$$

Obviously, $\frac{\partial L^{\mathrm{MAC}}}{\partial X_{f_{s,k}}} > 0$, which denotes Eq.(40) is monotonically increasing with respect to $X_{f_{s,k}}$. Then, $X_{f_{s,k}}$ can be updated by the gradient of the utility function, and the specific iterative equation can be written as:

$$X_{f_{s,k}}(j+1) = X_{f_{s,k}}(j) + d_{f_{s,k}} \left. \frac{\partial L^{\mathrm{MAC}}}{\partial X_{f_{s,k}}} \right|_{X_{f_{s,k}} = X_{f_{s,k}}(j)}, \quad (44)$$

where j is the iterative index, and $d_{f_{s,k}}$ indicates the strategy adjustment step size, which is used to control the speed of strategy adjustment.

Third, derive the first derivative of L^{MAC} with respect to $Y_{f_{s,k}}$ as follows:

$$\frac{\partial L^{\mathrm{MAC}}}{\partial Y_{f_{s,k}}} = \left(\alpha - \beta \sum_{k=1}^{K} \frac{\mathrm{Y}_{f_{s,k}} \lambda_{f_{s,k}} \left(e^{\theta_{f_{s,k}}} - 1 \right)}{\theta_{f_{s,k}}} \right) \lambda_{f_{s,k}} - \beta \frac{\lambda_{f_{s,k}} \left(e^{\theta_{f_{s,k}}} - 1 \right)}{\theta_{f_{s,k}}} \sum_{k=1}^{K} Y_{f_{s,k}} \lambda_{f_{s,k}} - \omega_{f_{s,k}} \lambda_{f_{s,k}} . \quad (45)$$

Then, derive the second derivative of L^{MAC} with respect to $Y_{f_{s,k}}$ as follows:

$$\frac{\partial^2 L^{\mathrm{MAC}}}{\partial {Y_{f_{s,k}}}^2} = -2\beta \frac{\lambda_{f_{s,k}}^2 \left(e^{\theta_{f_{s,k}}} - 1 \right)}{\theta_{f_{s,k}}}. \quad (46)$$

It is obvious that $\frac{\partial^2 L^{\mathrm{MAC}}}{\partial {Y_{f_{s,k}}}^2} < 0$. Thus, L^{MAC} is a convex function with respect to $Y_{f_{s,k}}$. To obtain the optimal utility function value and traffic flow access rate, we use the gradient iteration method to compute the extreme point of Eq. (40). The specific iterative equation is as follows:

$$Y_{f_{s,k}}(\tau+1) = Y_{f_{s,k}}(\tau) + \upsilon_{f_{s,k}} \left. \frac{\partial L^{\mathrm{MAC}}}{\partial Y_{f_{s,k}}} \right|_{Y_{f_{s,k}} = Y_{f_{s,k}}(\tau)}, \quad (47)$$

where τ is the iterative index, and $\upsilon_{f_{s,k}}$ indicates the strategy adjustment step size.

The specific algorithm is described as follows:

Algorithm 3 Gradient Iteration Algorithm for Traffic Flow Access Control

Step 1:
Initialize $w_{f_{s,k}}$, $d_{f_{s,k}}$, $\kappa = 0$, $Y_{f_{s,k}}(1) = 1$, $\hat{X}_{f_{s,k}}(0) = 0$ and $\hat{Y}_{f_{s,k}}(0) = 0$;
Step 2:
Repeat
$\kappa \Leftarrow \kappa + 1$;
Step 2a:
Initialize $j = 0$ and $X_{f_{s,k}}(1) = 0$;
Repeat
$j \Leftarrow j + 1$;
Update $X_{f_{s,k}}(j+1)$ by Eq.(44);
Until $\sum_{s=1}^{S} X_{f_{s,k}} \geq 1$
$\hat{X}_{f_{s,k}}(\kappa) \Leftarrow X_{f_{s,k}}(j+1)$;
Step 2b:
If Eq.(41) is satisfied:
Go to Step 2c;
Else:

Update $\omega_{f_{s,k}}$ by Eq.(42);
Go to Step 2a;
Step 2c:
Update $\theta_{f_{s,k}}$, $\widetilde{EC}_{f_{s,k}}$ and $\widetilde{EB}_{f_{s,k}}$ by Eq.(34), Eq.(30) and Eq.(38) respectively;
Step 2d:
Initialize $\tau = 0$;
Repeat
$\tau \Leftarrow \tau + 1$;
Update $Y_{f_{s,k}}(\tau{+}1)$ by Eq.(47);
Until $\left|Y_{f_{s,k}}(\tau{+}1) - Y_{f_{s,k}}(\tau)\right| < \varepsilon$;
$\hat{Y}_{f_{s,k}}(\kappa) \Leftarrow Y_{f_{s,k}}(\tau+1)$
Update $\theta_{f_{s,k}}$, $\widetilde{EC}_{f_{s,k}}$ and $\widetilde{EB}_{f_{s,k}}$ by Eq.(34), Eq.(30) and Eq.(38) respectively;
Until $\left|\hat{X}_{f_{s,k}}(\kappa) - \hat{X}_{f_{s,k}}(\kappa-1)\right| < \varepsilon$ and $|\hat{Y}_{f_{s,k}}(\kappa) - \hat{Y}_{f_{s,k}}(\kappa-1)| < \varepsilon$
Step 3:
Obtain $Y_{f_{s,k}} = \min\{\hat{Y}_{f_{s,k}}(\kappa), 1\}$.

C. Stackelberg Equilibrium

In the above description, the proposed problem has been formulated as the Stackelberg game model. Thus, it is very necessary to discuss the existence of Stackelberg equilibrium (SE).

As mentioned above, Algorithm 1 is used to solve the optimization problem between VMNOs and PMNO, which is modeled as the Stackelberg game known as "leader-follower model". First, in Step 1 of Algorithm 1, as the leaders, VMNOs make traffic flow access strategy $\mathbf{Y}$, and further, $EB(\mathbf{Y})$ is computed by Eq. (34) and Eq. (38). Second, in Algorithm 2, as the follower, PMNO makes resource allocation strategy $\boldsymbol{\delta}$ according to the EB requirement on each link and the average channel capacity $E[R]$ after Gaussian-like fitting. Then, we obtain the resources R allocated to each link and the EC constraint by Eq. (30). Third, in Algorithm 3, the current $\mathbf{Y}$ can be got through the link bandwidth splitting strategy $\mathbf{X}$, which is obtained by the gradient iteration. Finally, as described in Algorithm 1, if $\left|\mathbf{Y}^{(i+1)} - \mathbf{Y}^{(i)}\right| < \varepsilon$, the optimal solution of Algorithm 1 $\mathbf{Y}^* = \mathbf{Y}^{(i+1)}$. Otherwise, we continue to run Algorithm 1 with $\mathbf{Y}^{(i+1)}$. The specific relationship among all variables is shown as follows:

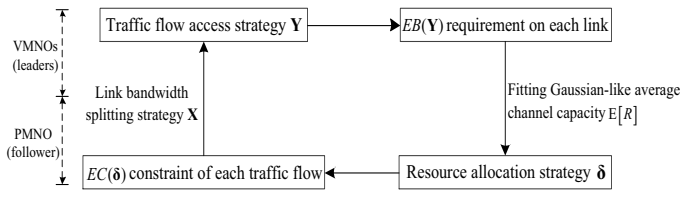


Fig. 7: The specific relationship among all variables.

Let $\boldsymbol{\delta}^*$ be a solution for problem (29) and $\mathbf{Y}^*$ be a solution for problem (36). Then, for all feasible $\boldsymbol{\delta}$ and $\mathbf{Y}$, the SE point $(\boldsymbol{\delta}^*, \mathbf{Y}^*)$ should satisfy the following conditions.

$$U_{\mathrm{ODA}}^{\mathrm{PHY}}(\boldsymbol{\delta}^*, EB(\mathbf{Y}^*)) \geq U_{\mathrm{ODA}}^{\mathrm{PHY}}(\boldsymbol{\delta}, EB(\mathbf{Y}^*)), \quad (48)$$

$$U^{\mathrm{MAC}}(\mathbf{Y}^*, EC(\boldsymbol{\delta}^*)) \geq U^{\mathrm{MAC}}(\mathbf{Y}, EC(\boldsymbol{\delta}^*)). \quad (49)$$

At the PMNO's side, since there is only one player, the resource allocation strategy $\boldsymbol{\delta}$ of the PMNO can be easily obtained by solving the problem (29). Then, at the VMNOs' side, the leaders derive their traffic flow access strategy $\mathbf{Y}$ by solving the problem (36). The specific description is given in the following:

1) The follower (PMNO) side analysis:

At the physical layer, given $\mathbf{Y}$ from the MAC layer, since the utility function is concave, we obtain the unique $\boldsymbol{\delta}$ through Algorithm 2. Further, we obtain the average link rate $\mathrm{E}[\mathrm{R}]$ on each link, which is almost proportional to the EB requirement.

Besides, in the resource virtualization phase, we compute $\mathrm{E}[\mathrm{R}]$ with the Gaussian-like fitting $\hat{R}$. Then, through Eq. (23) and Eq. (24), we know that after Gaussian-like fitting, $\widetilde{EC}$ is close to $\widetilde{EC}$ and is a linear function of $\mathrm{E}[\mathrm{R}]$. Therefore, the actual $\widetilde{EC}$ is nearly a linear function of $\mathrm{E}[\mathrm{R}]$.

2) The leaders (VMNOs) side analysis:

As shown in Eq.(40), the optimization problem (36) is reformulated to Lagrangian form, and the second derivative of L^{MAC} with respect to $Y_{f_{s,k}}$ has been given in Eq.(46). Then, we can find that $\frac{\partial^2 L^{\mathrm{MAC}}}{\partial {Y_{f_{s,k}}}^2} < 0$, thus L^{MAC} is a convex function. As a result, we can derive the traffic flow access strategy $\mathbf{Y}$ by Algorithm 2.

D. Complexity analysis of the total algorithm

The complexity of the proposed algorithm mainly comes from the Algorithm 3 of the MAC layer and the Algorithm 2 of the physical layer.

The idea of the linear search method with the time complexity $O(n)$ is used for the Algorithm 3 of the MAC layer. Assuming that Eq.(36b) can be satisfied after n iterations, the main execution time of the program is $SKn + SK$. Thus the time complexity of the Algorithm 3 can be expressed as $O(n)$.

For the Algorithm 2 of the physical layer, the total number of partition operations of N_{RB} RBs is $2^{N_{RB}-1}$, the maximum number of RB patterns in all RB partitions is N_{RB}, and the total number of user groups is $\sum_{i=1}^{K} C_K^i$. To reduce the computational complexity significantly, the FUA is used to solve the problem by dividing the overall users and RBs into M communities. Then, assuming that there are K_M users and N_{RP}^M RB patterns in the largest community, we can compute the number of all possible user groups as $\sum_{i=1}^{K_M} C_{K_M}^i$ and use the IHA to achieve the best match between user groups and RB patterns. Further, the time complexity in the largest community can be expressed as $O(K_M N_{RP}^M \sum_{i=1}^{K_M} C_{K_M}^i)$. Therefore, at the worst case, the time complexity of the Algorithm 2 after community processing can be expressed as $O(MK_M N_{RP}^M \sum_{i=1}^{K_M} C_{K_M}^i)$.

For the two-layer iterative process in the Algorithm 1, we can get equilibrium within a finite number of iterations. As

the simulation in Section VI, the result can be obtained after 3 iterations.

VI. PERFORMANCE EVALUATION AND ANALYSIS

In this section, we simulate a queueing system and demonstrate the performance of our cross-layer algorithm for the resource allocation of VMNs with customized services.

A. Simulation Setting

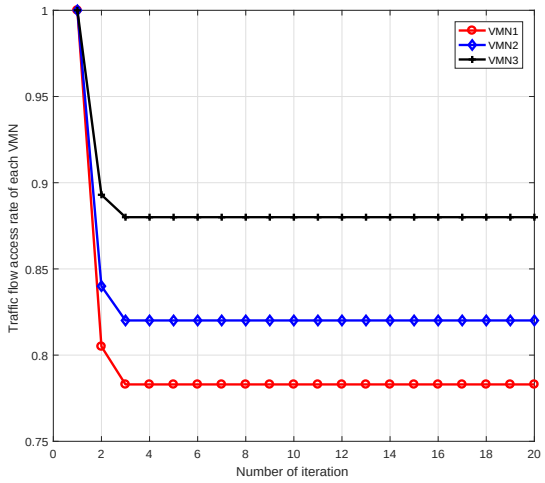
We conduct the simulations based on the LTE uplink and adopt the pedestrian test environment channel A suggested by ITU-R M.1225 [36]. The simulation parameters are listed in Table I. In addition, we use the scenario with perfect power control and define the transmitting SNR as E_s/σ^2 in the following.

TABLE I: SIMULATION PARAMETERS

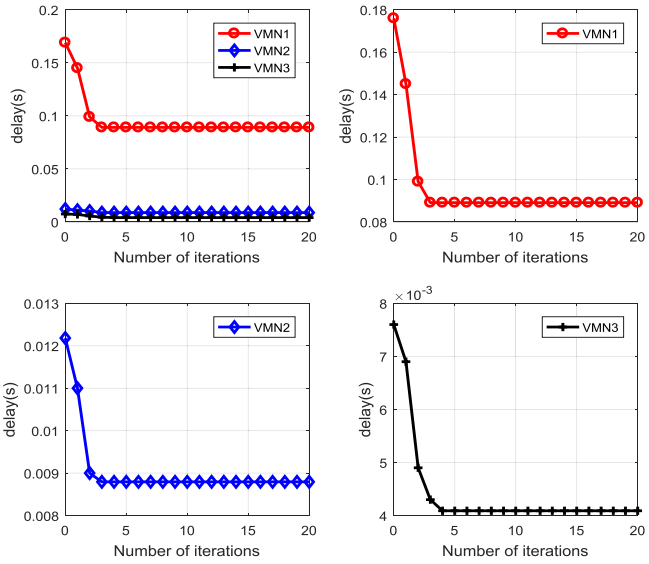
Parameter	Value
Channel model	ITU Ped-A
Carrier frequency	2GHz
Sampling frequency	1.92MHz
Maximum Doppler shift	10Hz
FFT size	128
OFDM symbols per frame	14
Subcarrier spacing	15kHz
System bandwidth	1.4MHz
Number of RBs N_{RB}	6
Number of coordinated cells N_c	4
N_{SC}^{RB}	12
RB configure	12×7
Number of users	6
UE transmitting antenna number	1
MIMO detector	MMSE
BS Receiving antenna number N_r	4
TTI duration	1ms
Scheduling period	10
Simulation frames	1000

B. Simulation results

To evaluate the performance of the proposed algorithm, we consider the simulation setups as shown in Table II, where the traffic flows of users belong to 3 VMNs with different customized delay-bounded QoS provisioning. For the delay performance simulations, we set packet size as 1 Kbits and adopt the M/G/1 queueing model.



(a) Traffic flow access rate of each VMN versus the number of iteration



(b) Maximum delay of each VMN versus the number of iteration

Fig. 8: The convergence of the proposed algorithm (SNR=6dB, $\beta = 0.1\beta_{\max}$).

As described in Section IV-C, we has set $\alpha = \rho f(\theta_{f_{s,k}})$. In the simulation section, considering the income is positively related to the cost, similar to the form of EB, we set

TABLE II: SIMULATIONS SETUP FOR FLOWS

Traffic flow / QoS performance	VMN1			VMN2			VMN3		
	User 1-2	User 3-4	User 5-6	User 1-2	User 3-4	User 5-6	User 1-2	User 3-4	User 5-6
$\lambda_{f_{s,k}}$(Kbps)	120	200	280	130	210	260	140	220	240
$D_s^{\max}$(s)	0.1			0.01			0.005		
P_{delay}^{out}	0.01								

$f\left(\theta_{f_{s,k}}\right) = \sum_{k=1}^{K} \frac{e^{\theta_{f_{s,k}}}-1}{\theta_{f_{s,k}}}$. Further, we normalize the income α by setting the income coefficient $\rho = \frac{1}{\max\limits_{s} \sum\limits_{k=1}^{K} \frac{e^{\theta_{f_{s,k}}}-1}{\theta_{f_{s,k}}}}$ and constrain the adjustment factor $0 < \beta < \frac{1}{\max\limits_{s,\mathbf{Y}_s=1} \sum\limits_{k=1}^{K} \widetilde{EB}_{f_{s,k}}}$ so that the range of cost is from 0 to 1. Because $\theta_{f_{s,k}}$ has a small change during the iterative processes, we compute the maximum value of β as $\beta_{\max} = [\beta_1, \cdots, \beta_n, \cdots, \beta_N]$ by the above constraint, where β_n is the maximum value of β in the n-th iteration and N is the total number of iterations.

1) Convergence of the proposed algorithm

In this section, to show the convergence of the proposed algorithm for the access rate of traffic flows and the maximum delay of each VMN, we set the arrival traffic flows of all VMNs as that of VMN2 in Table II. As shown in section IV, an interactive process between the PMNO and the VMNOs is formulated as a Stackelberg game model, which is simulated to demonstrate the evolution of access rate and delay performance.

From Figure 8(a) and 8(b), we can see that the initial access rates of all traffic flows are 1. Then the average rates of traffic flows that are admitted into VMNs exceed the corresponding EC constraints, which leads to the delay-bounded constraint violation. Then the access rates and the corresponding EC are adjusted until all traffic flows meet the delay-bounded constraints. From the figures,we can observe that the algorithm performance goes stabilized after 3 iterations.

2) Resource Efficiency of the Follower

In this section, the traffic flows of each VMN arrive as shown in Table II. Specifically, the follower sub-game in the proposed algorithm is based on the ODA strategy at the physical layer. For comparison purpose, the MCA strategy and the ACA strategy have been designed and implemented.

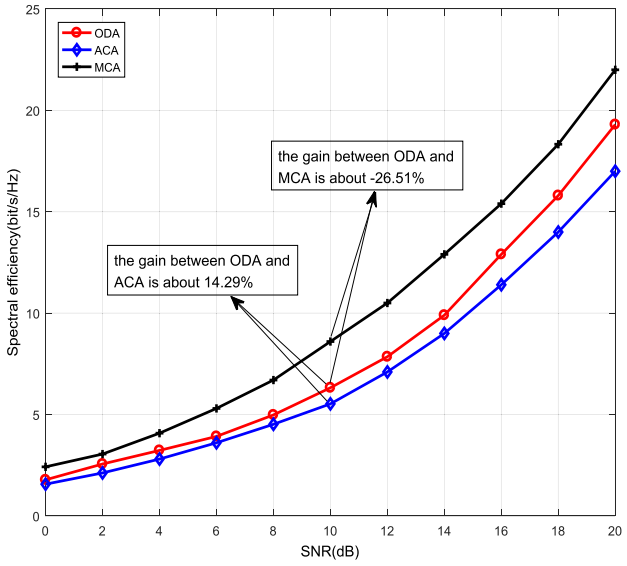


Fig. 9: The spectral efficiency versus SNR with different strategies of the follower ($\beta = 0.1\beta_{\max}$).

It can be observed from Figure 9 that MCA outperforms ODA and ACA, which is in accordance with the design goal of MCA stratergy. ODA is between MCA and ACA, which is because ODA mainly considers the traffic requiement to optimize the resource allocation. ACA makes no optimization to resource allocation, thus its performance is the worst. For example, when SNR=10dB, the gain between ODA and ACA is about 14.29%, and the gain between ODA and MCA is about -26.51%.

However, as we can see from Figure 10, for the traffic flow access rate, when SNR=10dB, the gain between ODA and ACA is about 6.83%, and the gain between ODA and MCA is about 14.67%. ODA strategy proposed in this paper allows more traffic flows access to the network than ACA and MCA. This is because ODA considers the user requirement in the design of the utility function at the physical layer. Then, it can meet the need of each link to the greatest extent, which is the best match with the users' needs. However, the performance of ACA and MCA mainly depends on the actual situation of traffic flows, and in this simulation, ACA is better than MCA.

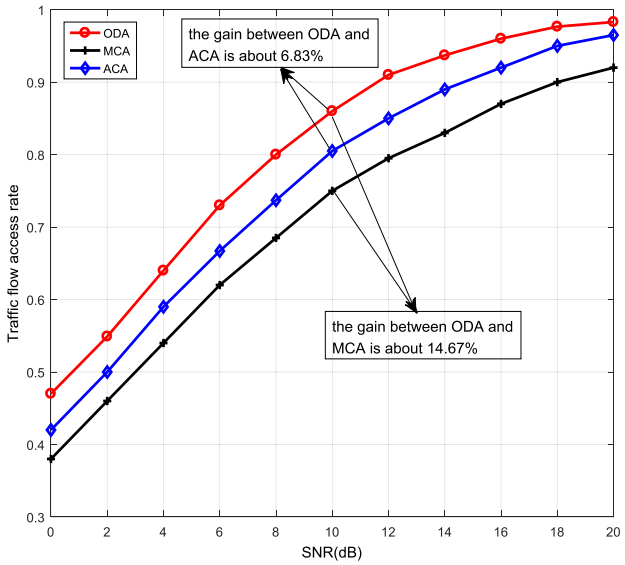


Fig. 10: Traffic flow access rate versus SNR with different strategies of the follower ($\beta = 0.1\beta_{\max}$).

3) Delay and Access Rate Performances of Leaders

Figure 11 plots the curve of the maximum delay of each VMN versus SNR firstly, where the arrival traffic flows of all VMNs are consistent with that of VMN2 in Table II. Then, specific to the delay with SNR=6dB and 12dB, the curves of cumulative distribution function (CDF) of delay in each VMN are plotted. Then from Figure 11(a), we can see that with the increase of SNR, the delay is getting smaller and smaller, and when the SNR is about 6dB, all VMNs meet the delay-bounded constraints basically. Meanwhile, it can be seen from Figure 11(b)-(d) that when SNR=6dB, not all traffic flows are admitted, but the maximum delay of each VMN is approaching to the delay-bounded constraint, thus the CDF of delay tends to be 1 near the delay-bounded constraint. When SNR=12dB, all traffic flows are admitted, and the resource exceeds the traffic request, thus the delay performance is improved obviously.

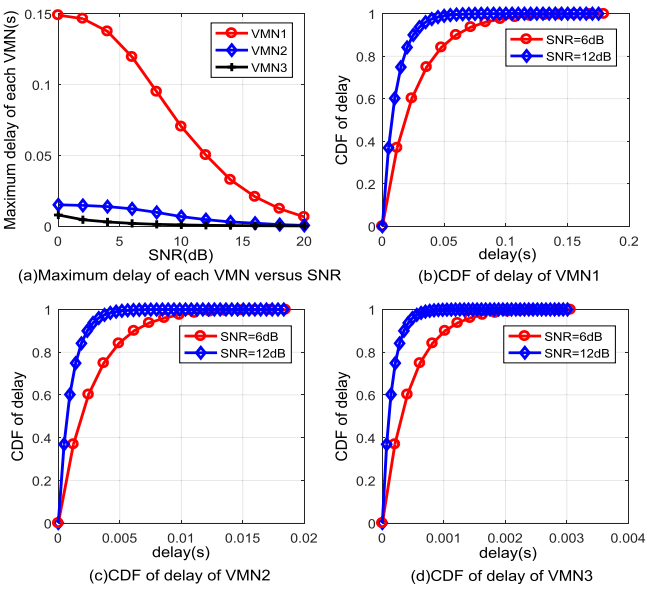


Fig. 11: Delay performance of each VMN ($\beta = 0.1\beta_{\max}$).

In Figure 12, to show the impact of β on the traffic flow access strategy, we simulate the traffic flow access rate versus SNR with $\beta = 0.01\beta_{\max}$, $0.1\beta_{\max}$, $0.5\beta_{\max}$ and $0.8\beta_{\max}$. Then, it is found that as β increases, the traffic flow access rate of VMN3 decreases, the traffic flow access rate of VMN2 is unchanged basically, and the traffic flow access rate of VMN1 increases. This is because the increase of β means that the unit profit brought by QoS provisioning is gradually reduced, and VMN3 can provide the best service, VMN2 is the second, and VMN1 is the worst. For example, when SNR=6dB, the ratios of the traffic flow access rates of different VMNs are 0.912:0.835:0.739, 0.893:0.830:0.771, 0.870:0.834:0.791 and 0.851:0.833:0.805 respectively.

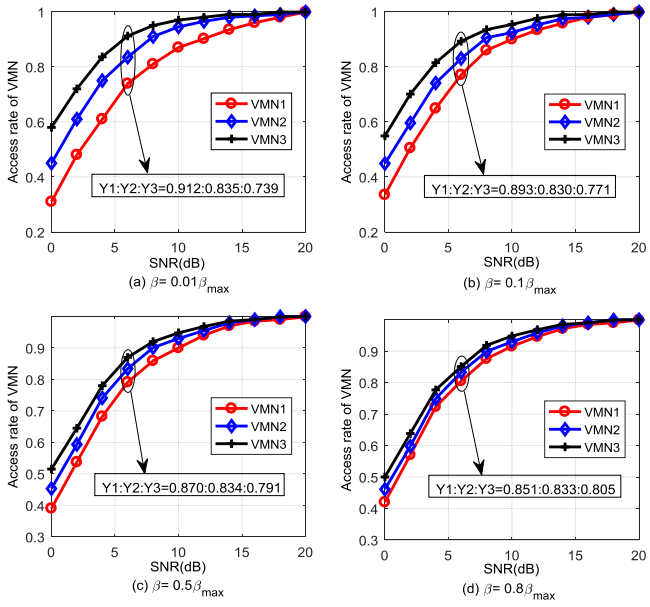


Fig. 12: Traffic flow access rate of each VMN versus SNR with different β values.

From Figure 11 and Figure 12, we can see that different maximum delay-bounded constraints make different VMNs show different service performances, including delay and access rate. Thus, the isolation among VMNs can be further proved.

VII. CONCLUSIONS

In this paper, a wireless resource virtualization scheme in uplink VMIMO-SC-FDMA systems is proposed. Through the analysis of wireless channel, a FSMC model is formed and EC for one slice scheduling period is calculated. Then we formulate the interactions between the VMNOs and the PMNO as a Stackelberg game, which considers benefits of all players. Finally, a dynamic algorithm with dual update is developed to search the Stackelberg equilibrium, which is the solution of the resource virtualization problem. Simulation results demonstrate that the proposed algorithm can improve resource efficiency and traffic flows' access rate under specific delay-bounded constraints. In the future, we plan to consider resource virtualization in cross-domain and investigate the uniform mechanism of resource virtualization and instantiation for end-to-end network slicing with customized delay-bounded QoS provisioning.

REFERENCES

[1] Z. Feng, C. Qiu, Z. Feng, Z. Wei, W. Li, and P. Zhang, "An effective approach to 5G: Wireless network virtualization," *IEEE Communications Magazine*, vol. 53, no. 12, pp. 53-59, Dec. 2015.

[2] A. Belbekkouche, M. M. Hasan, and A. Karmouch, "Resource Discovery and Allocation in Network Virtualization," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 1114-1128, 4th Quart., 2012.

[3] C. Liang, and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 358-380, 1st Quart., 2015.

[4] J. van de Belt, H. Ahmadi and L. E. Doyle, "Defining and Surveying Wireless Link Virtualization and Wireless Network Virtualization," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1603-1627, 3rd Quart., 2017.

[5] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "OpenRAN: A software-defined RAN architecture via virtualization," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 549-550, Oct. 2013.

[6] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel, "LTE wireless virtualization and spectrum management," *Wireless and Mobile Networking Conference* , pp.1-6, 13-15 Oct. 2010.

[7] L. Zhao, M. Li, Y. Zaki, and A. Timm-Giel, "LTE virtualization: From theoretical gain to practical solution," *23rd International Teletraffic Congress* , San Francisco, CA, 2011, pp. 71-78.

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

[9] X. Zhang and Q. Zhu, "Information-centric network virtualization for QoS provisioning over software defined wireless networks,"*MILCOM 2016 - 2016 IEEE Military Communications Conference*, Baltimore, MD, 2016, pp. 1028-1033.

[10] F. Fu and U. C. Kozat, "Stochastic Game for Wireless Network Virtualization,"*IEEE/ACM Transactions on Networking*, vol. 21, no. 1, pp. 84-97, Feb. 2013.

[11] X. Lu, K. Yang, and H. Zhang, "An elastic sub-carrier and power allocation algorithm enabling wireless network virtualization," *Wireless Personal Communications*, vol. 75, no. 4, pp. 1827-1849, Apr. 2013.

[12] X. Lu, K. Yang, Y. Liu, D. Zhou, and S. Liu, "An elastic resource allocation algorithm enabling wireless network virtualization," *Wireless Communications and Mobile Computing*, vol. 15, no.2, pp. 295-308, Feb. 2015.

[13] C. S. Chang, and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," *Proc. IEEE INFOCOM95*, vol. 3, 1995, pp. 1001-1009.

[14] C. S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091-1100, Aug. 1995.

[15] B. L. Mark and G. Ramamurthy, "Real-time estimation and dynamic renegotiation of UPC parameters for arbitrary traffic sources in ATM networks," *IEEE/ACM Transactions on Networking*, vol. 6, pp. 811-827, Dec. 1998.

[16] D. Wu, and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630-643. Jul. 2003.

[17] W. Yu, L. Musavian and Q. Ni, "Tradeoff Analysis and Joint Optimization of Link-Layer Energy Efficiency and Effective Capacity Toward Green Communications," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3339-3353, May 2016.

[18] A. H. Anwar, K. G. Seddik, T. ElBatt and A. H. Zahran, "Effective Capacity of Delay-Constrained Cognitive Radio Links Exploiting Primary Feedback," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7334-7348, Sep. 2016.

[19] J. Choi, "Effective Capacity of NOMA and a Suboptimal Power Control Policy With Delay QoS," *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1849-1858, Apr. 2017.

[20] Q. Zhang, and S. A. Kassam, "Finite-state Markov models for Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 47, pp. 1688-1692, Nov. 1999.

[21] M. Hassan, M. M. Krunz, and I. Matta, "Markov-based channel characterization for tractable performance analysis in wireless packet networks," *IEEE Transactions on Wireless Communications*, vol. 3, no. 3, pp. 821-831, May 2004.

[22] Y. Zhao, M. Zhao, L. Xiao, and J. Wang, "Capacity of Time-Varying Rayleigh Fading MIMO Channels," *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, 2005, pp. 547-551.

[23] S. H. Ting, K. Sakaguchi and K. Araki, "A Markov-Kronecker model for analysis of closed-loop MIMO systems," *IEEE Communications Letters*, vol. 10, no. 8, pp. 617-619, Aug. 2006.

[24] M. A. Ruder , D. Ding, U. L. Dang, A. V. Vasilakos, and W. H. Gerstacker, "Joint user grouping and frequency allocation for multiuser SC-FDMA transmission," *Physical Communication*, vol. 8, pp. 91-103, 2013.

[25] J. Fan, G.Y. Li, Q. Yin, B. Peng, and X. Zhu, "Joint User Pairing and Resource Allocation for LTE Uplink Transmission," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2838-2847, Aug. 2012.

[26] X. Lu, Q. Ni, W. Li and H. Zhang, "Dynamic User Grouping and Joint Resource Allocation With Multi-Cell Cooperation for Uplink Virtual MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3854-3869, Jun. 2017.

[27] K. Zhu, E. Hossain, and A. Anpalagan, "Downlink Power Control in Two-Tier Cellular OFDMA Networks Under Uncertainties: A Robust Stackelberg Game," *IEEE Transactions on Communications*, vol. 63, no. 2, pp. 520-535, Feb. 2015.

[28] H. Zhang, Y. Xiao, L. X. Cai, D. Niyato, L. Song and Z. Han, "A Multi-Leader Multi-Follower Stackelberg Game for Resource Management in LTE Unlicensed," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 348-361, Jan. 2017.

[29] H. Zhang, Y. Xiao, S. Bu, D. Niyato, F. R. Yu and Z. Han, "Computing Resource Allocation in Three-Tier IoT Fog Networks: A Joint Optimization Approach Combining Stackelberg Game and Matching," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1204-1215, Oct. 2017.

[30] S. Ji, L. Tang, M. Zhang and S. Du, "Dual power allocation optimization based on stackelberg game in heterogeneous network with hybrid energy supplies," *China Communications*, vol. 14, no. 10, pp. 84-94, Oct. 2017.

[31] B. Yang, Z. Li, S. Chen, T. Wang and K. Li, "Stackelberg Game Approach for Energy-Aware Resource Allocation in Data Centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 12, pp. 3646-3658, Dec. 1 2016.

[32] W. Cheng, X. Zhang and H. Zhang, "Statistical-QoS Driven Energy-Efficiency Optimization Over Green 5G Mobile Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3092-3107, Dec. 2016.

[33] P. A. Samuelson. "Economics: An Introductory Analysis", *McGraw-Hill Inc*, US , Dec, 1995, pp. 352-373.

[34] http://www.mathworks.co.uk/help/optim/examples/binary-integer-programming.html, Dec. 2012.

[35] V. D. Blondel, R. Lambiotte, "Fast unfolding of communities in large networks," *Computer Science*, 2008.

[36] Recommendation ITU-R M.1225, "Guidelines for evaluation of radio transmission technologies for IMT-2000," *International Telecommunication Union*, 1997.

Xiaofeng Lu received the B.Sc. degree from Sichuan University, Chengdu, China, in 1996, the M.Sc. degrees from Hunan University, Changsha, China, in 1999, and the Ph.D. degree in communication and information systems from the Huazhong University of Science and Technology, Wuhan, China, in 2006. From 1999 to 2003, he was a Research and Development Engineer with the Wuhan Research Institute of Post and Telecommunications. He is currently an Associate Professor with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China. His main research interests lie in the area of broadband wireless communications, covering topics, such as resource allocation and virtualization, MU-MIMO, and OFDMA.

Qiang Ni (M'04-SM'08) received the Ph.D. degree in engineering from the Huazhong University of Science and Technology, Wuhan, China. He is currently a Full Professor and the Head of the Communication Systems Research Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. He had authored over 180 research papers in international journals and conferences. His main research interests lie in the area of future generation communications and networking systems, including energy and spectrum efficient green wireless communications, non-orthogonal multiple access, 5G, massive MIMO, SDN, game theory, heterogeneous networks, cognitive radio network systems, cloud networks, energy harvesting, IoT, vehicular networks, and big data analytics. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a Contributor to the IEEE Wireless Standards.

Danping Zhao was born in Shanxi, China. She received the BSc degree in communication engineering from Xidian University, Xi'an, China, in 2016. She is currently pursuing the MSc degree in communication and information system in the State Key Laboratory of Integrated Services Networks at Xidian University, Xi'an, China. Her research interests include SC-FDMA, virtual MIMO and resource allocation of wireless communications.

Wenchi Cheng (M'14-SM'18) received the B.S. and Ph.D. degrees in telecommunication engineering from Xidian University, China, in 2008 and 2014, respectively, where he is an associate professor. He joined the Department of Telecommunication Engineering, Xidian University, in 2013, as an assistant professor. He worked as a visiting scholar at the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA, from 2010 to 2011. His current research interests include 5G wireless networks and orbital-angular-momentum based wireless communications. He has published more than 70 international journal and conference papers in IEEE Journal on Selected Areas in Communications, IEEE Magazines, IEEE INFOCOM, GLOBECOM, and ICC, etc. He received the Young Elite Scientist Award of CAST, the Best Dissertation (Rank 1) of China Institute of Communications, the Best Paper Award for IEEECIC ICCC 2018, the Best Paper Nomination for IEEE GLOBECOM 2014, and the Outstanding Contribution Award for Xidian University. He has served or serving as the Associate Editor for IEEE Access, the IoT Session Chair for IEEE 5G Roadmap, the Publicity Chair for IEEE ICC 2019, the Next Generation Networks Symposium Chair for IEEE ICCC 2019, the Workshop Chair for IEEE ICC 2019 Workshop on Intelligent Wireless Emergency Communications Networks, the Workshop Chair for IEEE ICCC 2017 Workshop on Internet of Things.

Hailin Zhang received the B.S. and M.S. degrees from Northwestern Polytechnic University, Xi'an, China, in 1985 and 1988, respectively, and the Ph.D. degree from Xidian University, Xi'an, in 1991. He is currently a Full Professor and the Head of the School of Telecommunications Engineering, Xidian University. He has authored more than 80 papers in telecommunications journals and proceedings. His main research interests lie in the area of broadband wireless communications, including massive MIMO, OFDM, and space-time coding.