Multi-Scale Hierarchical Resource Management for Wireless Network Virtualization

Huijuan Jiang, Tianyu Wang, *Member, IEEE*, and Shaowei Wang, *Senior Member, IEEE*

***Abstract*—To deal with the conflict between limited radio resources and dramatic growth of wireless traffic, wireless network virtualization (WNV) is proposed as an efficient network sharing solution in 5G era. In this paper, we propose a multi-time-scale hierarchical model for WNV, which consists of primary users (PUs) with low latency requirement, secondary users (SUs) with high data rate requirement and tertiary users (TUs) without strict QoS requirement, such as machine-to-machine communications. The resource allocation scheme is decomposed into inter-slice subchannels pre-assignment in large time period and intra-slice subchannels and power scheduling in small time slot. In large time period, the subchannels pre-assignment problem is formulated as an integer optimization problem considering the requirement of PUs, SUs, and TUs. In small time slot, the subchannels and power scheduling of PUs is formulated as a mixed optimization problem with integer variables, for which we adopt Lyapunov drift-plus-penalty function with tradeoff factor V to transform the problem into a Lyapunov optimization problem and propose a two-step algorithm consisting of a heuristic subchannel assignment procedure and a fast barrier-based power allocation procedure. SUs access spectrum in a static manner and TUs access spectrum in a cognitive way. Simulation results show that our proposed algorithm achieves great performance in terms of users' performance with almost linear computation complexity.**

***Index Terms*—Wireless network virtualization, Lyapunov optimization, resource allocation.**

I. Introduction

WIRELESS traffic has experienced explosive growth in recent years. As reported in [2], mobile data traffic will go through a sevenfold increase between 2016 and 2021, among which machine-to-machine (M2M) communications grow more dramatically. In addition, broadband speeds will nearly double by 2021 and 5G wireless networks need to satisfy a wide range of subscribers with various communication requirements in terms of latency, throughput, energy efficiency, scalability, security, etc [3]–[6].

To solve the imbalance between limited wireless resources and increasing demand of users, network sharing mechanism is promising in 5G era, which allows infrastructure and radio resources shared among wireless users [7]. Wireless Network Virtualization (WNV) was proposed in recent years as a promising network sharing framework by isolating network resources, e.g., WNV can be described as a mechanism in which physical network resources, such as physical infrastructure, wireless spectrum, energy, and antennas are virtualized and sliced into multiple slices which are isolated from each other [8]. In the WNV, infrastructure provider (InP) and mobile virtual network operators (MVNOs) play different roles for network sharing. InP is responsible to virtualize their infrastructure and radio resources into slices and assign them to different users sectors. In each slice, MVNO leases the virtual resources to their subscribers.

In WNV, current schemes can be classified into central schemes and distributed schemes. In central schemes [9], the InP plays a central role who directly virtualizes resources and assigns virtual resources to end users. In distributed schemes [10], the InP virtualizes resources into different slices and assigns these slices to different users sectors, and the MVNO of each slice allocates virtual resources to its users, which decomposes the virtual resource assignment problem into a hierarchical problem [11].

In this paper, we propose a multi-time-scale hierarchical model for WNV, which consists of primary users (PUs), secondary users (SUs) and tertiary Users (TUs) as shown in Fig. 1. The InP assigns part of the spectrum to PUs sector for a given geographic area and time period, within which no SUs and TUs are allowed to access. Then, the InP assigns part of the spectrum unutilized by PUs to SUs sector for a given geographic area and time period, within which no TUs are allowed to access. TUs can access any frequency band that is not assigned to PUs and SUs for a given geographic area and time period, as long as the specific band is idle. The spectrum resources virtualized and assigned to PUs, SUs and TUs sectors are named as PU slice, SU slice and TU slice, respectively. The PU slice provides service for PUs with low-latency QoS requirement, the SU slice provides services for SUs with high transmission data rate QoS requirement, and the TU slice provides services for TUs with a large number of local and temporary transmissions with no strict QoS requirement in capacity, latency, and reliability, such as massive IoT communications in the 5G era.

Manuscript received April 30, 2018; revised August 23, 2018; accepted October 12, 2018. Date of publication October 25, 2018; date of current version December 21, 2018. This work was partially supported by the National Natural Science Foundation of China (61671233, 61801208), the Jiangsu Science Foundation (BK20170650), the Postdoctoral Science Foundation of China (BX201700118, 2017M621712), the Jiangsu Postdoctoral Science Foundation (1701118B), and the Fundamental Research Funds for the Central Universities (021014380094). Part of this work was presented at the IEEE International Conference on Communications, Kansas City, MO, USA, May 20–24, 2018 [1]. The associate editor coordinating the review of this paper and approving it for publication was J. Liu. *(Corresponding author: Shaowei Wang.)*

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: mg1723063@smail.nju.edu.cn; tianyu.alex.wang@nju.edu.cn; wangsw@nju.edu.cn).

Digital Object Identifier 10.1109/TCCN.2018.2878028

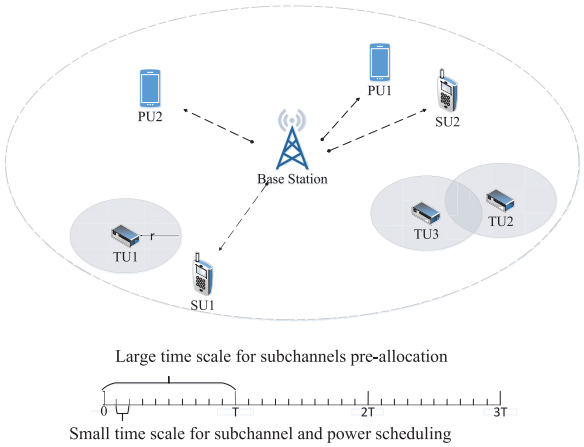

Fig. 1. Multi-scale hierarchical resource management model for wireless network virtualization.

We develop a hierarchical resource allocation scheme with two time scales. In large time period, the InP virtualizes spectrum band resources into three slices and assigns them to PUs, SUs and TUs sectors, respectively. In small time slot, MVNOs in PU slice allocate subchannals and power to end users to maximize the total utility while guaranteeing their QoS requirement. SUs access the subchannels in a static manner due to their stationary packets rate. In TU slice, we integrate a bunch of users in a specific frequency band. Each TU receives guiding information from the inter-slice subchannels allocation result and then accesses idle spectrum band opportunistically based on "listen-before-talk" scheme. Our contributions are summarized as follows:

- We propose a two-time-scale hierarchical scheme to reduce the complexity of the resource assignment task and increase independence among different slices effectively, where InP virtualizes spectrum resources into PU, SU and TU slices in large time period, and at the begin of each small time slot, MVNOs of PU slice assign subchannels and power to users to maximize total utility, SUs access the subchannels in a static manner and TUs access subchannels with the guidance of spectrum sensing result.
- To guarantee the network stability and balance the tradeoff between the network stability and the objective network performance with low complexity, we introduce Lyapunov drift-plus-penalty function with control parameter V to derive the tradeoff as $O[(1/V), O(V)]$.
- In TU slice, we integrate a bunch of TUs in one subchannels each of them can access subchannel opportunistically by cognitive radio to achieve a tradeoff between flexibility and reliability.
- To tackle the mixed optimization problem with integer variables, we propose a two-step process consisting of a heuristic subchannel assignment scheme and a fast barrier method for power allocation with nearly linear complexity.

The rest of the paper is organized as follows. Section II is the related work. In Section III, we present system model and formulate a mixed integer optimization problem. In Section IV, we transform the original optimization problem into a Lyapunov optimization problem by using Lyapunov drift-plus-penalty function, for which we introduce a heuristic subchannels assignment scheme and a fast barrier method for power allocation. Simulation results are provided in Section V and we conclude the paper in Section VI.

II. Related Work

Network slicing has attracted much attention in both academia and industry. 5G white paper gives several scenario examples providing different types of services to users with specific case including smart phones with high-throughput requirement, IoT communication requiring low-rate non-critical service and real-time communication asking for low latency [12]. Some works have implemented WNV customized for specific scenario. Reference [13] supports the requirement of vehicle-to-everything services, which improves transport fluidity, safety, and comfort on the road. And [14] supports network slicing for IoT services with high data rate, numerous devices connection and low service latency requirement.

With diverse users' QoS requirements, specific optimization problems are proposed for given user scenarios. Reference [15] gives a comprehensive survey on delay-aware control problem with several objectives and constraints including effective bandwidth, network stability, data drop rate, delay, power requirements and then implements Lyapunov stability drift approach and distributed stochastic learning to deal with the optimization problem. Reference [9] proposes a joint power and sub-carrier allocation problem to confirm queue stability with Lyapunov drift-plus-penalty algorithm. Dang *et al.* [16] aim to optimize energy efficiency with constraints on power consumption, queue stability and QoS requirement, and then transform the initial optimization problem via Lyapunov optimization approach with tradeoff factor V to trade off energy efficiency and delay and solves the optimization problem by Lagrange dual decomposition method. Reference [17] allocates resources with backhaul constraints for InP and the users' QoS requirement for wireless virtualization network.

As for slicing resources, [18] classifies the facet of slicing into three different level: spectrum-level, infrastructure-level and network-level. Reference [19] virtualizes power and spectrum and employs game-theory based schemes to solve the wireless resource allocation problem, while [20] proposes a hierarchical matching game-based scheme to satisfy efficient sub-channel allocation. Reference [21] proposes a novel virtual network embedding model to allocate three dimension resources consisting of computing, network and storage, and proposes two heuristic algorithms to solve the virtual network resource allocation problem. Reference [22] optimizes joint power, sub-carrier and antenna allocation problems for both perfect and imperfect channel knowledge cases.

Due to diverse QoS requirements of users' in wireless virtualization network and diverse type of slicing resources, several solution approaches are proposed to maximize the utility of resource allocation problem. Therein, [23] proposes a network slicing game where each slice tenant reacts to other tenants' decision to maximize its own utility. Reference [24]

proposes a mobile traffic forecasting model consisting of traffic analysis and predictor of mobile traffic, admission control decisions for network slice requests and adaptive correction of the forecasted mobile traffic load. Reference [25] adopts auction theory to create business models for heterogeneous scenario in the WNV.

Though quite a few works have discussed different issues facing the WNV, as far as the authors have known, the priority of users of different slices is not considered, which is the focus of this work. Besides, since the number of users in specific sector is stable, we assign the resource allocation task to InP and MVNOs to increase efficiency by introducing multi-time-scale hierarchical scheme. To solve the NP-hard optimization problem for PU and SU slice, we propose a fast barrier method with nearly linear complexity.

III. System Model

We consider an OFDM system with three types of users, i.e., PUs, SUs and TUs. For each time period *T*, total *L* subchannels are virtualized into three network slices according to the current user requirement, serving PUs, SUs and TUs, respectively. Within a time period, each slice manages its own subchannels for each time slot, by using centralized scheduling schemes or distributed cognitive radio techniques. We denote by L_p, L_s and L_t as the number of subchannels allocated to the PU slice, the SU slice and the TU slice, respectively.

A. Primary User Model

PUs are low-latency users with strict packet delay constraint, e.g., self-driving vehicles, medical equipments and industrial IoT devices. For simplicity without losing generality, we consider the downlink transmissions.

1) Large Time Period: We assume that the PU packets arrive at a Poisson process with arrival rate λ_p, and all packets are buffered within the network with an unlimited buffer size. The service rate of PU slice is given by $L_p, \mu_{p,0}$, in which $\mu_{p,0}$ represents the capacity of each PU subchannel. Due to the multiplexing gain, the spectrum efficiency increases with the number of subchannels. For simplicity, we assume that

$$\mu_{p,0} = \alpha log\big(1 + \beta L_p\big), \tag{1}$$

in which α and β are scaling factors. Thus, the service rate of the PU slice is given by

$$\mu_p\big(L_p\big) = L_p \alpha log\big(1 + \beta L_p\big). \tag{2}$$

Therefore, PUs packets arriving and serving process can be formulated as an $M/M/1/\infty$ queueing system with arrival rate λ_p and service rate μ_p. To ensure the occupation rate $\rho_p = \lambda_p/\mu_p < 1$, we assume the total subchannels *L* is enough such that $\lambda_p < \mu_p L$. For any PU packets arriving at the queue, the probability that there are *n* packets in the system is given by

$$p_n = \big(1 - \rho_p\big)\rho_p^n, n \geq 0. \tag{3}$$

The delay of the $(n+1)-th$ arriving packet is then given by the total service time of $n+1$ packets. As the service time of each PU packet follows a common negative exponential distribution with mean $1/\mu_p$, the total service time follows an Erlang-$(n+1)$ distribution with mean $(n+1)/\mu_p$. Thus the mean waiting time is yielded by:

$$W_{pu} = \frac{\rho_p/\mu_p}{1 - \rho_p}, \tag{4}$$

and the cumulative distribution function is given by

$$F_{n+1}(t) = 1 - \sum_{k=0}^{n} \frac{\big(\mu_p t\big)^k}{k!} e^{-\mu_p t}, t \geq 0. \tag{5}$$

Therefore, the probability that the delay of the arriving packet exceeds the threshold D_p is given by

$$\begin{aligned} Pr\big\{ d_{pu} \geq D_p \big\} &= \sum_{n=0}^{\infty} p_n \big[1 - F_{n+1}\big(D_p\big)\big] \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{\big(\mu_p D_p\big)^k}{k!} e^{-\mu_p D_p} \big(1 - \rho_p\big)\rho_p^n \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{\big(\mu_p D_p\big)^k}{k!} e^{-\mu_p D_p} \big(1 - \rho_p\big)\rho_p^n \\ &= \sum_{k=0}^{\infty} \frac{\big(\mu_p \rho_p D_p\big)^k}{k!} e^{-\mu_p D_p} \\ &= e^{-\mu_p (1-\rho_p) D_p}. \end{aligned} \tag{6}$$

Consider PUs' strict requirement in latency, we regard delay as the optimization indicator for PUs. Therefore, for the inter-slice subchannels assignment in large time period, we try to minimize the mean delay of PUs while guaranteeing the probability that the arriving packet exceeds the threshold D_p.

2) Small Time Slot: We consider PU slice serves *N* PUs, and the number of subchannels in PU slice is L_p. Since the subchannels are orthogonal, signal in different subchannels would not interference each other. Let $c_{n,l}$ denote the channel gain of *n*th PU in *l*th subchannel, N_0 denote the PSD of additive white Gaussian noise (AWGN) and $\Gamma = -ln(5BER)/1.5$ is specified BER for an uncoded MQAM, the signal-to-interference-plus-noise (SINR) for *n*th PU in *l*th subchannel is $H_{n,l} = \frac{|c_{n,l}|^2}{\Gamma(N_0 W/L)}$. Then we can derive the throughput of *n*th PU over *l*th subchannel

$$r_{n,l} = \rho_{n,l} log\big(1 + p_{n,l} H_{n,l}\big), \tag{7}$$

where $\rho_{n,l}$ represents whether *l*th subchannel is assigned to *n*th PU, which is a binary variable, i.e., $\rho_{n,l} \in \{0, 1\}$, and $p_{n,l}$ denotes the power allocated to *n*th users over *l*th subchannel. The MVNO of PU slice assigns subchannel and power to PUs each small time slot *t*, $t \in \{0, 1, 2, \ldots, T\}$.

For simplification, we assume that the SINR is constant during one small time slot, and data packets of each user arrive each time slot, denoted by $\mathbf{A(t)} = \{A_1(t), A_2(t), \ldots, A_N(t)\}$. **A(t)** follows Possion distribution with mean λ_{pu} and is independent and identically distributed (i.i.d.). $\mathbf{Q(t)} = \{Q_1(t), Q_2(t), \ldots, Q_N(t)\}$ denotes data packets stored in the queue of *n*th user, this yields,

$$Q_n(t+1) = [Q_n(t) - R_n(t)]^+ + A_n(t), \forall n. \tag{8}$$

The arrive-departure is depicted as Fig. 2.

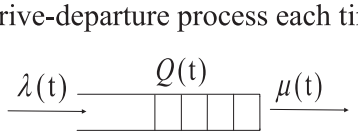

Fig. 2. Arrive-departure process of PU each time slot.

The stability of network is guaranteed if all users's queue stability is guaranteed, and the stability of *n*th user's queue is guaranteed if

$$\lim_{t\to\infty}\frac{E|Q_n(t)|}{t}=0,\forall n. \tag{9}$$

Definition 1: A discrete time process $Q(t)$ is mean rate stable if

$$\lim_{t\to\infty}\frac{E|Q(t)|}{t}=0. \tag{10}$$

To satisfy the QoS requirement, the probability that the queue length of each user is larger than the upper bound of queue length Q^{up} should be smaller than ε, that is

$$\lim_{t\to\infty}Pr\{Q_m(t)\geq Q^{up}\}\leq\varepsilon,\forall n. \tag{11}$$

Besides, each user's throughput each small time slot is required to exceed lower bound R^{low}, that is,

$$R_n(t)\geq R^{low},\forall n. \tag{12}$$

In small time slot, we allocate subchannels and power among PUs with various packets arrival rate and packets stored in the queue to minimize the total data packets stored in the queue while guaranteeing the QoS requirement consisting of the stability of network, the delay of each PU and the throughput of each PU each time slot.

B. Secondary User Model

SUs are broadband users with a stationary data rate requirement, e.g., high-definition video users and virtual reality users. We also consider the downlink transmissions. We assume that each SU requires a logical channel with a stationary r_s packets rate. Also, due to the multiplexing gain, we assume that the packet rate of each subchannel is given by

$$\mu_{s,0}=\alpha log(1+\beta L_s). \tag{13}$$

Thus, the total packet rate of the SU slice is given by

$$\mu_s=L_s\alpha log(1+\beta L_s). \tag{14}$$

Therefore, the SU slice can support total $N_s=\mu_s/r_s$ SUs.

In each time period, the InP allocates L_s subchannels to SU slice to maximize the number of SUs that the SU slice can support. In each time slot, SUs access the subchannels in a static manner due to their stationary pakcets rate.

C. Tertiary User Model

TUs are IoT users without strict QoS requirement in terms of capacity, latency and reliability. For simplicity, we consider the uplink transmission. Consider the limited spectrum resources, a finite buffer size *K* is applied and the arriving and serving process is formulated as an *M*/*M*/1/*K* queueing system. We consider that TU slice serve N_t TUs, thus $n_t=N_t/L_t$ TUs are integrated in one subchannel. The TUs opportunistically access the spectrum band by using spectrum sensing technique. The sensing capability is characterized by false alarm probability p_f and mis-detection probability p_m, which is assumed to be the same for all users. We assume that all users are synchronized in time slots, the packets of each TU arrive at a Possion distribution with arrival rate λ_t and the transmission time of one TU packet is constant, denoted as $1/\mu_t$. Also, we assume that all users have the same interference radius *r*, i.e., any two transmitting users will interfere with each other only if their distance is within *r*.

Consider a subchannel with total n_t users. We denote p_{tran} as the probability that a user transmits in a time slot, and p_{idle} as the probability that the band is decided to be idle during a time slot. For simplicity, we assume that detection of different user signals are independent from each other. Thus, we have

$$p_{idle}=\left[\left(1-\frac{r^2}{R^2}\right)+\frac{r^2}{R^2}p_{tran}p_m+\frac{r^2}{R^2}(1-p_{tran})\big(1-p_f\big)\right]^{n_t-1}. \tag{15}$$

For the Poisson process of TUs, the probability that *n* packets arrive at a single TU during a time slot is given by

$$q_n=\frac{(\lambda_t/\mu_t)^n}{n!}e^{-\lambda_t/\mu_t}. \tag{16}$$

We denote $\pi_n, n=0,1,\ldots$ as the steady probability that the TU has *n* packets in the buffer at the beginning of each time slot, and $p_{i,j}$ as the transition probability from state π_i to state π_j. Thus, we have

$$p_{i,j}=\begin{cases}q_j & i=0\\ q_0p_{idle} & i\geq 1, j=i-1\\ q_{j-i+1}p_{idle}+q_{j-i}(1-p_{idle}) & i\geq 1, j\geq i\\ 0, & \text{others}\end{cases},$$

and the corresponding transition matrix is defined as

$$\mathbf{P}=\begin{bmatrix}p_{0,0} & p_{1,0} & p_{2,0} & \cdots\\ p_{0,1} & p_{1,1} & \cdots & \\ p_{0,2} & \cdots & & \\ \vdots & & & \end{bmatrix}. \tag{17}$$

Thus, we have that the steady probability vector $\pi=(\pi_0,\pi_1,\ldots,)^T$ satisfies

$$\pi\mathbf{P}=\pi, \tag{18}$$

and

$$\pi\mathbf{1}=1, \tag{19}$$

where $\mathbf{1}=(1,1,\ldots)^T$. Combining equations (18) and (19), the steady probability π_n can be numerically calculated as a function of p_{idle}, i.e., $\pi_n(p_{idle})$, and thus, as a function of p_{tran}, i.e., $\pi_n(p_{tran})$. Note that the probability that the user needs to transmit in a time slot is given by $1-\pi_0$. We have

$$(1-\pi_0)p_{idle}=p_{tran}. \tag{20}$$

Therefore, p_{tran} can be numerically calculated from (20).

For any TUs, the probability that the band is correctly decided to be idle is given by

$$p_{idle}^{+} = \left[\left(1 - \frac{r^2}{R^2}\right) + \frac{r^2}{R^2}(1 - p_{tran})\big(1 - p_f\big)\right]^{n_t - 1}, \quad (21)$$

The throughput of each TU is then given by

$$r_t = \lambda_0 \cdot (1 - \pi_0) \cdot p_{idle}^{+}. \quad (22)$$

Consider one TU, the occupy rate is $\rho_t = \lambda_t / r_t$. For any packets arriving, the probability that there are $n \in [0, K]$ packets in the buffer is given by

$$w_n = \rho_t^n \frac{1 - \rho_t}{1 - \rho_t^{K+1}}, 0 \le n \le K. \quad (23)$$

Thus average date drop probability of TU is

$$Pr\big\{\overline{d_{tu}}\big\} = w_K. \quad (24)$$

In each time period, to guarantee the performance of TU packets, the InP allocate subchannnels to TU slice to guarantee the average data drop rate of TUs. In each time slot, each TU senses the subchannel which it belongs to and accesses the subchannel with probability p_{tran}.

D. Problem Formulation

The resource allocation problem is decomposed into inter-slice subchannels assignment in large time period and intra-slice subchannels and power scheduling in small time slot.

1) Large Time Period: In large time period, the InP assigns subchannels to PU, SU and TU slice to optimize the performance of network. We try to minimize the average delay of PUs and maximize the number of SUs that the SU slice can support while guaranteeing the delay requirement of PUs and the average packet drop rate of TUs. The optimization task can be formulated as follows:

$$\begin{aligned} OP1 \quad & \max_{L_p, L_s, L_t} \big[W_{pu}, N_{su}\big] \\ s.t. \quad & C_1 : Pr\big\{d_{pu} \ge D_p\big\} \le P_p, \\ & C_2 : Pr\big\{\overline{d_{tu}}\big\} \le \varepsilon_t, \\ & C_3 : L_p + L_s + L_t \le L, \end{aligned} \quad (25)$$

where W_{pu} and N_{su} represent the average delay of PUs and the number of SUs that the SU slice can support, respectively; C_1 means the probability that PUs' delay d_{pu} exceeds delay threshold D_p is not larger than the probability threshold P_p. C_2 informs that the average data drop rate of TUs can not be larger that the threshold ε_t. C_3 indicates the spectrum resources constraint. To convert this multi-objective optimization into single-object optimization, we apply a linear weighted sum method, then the utility function is converted into $u = -\alpha_p \widehat{W_{pu}} + \alpha_s \widehat{N_{su}}$, where the weighted factors α_p and α_s represent the importance of PU and SU utility, respectively. Consider the magnitude of delay and numbers of SUs, the average delay of PUs and the numbers of SU that the SU slice can support are normalized to an order of magnitude before summing up, i.e., $\widehat{W_{pu}}$ and $\widehat{N_{su}}$. We adopt enumeration method to solve *OP*1 by enumerating L_p and L_s.

2) Small Time Slot: In each small time slot, SUs access the spectrum in a static manner and TUs access the spectrum opportunistically by cognitive radio. For PU slice, we try to minimize the total queue length of PUs while guaranteeing the QoS requirement of PUs and total transmission power budget of PU slice. Mathematically, the optimization problem is described as follows,

$$\begin{aligned} OP2 \quad & \min_{p_{n,l}, \rho_{n,l}} \sum_{n=1}^{N} Q_n(t) \\ s.t. \quad & C_1 : \lim_{t \to \infty} \frac{E|Q_n(t)|}{t} = 0, \forall n, \\ & C_2 : \lim_{t \to \infty} Pr\{Q_n(t) \ge Q^{up}\} \le \varepsilon, \forall n, \\ & C_3 : R_n(t) \ge R^{low}, \forall n, \\ & C_4 : \sum_{n=1}^{N} \sum_{l=1}^{L} \rho_{n,l} p_{n,l} \le P_p^t, \\ & C_5 : p_{n,l} \ge 0, \forall n, \forall l, \\ & C_6 : \sum_{n=1}^{N} \rho_{n,l} = 1, \forall l, \\ & C_7 : \rho_{n,l} \in \{0, 1\}, \forall n, \forall l, \end{aligned} \quad (26)$$

where C_1, C_2 and C_3 guarantee PUs' QoS requirement, among which C_1 guarantees the stability of PUs, C_2 constrains the queue length of each PU and C_3 guarantees that each PU's throughput exceeds lower bound. C_4 and C_5 are the transmission power constraints, where P_p^t is the maximum allowable total power each small time slot. C_6 and C_7 inform that each subchannel cannot be shared among different users each time.

IV. Proposed Primary User Algorithms

In this section, we introduce an efficient algorithm for PUs subchannels and power allocation problem with low complexity. First, We use lyapunov optimization to transform OP2 into an equivalent drift-plus-penalty minimization problem. Then, we propose a two-step process consisting of a heuristic subchannels allocation procedure and a fast barrier based power allocation procedure to solve the mixed integer optimization problems.

A. Lyapunov Optimization

Lyapunov drift-plus-penalty function introduces a time average penalty and tradeoff factor *V* to trade off the time average penalty with *O*(1/*V*) and average queue length with *O*(*V*) [26]. This method aims to maximize network utility subject to network stability.

According to [27], C_1 in *OP*2 can be satisfied if $\lim_{t \to \infty} E\{Q_n(t)\} \le Q^{up}\varepsilon, \forall n$. Consider Poisson process, we have $E\{Q_n(t)\} = \lambda_{pu} - R_n(t)$, then C_1 in *OP*2 is converted into

$$\lim_{t \to \infty} \lambda_{pu} - R_n(t) \le Q^{up}\varepsilon, \forall n. \quad (27)$$

To tackle this constraint, we introduce virtual queue [28], evolving $H_n(t)$ with $H_n(t) = 0$ and update the equation as follows,

$$H_n(t+1) = [H_n(t) + y_n(t)]^{+}, \forall n, \quad (28)$$

where $y_n(t) = \lambda_{pu} - R_n(t) - Q^{up}\varepsilon$. Then $Q_n(t)$ and $H_n(t)$ are actual and virtual queues, respectively.

Let $\Theta(t) \triangleq (\mathbf{Q(t)}, \mathbf{H(t)})$, then Lyapunov function is represented as

$$L(\Theta(t)) \triangleq \frac{1}{2}\left\{\sum_{n=1}^{N}(Q_n(t))^2 + \sum_{n=1}^{N}(H_n(t))^2\right\}. \quad (29)$$

This denotes a scalar measure of queue congestion in PU slice. Larger $L(\Theta(t))$ implies larger queue backlog.

Denote $\Delta(\Theta(t))$ as the conditional Lyapunov drift in time slot *t*:

$$\Delta(\Theta(t)) \triangleq E\{L(\Theta(t+1)) - L(\Theta(t))|L(\Theta(t))\}, \quad (30)$$

which depends on random channel states and control actions in reaction to channel states. Then the drift-plus-penalty function of *OP*2 is $\Delta(\Theta(t)) + VE\{\sum_{n=1}^{N} Q_n(t)|\Theta(t)\}$.

Theorem 1: For general control policy, an upper bound of the drift-plus-penalty function exists, that is,

$$\Delta(\Theta(t)) + VE\left\{\sum_{n=1}^{N} Q_n(t)|\Theta(t)\right\} \le B - \sum_{n=1}^{N} H_n(t)R_n(t) - \sum_{n=1}^{N}(\lambda_{pu} - Q^{up}\varepsilon)R_n(t) - \sum_{n=1}^{N} VQ_n(t)R_n(t). \quad (31)$$

Then, *OP*2 can be converted into

$$\begin{aligned} \max_{p_{n,l},\rho_{n,l}} & \sum_{n=1}^{N}\sigma_n(t)\left(\sum_{l=1}^{L}\rho_{n,l}r_{n,l}\right) \\ s.t.\ \ C_1 &: R_n(t) \ge R^{low}, \forall n, \\ C_2 &: \sum_{n=1}^{N}\sum_{l=1}^{L}\rho_{n,l}p_{n,l} \le P_p^t, \\ C_3 &: p_{n,l} \ge 0, \forall n, \forall l, \\ C_4 &: \sum_{n=1}^{N}\rho_{n,l} = 1, \forall l, \\ C_5 &: \rho_{n,l} \in \{0,1\}, \forall n, \forall l, \end{aligned} \quad (32)$$

where $\sigma_n(t) = \lambda_{pu} - Q^{up}\varepsilon + H_n(t) + VQ_n(t)$.

B. Subchannel Allocation

Due to the binary variable $\rho_{n,l}$, (32) is NP-hard with mixed integer programming. To solve this problem, we propose a heuristic subchannel allocation method to remove $\rho_{n,l}$. In the subchannel allocation process, we preset that total transmission power is divided equally among subchannels, i.e., $p_{n,l}^{pre} = \frac{P_p^t}{N}$. Then the presetting throughput of *n*th user over *l*th subchannel is

$$r_{n,l}^{pre}(t) = log\Big(1 + p_{n,l}^{pre}H_{n,l}(t)\Big). \quad (33)$$

Let $\Omega_n(t)$ denotes subchannels allocated to *n*th users in time slot *t*. We propose a heuristic subchannels allocation method to remove the integer variable. We prefer to allocate subchannel to user with larger $\sigma_n(t)$ and maximal presetting rate in this subchanel among all users. The subchannels allocation is shown in Table I.

TABLE I
SUBOPTIMAL SUBCHANNEL ALLOCATION

1.**Initialization**:
2. Set $\Omega_n(t) = \varnothing, R_n = 0, \forall n, \mathcal{N}_r = \mathcal{N}, \mathcal{L}_r = \mathcal{L}$.
3.**First Allocation:**
4. **while** $\mathcal{N}_r \neq \varnothing$
5. Find n^*, l^* that $\sigma_n^*(t)r_{n^*,l^*}^{pre} \ge \sigma_n(t)r_{n,l}^{pre}, \forall n \in \mathcal{N}_r, l \in \mathcal{L}_r$;
6. $\Omega_{n^*}(t) = \Omega_{n^*}(t) \cup l^*, \mathcal{N}_r = \mathcal{N}_r \setminus n^*, \mathcal{L}_r = \mathcal{L}_r \setminus l^*, R_n = r_{n^*,l^*}^{pre}$
7. **end while**
8.**Second Allocation:**
9. **while** $\mathcal{L}_r \neq \varnothing$
10. Find $n^* = arg\min_{n\in\mathcal{N}} R_n/\sigma_n(t)$;
11. Find l^* that $r_{n^*,l^*}^{pre} \ge r_{n^*,l}^{pre}, \forall l \in \mathcal{N}_r$;
12. $\Omega_{n^*}(t) = \Omega_{n^*}(t) \cup l^*, \mathcal{L}_r = \mathcal{L}_r \setminus l^*, R_n = R_n + r_{n^*,l^*}^{pre}$.
13. **end while**

C. Power Allocation

Given subchannel assignment, the optimization problem (32) is converted into

$$\begin{aligned} \max_{p_{n,l}} & \sum_{n=1}^{N}\sigma_n(t)\sum_{l\in\Omega_n} log\big(1 + p_{n,l}H_{n,l}\big) \\ s.t.\ \ C_1 &: R_n(t) \ge R^{low}, \forall n, \\ C_2 &: \sum_{n=1}^{N}\sum_{l\in\Omega_n} p_{n,l} \le P_p^t, \\ C_3 &: p_{n,l} \ge 0, \forall n, \forall l. \end{aligned} \quad (34)$$

Eq. (34) is a convex optimization problem since the objection function is convex and all constraints are affine. Barrier method is one of standard convex optimization techniques to solve this kind of optimization problem. We convert the original problem into a sequence of unconstrain minimization problems by introducing logarithmic barrier function with a given parameter *t*. The barrier method is decomposed into centering step and Newton step. The centering step is the outer iteration to compute the central point. With the increase of *t*, the central point is closer to the optimal solution of the original problem. The complexity of Newton method lies in the inner iteration which needs to inverse matrix with complexity $O(L^3)$.

Since the subchannels in PU slice is numerous, the complexity of the barrier method is intolerant for each time slot. We propose a fast barrier method that reduces the complexity to be linear to the number of subchannels.

We convert the optimization problem into the following unconstrained minimization problem by using barrier method.

$$\min \Psi(P) = -tf(P) + \phi(P), \quad (35)$$

where $\mathcal{P} = \{p_{1,1}, \ldots, p_{n,l} \ldots, p_{N,L}\}$ is a $N\times L$ matrix, among which $p_{n,l}$ denotes the power allocated to *n*th PU over *l*th subchannel. And

$$f(P) = \sum_{n=1}^{N}\sigma_n(t)\sum_{l\in\Omega_n} log\big(1 + p_{n,l}H_{n,l}\big). \quad (36)$$

TABLE II
THE BARRIER METHOD

1.**Initialization**
2. Feasible point: P; Trade-off parameter: $t > 0, \mu > 1$;
3. Tolerance: $\epsilon > 0, \epsilon_n > 0; \alpha \in (0, 1/2), \beta \in (0, 1)$.
4. **Outer loop: centering step**
5. **while:** $(N+1)/t < \varepsilon$
6. Update $P^*(t) = P$, P is computer by Newton step.
7. **Inner loop: Newton step**
8. Compute ΔP_{nt} and $\lambda^2 := - \bigtriangledown \Psi_t(P) \Delta P_{nt}$.
9. **while:** $\lambda^2/2 > \epsilon_n$
10. **Backtracking line research:** $s := 1$
11. **while**: $\Psi_t(P + s\Delta P_{nt}) > \Psi_t(P) - \alpha s \lambda^2$
12. $s := \beta s$.
13. **End while**
14. Update: $P := P + s\Delta P_{nt}$.
15. **End while**
16. Update: $t := \mu t$.
17. **End while**

$\phi(P)$ is the logarithmic barrier function of (34)

$$\phi(P) = -\sum_{n=1}^{N} \sum_{l \in \Omega_n} log\Big(log\big(1 + p_{n,l} H_{n,l}\big) - R^{low}\Big) - log\left(P_p^t - \sum_{n=1}^{N} \sum_{l \in \Omega_n} p_{n,l}\right) - \sum_{n=1}^{N} \sum_{l \in \Omega_n} log p_{n,l}. \tag{37}$$

The process of barrier method is shown as Table II. In each centering step, we use Newton method to compute Newton step ΔP_{nt}, denoted as

$$\nabla^2 \Psi(P) \Delta P_{nt} = -\nabla \Psi(P), \tag{38}$$

where $\nabla^2 \Psi(P)$ is the Hessian matrix and $\nabla \Psi(P)$ is the gradient of $\Psi(P)$. Apparently, the complexity of computing Newton step ΔP_{nt} by matrix inversion directly is too high, we develop a fast barrier method with lower complexity. Denote

$$f_0 = P_p^t - \sum_{n=1}^{N} \sum_{l \in \Omega_n} p_{n,l},$$
$$f_n = R_n(t) - R^{low}, \forall n. \tag{39}$$

The Hessian of $\Psi(x)$ is

$$\frac{\partial^2 \Psi_t(P)}{\partial p_{k,n}^2} = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_N \end{bmatrix} + \frac{\bigtriangledown f_0 \bigtriangledown f_0^T}{f_0^2} + \sum_{n=1}^{N} \frac{\bigtriangledown f_n \bigtriangledown f_n^T}{f_n^2} = D + \sum_{k=1}^{K} g_i g_i^T, \tag{40}$$

where $\mathcal{D} = diag(D_1, D_2, \ldots, D_N)$ and

$$D_n = \frac{1}{p_{n,l}^2} + \left(t + \frac{1}{f_n}\right) \frac{H_{n,l}^2}{\big(1 + p_{n,l} H_{n,l}\big)^2}. \tag{41}$$

$K = 1 + N$ and

$$g_i = \begin{cases} \frac{\bigtriangledown f_0}{f_0}, & i = 1 \\ \frac{\bigtriangledown f_n}{f_n} & n = 1, \ldots, N, i = n + 1. \end{cases} \tag{42}$$

TABLE III
THE FAST BARRIER METHOD

1. **Step 1:**
2. Decompose Λ_0, $\Lambda_0 = \Lambda_1 + g_1 g_1^T$;
3. Then $\Delta P = \upsilon_1^1 - \frac{g_1^T \upsilon_1^1}{1 + g_1^T \upsilon_2^1} \upsilon_2^1$,
4. where $\Lambda_1 \upsilon_1^1 = g_0$ and $\Lambda_1 \upsilon_2^1 = g_1$.
5. **Step 2:**
6. Decompose Λ_1, $\Lambda_1 = \Lambda_2 + g_2 g_2^T$.
7. Then $\upsilon_i^1 = \upsilon_i^2 - \frac{g_2^T \upsilon_i^2}{1 + g_2^T \upsilon_3^2} \upsilon_3^2, i = 1, 2,$
8. where $\Lambda_2 \upsilon_i^2 = g_{i-1}, i = 1, 2, 3.$
⋮
9. **Step k:**
10. Decompose Λ_{k-1}, $\Lambda_{k-1} = \Lambda_k + g_k g_k^T$.
11. Then $\upsilon_i^{k-1} = \upsilon_i^k - \frac{g_k^T \upsilon_i^k}{1 + g_k^T \upsilon_{k+1}^k} \upsilon_{k+1}^k, i = 1, \ldots, k,$
12. where $\Lambda_k \upsilon_i^k = g_{i-1}, i = 1, \ldots, k+1.$
⋮
13. **Continue** this process to Kth step.
14. **Step K+1:**
15. We obtain $K+1$ matrix $\Lambda_K \upsilon_i^K = g_{i-1}, i = 1, \ldots, K+1$
16. and variables $\upsilon_i^{k-1}, i = 1, \ldots, k$ in the kth step.
17. Then by υ_i^M, we can derive $\triangle P$.

We propose a $(K+1)$-step iterative algorithm to compute Newton step with complexity $O(K^2 L)$. By decomposing

$$\Lambda_i = \Lambda_{i+1} + g_{i+1} g_{i+1}^T, i = 0, 1, \ldots, K-1, \tag{43}$$

where

$$\Lambda_i = D + \sum_{j=i+1}^{K} g_j g_j^T, i = 0, 1, \ldots, K-1. \tag{44}$$

the Hessian matrix can be converted into

$$\Lambda_0 = D + \sum_{i=1}^{K} g_i g_i^T, \tag{45}$$

where $\Lambda_0 = \nabla^2 \Psi(P)$ and $g_0 = -\nabla \Psi(P)$.

The fast barrier process is shown as Table III, where we derive $K+1$ matrix system $\Lambda_K = \upsilon_i^K g_{i-1}$, k variables $g_i^{k-1}, i = 1, \ldots, k$ in step $k-1$ and $m+1$ variables $g_i^k, i = 1, \ldots, k+1$ in step k. Obviously, the matrix inversion is imperative after we solve the $K+1$ matrix $\Lambda_K \upsilon_i^K = g_{i-1}, i = 1, \ldots, K+1$, which cost $O(LK)$. Since $\mathcal{D}$ is a diagonal matrix, we can obtain $\upsilon_i = D^{-1} g_i, i = 1, \ldots, L$, with complexity $O(L)$. Therefore, the total complexity is $O(K^2 L)$ [29], [30].

V. SIMULATION RESULTS

We verify our proposed multi-scale hierarchical resource management scheme via numerical experiment. Consider an OFDM-based wireless virtualization network, where all users are located in a circle area with radius 20km. The path loss exponent is 4, the variance of shadowing effect is 10dB and the amplitude of multipath fading is Rayleigh. The simulation parameters are given in Table IV, including the packet arrival rate λ_p, the weighted factor α_p, α_s, the false alarm and mis-detection probability p_f, p_m, and the tradeoff factor V in Lyapunov drift-plus-penalty function.

Firstly, we study the performance of PUs in time slot. In Fig. 3, the average queue length of PUs is shown as a

TABLE IV
SIMULATION PARAMETERS

r = 10km	Interference range of TUs
L = 400	Total number of subchannels
μ_t = 10	Capacity of a subchannel to support TUs
λ_p = 400	Total arrival rate of PU packets
r_s = 1	Stationary packets rate SUs
λ_t = 0.002	Arrival rate of TU
D_p = 3	Delay threshold of PUs
P_p = 0.2	Probability threshold of PUs
N_t = 3000	Total number of TUs
ε_t = 0.01	Data drop rate threshold of TUs
p_m = 0.01	Mis-detection probability
p_f = 0.01	False alarm probability
$N_0 = 10^{-15}$W	Noise power
$P_p^T = 10^{-5}$W	Power constraint of PUs
N = 50	Number of PUs
λ_{pu} = 8	Arrival rate of PUs
R^{low} = 2	Throughput lower bound of PUs

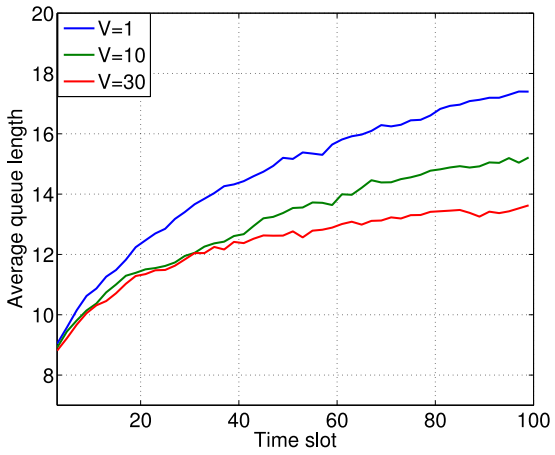

Fig. 3. The average queue length of PUs as a function of time slot with tradeoff factor V changing from 1 to 30.

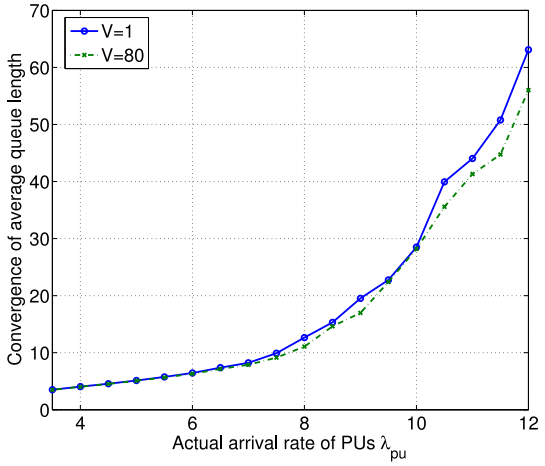

Fig. 4. The convergence of average queue length of PUs as a function of the arrival rate with tradeoff factor $V = 1$ and $V = 80$.

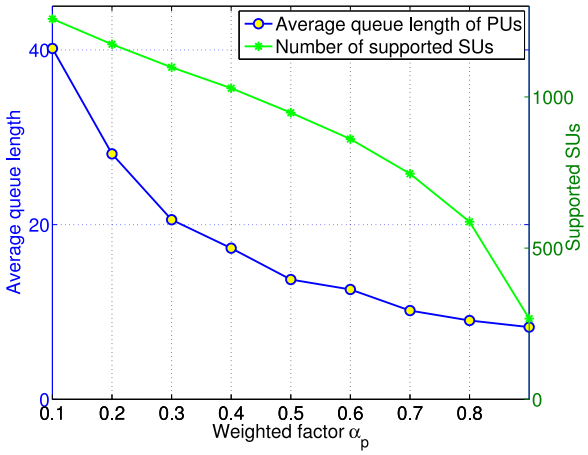

Fig. 5. The convergence of average queue length of PUs and the number of SUs that SU slice can support as a function of weighted factor α_p.

function of small time slot with the tradeoff factor V taking the value of 1, 10 and 30, respectively. We can see that as time goes on, the general trend of average queue length converges, with light fluctuation in a narrow range. The average queue length fluctuates because of the changing arrival rate each time slot. Besides, with the increase of tradeoff factor V, the queue length converges to a small value at a faster rate due to the preference for the objective function, which indicates the tradeoff between average queue length of PUs and the network stability.

Fig. 4 shows the convergence of average queue length of PUs as a function of the arrival rate with tradeoff factor $V = 1$ and $V = 80$. As the increase of arrival rate, the convergence of average queue length increases due to the limited capacity. The capacity of resources in PU slice is implied by the slope of queue length. The difference of convergence of average queue length with $V = 1$ and $V = 80$ is slight for small arrival rate since the system capacity is enough for small arrival rate.

Then we study how the weighted factors of PUs and SUs impact the performance of PUs and SUs. Fig. 5 shows the convergence of average queue length of PUs and the number of SUs that SU slice can support as a function of weighted factor α_p. It's obvious that with the increase of α_p, the average queue length of PUs and the number SUs that SU slice can support decrease, while the slope of PU's queue length becomes smaller and smaller and the slope of SUs supported becomes larger and larger. That is because with the increase of L_p, the expected delay of PU decreases sharply and then stabilizes in a small value, while the SUs that the SU slice can support decreases at a constant speed.

Finally, we investigate the impact of TUs' spectrum sensing capacity to system performance. Fig. 6 informs the delay of PU and the number of SUs that the SU slice can support as a function of p_f with $p_m = 0.01$. It can be seen that as the increase of false alarm probability, the performance of PUs and SUs decrease with a smaller slope when p_f is small and a larger slope when p_f is large. Especially when $p_f > 0.1$, the network performance declines dramatically until the delay of PUs increases to infinity and the number of SUs that SU slice can support decreases to zero.

Fig. 7 illustrates the delay of PUs and the number of SUs that the SU slice can support as a function of p_m with $p_f = 0.01$. When the mis-detection probability is small, the performance of PUs and SUs decrease lightly as the increase of p_m. When the mis-detection probability is large, the number of SUs that the SU slice can support decreases dramatically to

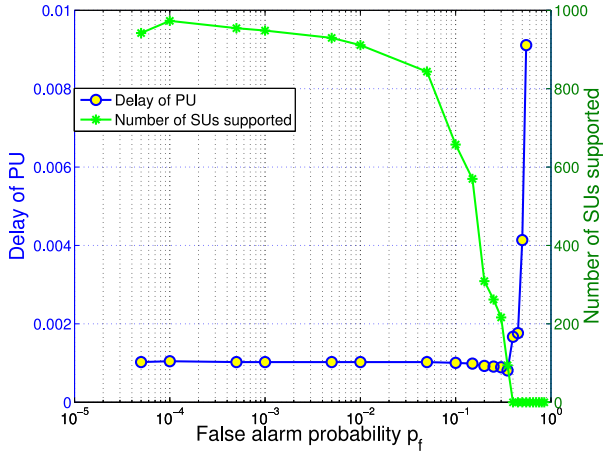

Fig. 6. The delay of PUs and the number of SUs that the SU slice can support as a function of p_f with $p_m = 0.01$.

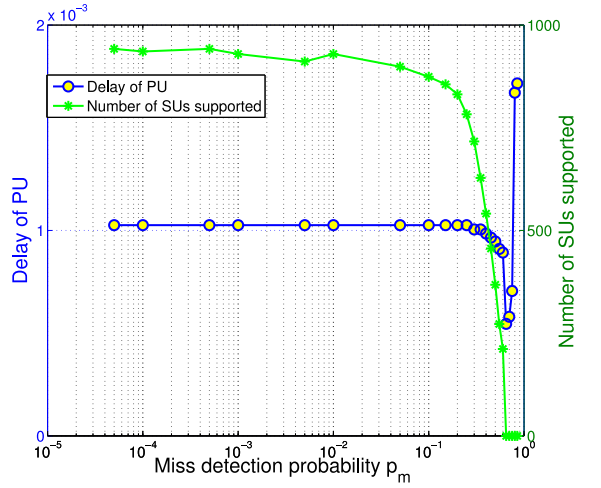

Fig. 7. The delay of PUs and the number of SUs that the SU slice can support as a function of p_m with $p_f = 0.01$.

zeros while the delay of PUs improves and then degenerates. That is because the number of necessary TU subchannels with increasing p_m under guaranteed data drop rate increases not so dramatically compared with the number of TU subchannels with increasing p_f. Due to the characteristics of PUs' delay function, the InP prefers to allocate more subchannels to PU slice, so that the delay of PUs decreases slightly and then increases to infinity.

VI. Conclusion

In this paper, we proposed a two-time-scale resource management scheme, in which frequency resources are statically allocated to different slices in each time period, and each network slice schedules its own subchannels and power in each time slot. We have defined three types of users, PUs, SUs and TUs, and focus on packet delay, data rate and IoT throughput. To satisfy the diverse requirements of different users and maximize the network utility, we have proposed a large-time-period frequency allocation algorithm, a small-time-slot subchannels and power scheduling algorithm using Lynapunov optimization. A two-step procedure is developed to tackle the intractable optimization task, which consists of a heuristic subchannels assignment scheme and a fast barrier method for power allocation with nearly linear complexity. Simulation results show the effectiveness and the efficiency of our proposed method. However, we do not fully consider the fairness among users in each slice, which should be investigated in future work.

Appendix
Proof of Theorem 1

For $Q_n(t)$, since

$$\left\{[Q-R]^+ + A\right\}^2 \leq Q^2 + R^2 + A^2 - 2Q(R-A), \quad (46)$$

we can derive that

$$Q_n(t+1)^2 \leq Q_n(t)^2 + R_n(t)^2 + A_n(t)^2 - 2Q_n(t)(R_n(t) - A_n(t)), \forall n. \quad (47)$$

Applying to all PU slice users, we can obtain that

$$\sum_{n=1}^{N} \frac{Q_n(t+1)^2 - Q_n(t)^2}{2} \leq \sum_{n=1}^{N} \frac{R_n(t)^2 + A_n(t)^2}{2} - \sum_{n=1}^{N} Q_n(t)(R_n(t) - A_n(t)). \quad (48)$$

Similarly, for $H_n(t)$, we can derive that

$$\sum_{n=1}^{N} \frac{H_n(t+1)^2 - H_n(t)^2}{2} \leq \sum_{n=1}^{N} \frac{y_n(t)^2}{2} + \sum_{n=1}^{N} H_n(t) y_n(t). \quad (49)$$

Summing up (48) and (49) yields

$$\begin{aligned} &L(\Theta(t+1)) - L(\Theta(t)) \\ &\leq \sum_{n=1}^{N} \frac{R_n(t)^2 + A_n(t)^2}{2} + \sum_{n=1}^{N} \frac{y_n(t)^2}{2} \\ &\quad - \sum_{n=1}^{N} Q_n(t)(R_n(t) - A_n(t)) + \sum_{n=1}^{N} H_n(t) y_n(t). \end{aligned} \quad (50)$$

Then the drift-plus-penalty function can be represented as

$$\begin{aligned} &\Delta(\Theta(t)) + VE\left\{\sum_{n=1}^{N} Q_n(t) | \Theta(t)\right\} \\ &\leq B - \sum_{n=1}^{N} \left(\lambda_{pu} - Q^{up}\varepsilon\right) R_n(t) \\ &\quad - \sum_{n=1}^{N} H_n t R_n(t) - \sum_{n=1}^{N} V Q_n(t) R_n(t), \end{aligned} \quad (51)$$

where

$$\begin{aligned} B = &\sum_{n=1}^{N} \frac{\left(\lambda_{pu} - Q^{up}\varepsilon\right)^2 + R_n(t)^2}{2} + \sum_{n=1}^{N} H_k(t)\left(\lambda_{pu} - Q^{up}\varepsilon\right) \\ &+ \sum_{n=1}^{N} \frac{R_n(t)^2 + A_n(t)^2}{2} + V \sum_{n=1}^{N} Q_n(t) A_n(t). \end{aligned} \quad (52)$$

REFERENCES

[1] H. Jiang, T. Wang, and S. Wang, "Three-tier hierarchical model of dynamic spectrum sharing based on hybrid authorization using geolocation database and cognitive radio," in *Proc. IEEE ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.

[2] "Cisco visual network index: Forecast and methodology, 2016–2021," San Jose, CA, USA, Cisco, White Paper, Jun. 2017.

[3] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1923–1940, 4th Quart., 2015.

[4] S. Wang and C. Ran, "Rethinking cellular network planning and optimization," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 118–125, Apr. 2016.

[5] J. Liu, S. Zhang, N. Kato, H. Ujikawa, and K. Suzuki, "Device-to-device communications for enhancing quality of experience in software defined multi-tier LTE-A networks," *IEEE Netw.*, vol. 29, no. 4, pp. 46–52, Jul./Aug. 2015.

[6] X. Sun and S. Wang, "Resource allocation scheme for energy saving in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4407–4416, Aug. 2015.

[7] L. Zhang *et al.*, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 44–51, Oct. 2017.

[8] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2014.

[9] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Dynamic resource provisioning with stable queue control for wireless virtualized networks," in *Proc. IEEE PIMRC*, Hong Kong, Dec. 2015, pp. 1856–1860.

[10] K. Zhu and E. Hossain, "Virtualization of 5G cellular networks as a hierarchical combinatorial auction," *IEEE Trans. Mobile Comput.*, vol. 15, no. 10, pp. 2640–2654, Oct. 2016.

[11] W. Chen, X. Xu, C. Yuan, J. Liu, and X. Tao, "Virtualized radio resource pre-allocation for QoS based resource efficiency in mobile networks," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.

[12] J. Erfanian and B. Daly, "5G white paper," Frankfurt, Germany, NGMN Alliance, White Paper, Mar. 2015.

[13] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 38–45, Dec. 2017.

[14] J. Ni, X. Lin, and X. Shen, "Efficient and secure service-oriented authentication supporting network slicing for 5G-enabled IoT," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 644–657, Mar. 2018.

[15] Y. Cui, V. K. N. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems' large deviation theory, stochastic Lyapunov drift, and distributed stochastic learning," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1677–1701, Mar. 2012.

[16] T. Dang, Y. Mo, Y. Sun, and M. Peng, "Energy-efficient resource allocation in delay-aware wireless virtualized networks," in *Proc. IEEE WCSP*, Nanjing, China, Oct. 2017, pp. 1–6.

[17] T. LeAnh, N. H. Tran, D. T. Ngo, and C. S. Hong, 'Resource allocation for virtualized wireless networks with backhaul constraints," *IEEE Commun. Lett.*, vol. 58, no. 3, pp. 1677–1701, Jan. 2017.

[18] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 4, pp. 462–476, Sep. 2016.

[19] X. Zhang and Q. Zhu, "Game-theory based power and spectrum virtualization for optimizing spectrum efficiency in mobile cloud-computing wireless networks," *IEEE Trans. Cloud Comput.*, to be published, doi: 10.1109/TCC.2017.2727044.

[20] S. M. A. Kazmi, N. H. Tran, T. M. Ho, and C. S. Hong, "Hierarchical matching game for service selection and resource purchasing in wireless network virtualization," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 148–151, Jan. 2018.

[21] P. Zhang, H. Yao, and Y. Liu, "virtual network embedding based on computing, network and storage resource constraints," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2017.2726120.

[22] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Resource provisioning in wireless virtualized networks via massive-MIMO," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 237–240, Feb. 2015.

[23] P. Caballero, A. Banchs, G. Veciana, and X. Costa-Pérez, "Network slicing games: Enabling customization in multi-tenant networks," in *Proc. IEEE INFOCOM*, Atlanta, GA, USA, May 2017, pp. 1–9.

[24] V. Sciancalepore *et al.*, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE INFOCOM*, Atlanta, GA, USA, May 2017, pp. 1–9.

[25] U. Habiba and E. Hossain, "Auction mechanisms for virtualization in 5G cellular networks: Basics, trends, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2264–2293, 3rd Quart., 2018.

[26] M. Neely, *Stochastic Network Optimization with Application to Communication and Queuing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[27] A. Mukherjee, "Queue-aware dynamic on/off switching of small cells in dense heterogeneous networks," in *Proc. IEEE GLOBECOM Workshops*, Atlanta, GA, USA, Dec. 2013, pp. 182–187.

[28] Y. Li, M. Sheng, Y. Shi, X. Ma, and W. Jiao, "Energy efficiency and delay tradeoff for time-varying and interference-free wireless networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 5921–5931, Sep. 2014.

[29] J. Dai and S. Wang, "Clustering-based spectrum sharing strategy for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 228–237, Jan. 2017.

[30] S. Wang, W. Shi, and C. Wang, "Energy-efficient resource management in OFDM-based cognitive radio networks under channel uncertainty," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3092–3102, Sep. 2015.

Huijuan Jiang received the B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2017. She is currently pursuing the M.S. degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. Her research interests include wireless communications and machine learning and resource allocation in cognitive radio networks.

Tianyu Wang received the Ph.D. degree from Peking University, Beijing, China, in 2016. He is currently an Associate Researcher with the School of Electronic Science and Engineering, Nanjing University, China. He has published over 30 IEEE journal and conference papers. His current research interest includes network slicing, load balancing, and machine learning in wireless networks. He was a recipient of the Best Paper Award from the IEEE ICC'15, IEEE GLOBECOM'14, and ICST ChinaCom'12.

Shaowei Wang (S'06–M'07–SM'13) received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. In 2006, he joined the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, as a Faculty Member. From 2012 to 2013, he was a Visiting Scholar/Professor with Stanford University, Stanford, CA, USA, and the University of British Columbia, Vancouver, BC, Canada. He has published over 100 papers in leading journals and conference proceedings in his research areas. He organized the Special Issue on Enhancing Spectral Efficiency for LTE-Advanced and Beyond Cellular Networks for IEEE WIRELESS COMMUNICATIONS, and the Feature Topic on Energy-Efficient Cognitive Radio Networks for *IEEE Communications Magazine*. He is on the Editorial Board of *IEEE Communications Magazine*, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *Wireless Networks* (Springer). He serves/served on the technical or executive committees of reputable conferences, including IEEE INFOCOM, IEEE ICC, IEEE GLOBECOM, and IEEE WCNC.