Energy-Aware Scheduling for Multi-Hop Cognitive Radio Networks

Jinlin Peng, Hao Yue, Kaiping Xue, *Senior Member, IEEE*, Ying Luo, Peilin Hong, and Yuguang Fang, *Fellow, IEEE*

***Abstract*—Cognitive radio (CR) technology, which enables unlicensed secondary users to opportunistically access the unused licensed spectrum, has attracted more and more attention from both academia and industry due to its potential to significantly improve the spectrum utilization. Considering both temporal and spatial variations of spectrum availability, this paper focuses on improving the energy efficiency in CR networks by opportunistically serving the delay-tolerant data only when enough spectrum is available. Based on this idea, a stochastic optimization problem is formulated to integrate the power control, link scheduling, and routing, which minimizes the expected power consumption while guaranteeing the system stability. To obtain the solution, we use the Lyapunov optimization technique and design an online algorithm, which solves a sub-problem without future knowledge of the related stochastic models (e.g., random data arrival and spectrum supply). Besides, in view of the NP-hardness of the sub-problem, we also develop a heuristic algorithm based on branch-and-bound framework to obtain the approximate solution with low computing complexity. Theoretical analysis shows that our algorithm offers an explicit tradeoff between energy consumption and delay performance. Numerical results also confirm the effectiveness of our solutions.**

***Index Terms*—Cognitive radio, energy consumption, cross-layer scheduling, time varying spectrum supply, stochastic optimization.**

I. Introduction

THE RAPID growth in popularity of wireless devices, such as smartphones and tablets, and the surge of various mobile applications, such as online social networking and mobile gaming, have resulted in recent booming of data services. The ever growing data services directly lead to the exponential increase in data traffic and the increasing demand for spectrum resource in wireless networks. In parallel with that, recent studies [1], [2] show that even in the most crowed region of big cities (e.g., Washington, Chicago, New York City, etc.), many licensed spectrum blocks are not efficiently utilized in certain geographical areas and are idle most of the time, mainly due to the static spectrum allocation regulation of Federal Communications Commission (FCC). Such circumstances motivate FCC to open up the licensed spectrum bands and search for new innovative technologies to encourage dynamic use of the under-utilized spectrum. As one of the most promising solutions, cognitive radio (CR) technique enables unlicensed secondary users (SUs) to opportunistically access vacant licensed spectrum as long as it does not disrupt the quality of service of licensed spectrum holder, which can significantly improve the utilization of spectrum resource.

Another direct result of the ever growing data traffic is the increase of the associated energy consumption in wireless networks. For example, telecommunication data volume increases approximately by an order of 10 every 5 years, which causes an increase of the associated energy consumption by approximately 16-20 percent per annum [3]. On the one hand, many types of multi-hop wireless networks (e.g., ad hoc networks and sensor networks) are battery-powered and are thus constrained with energy at each node. On the other hand, the sharp rise in energy cost and carbon dioxide (CO_2) emission of information and communications technology (ICT) is more and more obvious. Therefore, optimizing the energy efficiency of wireless communications is increasingly urgent and important since it can not only reduce environmental impact, but also reduce overall network costs to make communication more practical and affordable.

In this paper, we focus on the energy-efficient communications in cognitive radio networks (CRNs). Specifically, we develop an energy-aware cross-layer scheduling strategy to minimize the energy consumption in CRNs. In the literature, some research works [4]–[14] have explored energy-efficient communications over traditional wireless networks. However, none of them considers the uncertain spectrum supply, which is one salient feature of CRNs. In CRNs, since SUs must evacuate licensed bands whenever primary services are active, it is much more challenging to perform power allocation, link scheduling, and routing than in traditional wireless networks. Considering the spatial uncertainty of spectrum availability in CRNs, some efforts [15]–[22] have been devoted to cross-layer design for maximizing network throughput or minimizing the usage of

Manuscript received February 17, 2016; revised August 6, 2016; accepted September 21, 2016. Date of publication October 3, 2016; date of current version December 30, 2016. This work of K. Xue and P. Hong is partially supported by the National Natural Science Foundation of China under Grant No. 61379129 and No. 61671420, and the Fundamental Research Funds for the Central Universities. The work of Y. Fang was partially supported by the U.S. National Science Foundation under Grant CNS-1343356. The associate editor coordinating the review of this paper and approving it for publication was L. DaSilva.

J. Peng is with Huawei Shanghai Research Institute, Shanghai 201206, China (e-mail: pengjinlin@huawei.com).

H. Yue is with the Department of Computer Science, San Francisco State University, San Francisco, CA 94132 USA (e-mail: haoyue@sfsu.edu).

K. Xue, Y. Luo, and P. Hong are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: kpxue@ustc.edu.cn; yingluo@mail.ustc.edu.cn; plhong@ustc.edu.cn).

Y. Fang is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: fang@ece.ufl.edu).

Digital Object Identifier 10.1109/TCCN.2016.2614838

licensed spectrum. However, none of them focus on energy efficiency in CRNs. Recently, Li *et al.* [23] have investigated the minimum energy consumption problem by exploring joint frequency allocation, link scheduling, routing, and transmission power control for multi-hop CRNs. This approach is referred to as the *traditional static real-time transmission strategy* in this paper since it collects the available spectrum and data arrival information, solves the optimization problem, and then immediately transmits all the data from the source to the destination. Note that in CRNs, the spectrum availability is time-varying due to the activity of primary users (PUs), which is ignored in [23]. Besides, in practice, most of the services, such as video streaming, are delay-tolerant. Therefore, based on the Shannon capacity theorem, we can further reduce the network energy consumption through scheduling the transmission of delay-tolerant traffic when the network nodes have more available spectrum bands. It means that the traditional static real-time transmission strategy may not be optimal any more after we consider the time dimension.

In this paper, we develop a dynamic energy-aware scheduling strategy by considering both temporal and spatial uncertainty of spectrum availability in CRNs. In view of the randomness in data arrival and spectrum supply, a stochastic optimization problem that minimizes the time-average expected energy consumption while stabilizing the network is formulated. Based on the Lyapunov optimization technique [24]–[27], we design an online algorithm to obtain the solution. Our contributions are summarized as follows:

- We study the average energy expenditure minimization problem by considering both temporal and spatial spectrum features in multi-hop CRNs. Specifically, we propose to reduce the network energy consumption through scheduling the transmission of delay-tolerant traffic when the network nodes have more available spectrum resources.
- Mathematically, a cross-layer stochastic optimization framework is formulated, which jointly considers power allocation at physical layer, link scheduling at link layer, and flow routing at network layer. Based on Lyapunov optimization technique, we develop an energy-efficient scheduling algorithm that solves a sub-problem without future spectrum supply information. Besides, in view of the NP-hardness of the sub-problem, we develop an algorithm based on branch-and-bound framework to obtain the approximate solution.
- Through theoretical analysis and simulation results, we show that our algorithm can not only stabilize the system, but also offer an explicit tradeoff between the energy consumption and delay performance. Furthermore, numerical results show that our algorithm outperforms the traditional static real-time algorithm in terms of energy efficiency.

The rest of this paper is organized as follows. The related work is reviewed in Section II. In Section III, we describe the system model in detail and mathematically formulate our problem. In Section IV, we design an energy-efficient scheduling strategy based on Lyapunov optimization technique. We describe an approximate algorithm based on branch-and-bound to the sub-problem at each time slot in Section V. We conduct simulations and evaluate the performance of our proposed algorithms in Section VI. Finally, we draw the concluding remarks in Section VII.

II. Related Work

Recently, CRs have drawn intensive attention due to the potential to significantly improve the spectrum efficiency [30]–[36]. However, one key obstacle to the deployment of CRNs lies in the uncertainty of spectrum supply. Since SUs must evacuate the licensed bands whenever primary services become active, the return of primary services has significant impact on how to perform opportunistic spectrum accessing, scheduling and interference avoidance, and routing in multi-hop CRNs. To overcome this obstacle, some efforts have been made to cross-layer optimization in CRNs by considering the uncertainty of spectrum availability. Tang *et al.* [15] studied the joint spectrum allocation and link scheduling problems with the objectives of maximizing throughput and achieving certain fairness in CRNs. Michelusi and Mitra [16] proposed a cross-layer framework to jointly optimize spectrum sensing and access in agile wireless networks, in which, sensing and access are jointly controlled to maximize the SU throughput, with constraints on PU throughput degradation and SU cost. Hou *et al.* [17] and Shi *et al.* [18] investigated the joint power control, frequency scheduling and routing problem in order to minimize the network-wide spectrum resource and presented a centralized algorithm for spectrum sharing in CRNs. In their following work, Shi and Hou [19] also provided a distributed approach to the same problem. Considering the uncertain spectrum supply, Pan *et al.* [20]–[22] modeled the vacancy of licensed bands as a series of random variables and minimized the usage of licensed spectrum to support CR sessions with rate requirements at certain confidence levels.

On the other hand, in view of the importance of energy saving, many research works have been constructed to minimize energy consumption in traditional wireless networks while meeting certain service requirements at different layers of protocol stack, e.g., energy-efficient scheduling and MAC schemes [4], [5], Energy harvesting relay and power allocation [6], and energy-efficient routing protocols [7]–[9]. All of these efforts only focus on single layer optimization. As we know, power control has a profound impact on interference among nodes and on scheduling. Moreover, power control and scheduling determine link capacities, which, in turn, affect routing. Thus, a network problem is inherently cross-layer in nature and calls for joint consideration of power control, scheduling, and routing. Based on this observation, Cruz and Santhanam [10] and Bhatia and Kodialam [11] studied the joint routing, scheduling and power control problem for power efficient communications over multi-hop wireless networks. Meanwhile, Oh *et al.* [14] proposed a distributed algorithm for BS switching on/off by considering network impact, which is an effect caused by turning off a BS. To overcome the centralized computation burden [10] and [11], Lin *et al.* [12] designed a distributed algorithm with low computational complexity. By considering dynamic spectrum

and renewable energy resource availability, Liao *et al*. [13] formulated both offline and online energy cost minimization problems and gave the corresponding control algorithms.

In the literature, some research efforts have also been devoted to energy-efficient communications in CRNs. Jiang *et al.* [28] propose several energy-efficient solutions to spectrum sensing, spectrum sharing, and secondary network deployment in non-cooperative CRNs. Li *et al.* [23] has investigated the minimum energy consumption problem with joint consideration of frequency allocation, link scheduling, routing, and transmission power control in multi-hop CRNs. Bayhan and Alagoz [29] has also studied energy efficient scheduling in centralized cognitive radio networks. Different from the existing works, we use the harvested spectrum to serve delay-tolerant traffic and reduce the network power consumption by considering both temporal and spatial variations of spectrum availability in multi-hop CRNs.

III. System Model and Problem Formulation

A. System Architecture and Motivation

The potential to significantly improve the utilization of spectrum resource has prompted a few interesting research problems on the implementation of CRs in both ad hoc networks and cellular networks [30]–[36]. Unfortunately, all of these existing works commonly assume that SUs have already been equipped with CRs, which can explore licensed spectrum bands, reconfigure RF, switch frequency across a wide range of spectrum (e.g., from 20 MHz to 2.5 GHz [37]), and exchange packets over non-contiguous spectrum bands. This may be possible in theory, but in practice currently it is very difficult to implement such high cognitive capability in lightweight radios in mobile devices such as cell phones. Besides, the cognitive functionalities may significantly increase energy consumption, which is also undesirable for energy-constrained mobile devices.

Based on the above observations, we have designed a new architecture for CRNs in [20] and [38], called cognitive capacity harvesting (CCH) networks. As shown in Fig. 1, a CCH consists of three types of entities: Second Service Provider (SSP), Secondary Users (SUs) and static Relay Stations (RSs). The SSP is an independent service provider with its own basic spectrum bands. It can also harvest/purchase some vacant licensed spectrum, such as TV white spaces, from other license holders and then allocate all the available spectrum resources to the CCH network for enhancing its services for SUs. A SU can be any traditional device with existing accessing technology (e.g., laptops or desktop computers using Wi-Fi, cell phones using GSM/GPRS, smart phones using 3G/4G/LTE, etc.) or CR device using CR technology. To facilitate the access of SUs with or without cognitive capability, the SSP deploys a collection of RSs equipped with multiple cognitive radios, which form the basic infrastructure and can tune to any basic band or harvested band for communications. If a SU has cognitive capability, it can communicate with RSs over both basic bands and harvested bands. Otherwise, they communicate over the basic bands. Some basic bands are reserved to establish the common control channels, through which important signaling information can be exchanged.

Since most of energy in CCH is consumed by the RSs, in this paper, we focus on developing a dynamic cross-layer scheduling strategy to minimize the time-average power expenditure of RSs in CCH. Considering the time-varying spectrum supply feature and the delay-tolerant traffic, the *traditional static real-time transmission strategy* (refer to Section I, i.e., collect the available spectrum and data arrival information, then immediately transmit all the arrival data from source to destination with the minimum energy consumption) may not be optimal, which motivates us to design a dynamic scheduling strategy. Here, we give a simple example to show the superiority of dynamic scheduling strategy to traditional static real-time transmission strategy. As illustrated by Fig. 2, we consider a time-slotted system where node A transmits data to node B. To simplify the analysis, we only focus on two slots. Assume that the data arrives at time slot t_1 and slot t_2 are 3 units and 1 unit, respectively. The number of harvested spectrum bands, whose bandwidth is equal to 1 unit, are 1 and 3 at slot t_1 and slot t_2, respectively. For illustration purpose, we use Shannon capacity $C = W\log(1+\frac{hP}{\eta W})$, where W is the bandwidth, h is the channel state, P is the allocated power and η is the noise power density. Without loss of generality, we assume $h = 1$ and $\eta = 1$. Under the traditional static real-time transmission strategy, node A will transmit 3 units data to node B over 1 available spectrum band at time slot t_1, thus the power consumption is $P_1 = 2^{3/1} - 1 = 7$ units. Similarly, node A will transmit 1 unit data to node B with 3 available spectrum bands at time slot t_2, thus the power consumption is $P_2 = 3 \times (2^{1/3} - 1) = 0.7798$ units. We can easily obtain the average power consumption of the traditional static real-time strategy is 3.8899 units. However, as a comparison, if node A does not transmit any data at time slot t_1 and transmits all the data at time slot t_2, we have $P_1 = 0$ unit and $P_2 = 3 \times (2^{4/3} - 1) = 4.5595$ units. In this case, the average power consumption is 2.2797 units, which is much lower than that for the traditional static real-time strategy.

B. Mathematical Modeling

In this subsection, we present our mathematical model for a more general case. Consider a typical CCH network consisting of an SSP, a group of RSs and SUs, and a set of spectrum bands (including basic bands and harvested bands) as shown in Fig. 1 [20], [38]. Let $\mathcal{N}$ and $\mathcal{B}$ denote the set of RSs and the set of spectrum bands in the network, respectively. We assume that the total number of RSs is $|\mathcal{N}| = N$ and the bandwidth of each frequency band $m \in \mathcal{B}$ is W_m.

The entire system operates in a time-slotted fashion. To avoid causing serious interference to any PU, all RSs adopt the overlay spectrum sharing approach, which leads to the temporal and spatial uncertainty of the spectrum availability in CCH. At time slot t, each RS i senses the spectrum and finds a set of available frequency bands $\mathcal{M}_i(t) \subseteq \mathcal{B}$. The band availability state $\mathcal{M}_i(t)$ evolves over time according to the activity of PUs, thus may not be the same at different RSs and at different

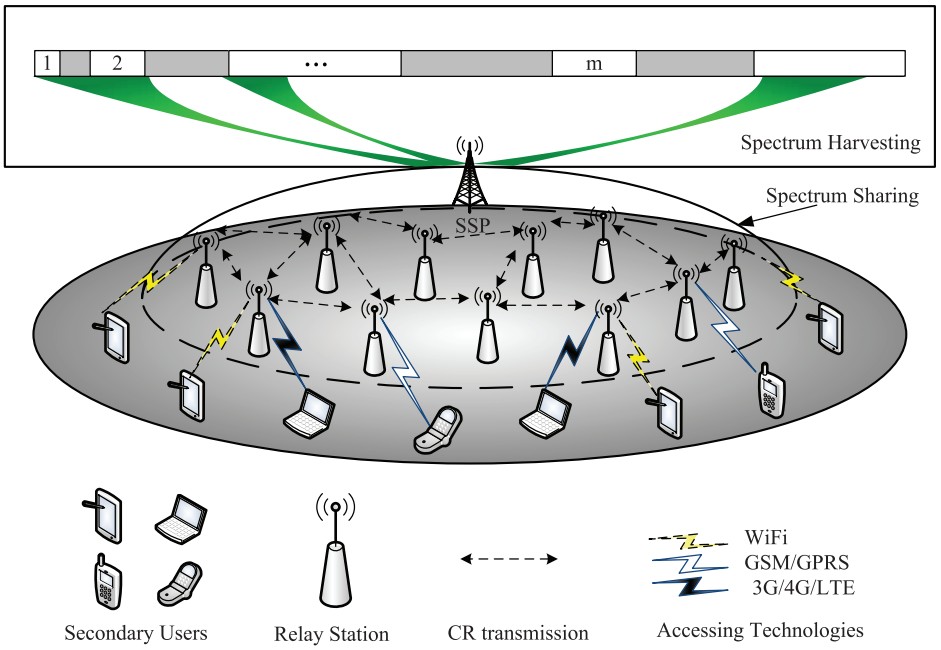
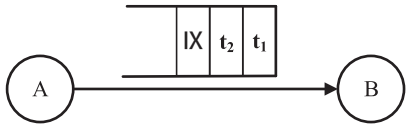


Fig. 1. System architecture of CCH.



	Harvested spectrum band	Data arrival	Traditional static real-time transmission strategy		Dynamic scheduling strategy	
			Power consumption	Avg. power consumption	Power consumption	Avg. power consumption
t_1	1	3	$2^{3/1}$-1=7	(7+0.7798)/2=3.8899	0	(0+4.5595)/2=2.2797
t_2	3	1	$3*(2^{3/1}-1)=0.7798$		$3*(2^{4/3}-1)=4.5595$	

Fig. 2. The single link communication scenario.

time slots. This is the biggest different characteristic of cognitive multi-hop wireless networks from traditional multi-hop networks. It is the time-varying spectrum supply feature that provides the dynamic scheduling strategy with energy saving possibility. It should be noted that better sensing techniques employed, the smaller interference that can be potentially generated to PUs. We assume that sensing techniques of deployed static RSs are good enough in this paper.

We consider the following power propagation model [38]:

$$P_r = \gamma \cdot d^{-n} \cdot P_t, \tag{1}$$

where P_t is the transmission power at the sender, P_r is the received power at the receiver, γ is an antenna related constant, n is the path loss factor, and d is the distance between the sender and receiver. We assume the data transmission over a link is successful if and only if the received power exceeds the sensitivity α_1. Then according to (1), the maximum transmission range $R_T^{\max} = (\gamma P_{max}/\alpha_1)^{1/n}$ for a RS when it uses the maximum transmission power P_{max}. Besides, we use the primary interference model, that is, the interference is intolerable if and only if it exceeds a toleration threshold, say α_2 at a receiver. Similarly, we have the maximum interference range $R_I^{\max} = (\gamma P_{max}/\alpha_2)^{1/n}$. Obviously, $\alpha_2 < \alpha_1$.

Transmission Constraints. Based on the definition of transmission range, the set of RSs that can transmit to RS i on band $m \in \mathcal{M}_i(t)$ at time slot t with maximum transmission power is

$$\mathcal{T}_i^m(t) = \left\{ j | d_{ij} \le R_T^{\max}, j \in \mathcal{N} \backslash \{i\}, m \in \mathcal{M}_j(t) \right\}, \tag{2}$$

where d_{ij} is the distance between RSs i and j. Note that $\mathcal{T}_i^m(t)$ is also the set of RSs to which RS i can transmit. Similarly, based on the definition of interference range, the set of RSs that can interfere with RS i on band $m \in \mathcal{M}_i(t)$ at time slot t with maximum transmission power is

$$\mathcal{I}_i^m(t) = \left\{ j | d_{ij} \le R_I^{\max}, j \in \mathcal{N}, m \in \mathcal{M}_j(t) \right\}. \tag{3}$$

Note that $\mathcal{I}_i^m(t)$ is also the set of RSs with which RS i can interfere.

We denote

$$x_{ij}^m(t) = \begin{cases} 1 & \text{If RS } i \text{ transmits data to RS } j \\ & \text{on band } m \text{ at time slot } t, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Since one RS i can use a band $m \in \mathcal{M}_i(t)$ for transmission to only one RS at any time, we have

$$\sum_{j \in \mathcal{T}_i^m(t)} x_{ij}^m(t) \le 1. \tag{5}$$

In order to achieve successful transmission from RS i to RS j on band m, the following power allocation constraints must be satisfied for the transmission link $i \to j$ and interfering link $k \to h$, where $k \in \mathcal{I}_j^m(t), k \ne i$, and $h \in \mathcal{T}_k^m(t)$. If $x_{ij}^m(t) = 1$, we have $p_{ij}^m(t) \in [(\frac{d_{ij}}{R_T^{\max}})^n P_{max}, P_{max}]$ and $p_{kh}^m(t) \le (\frac{d_{kj}}{R_I^{\max}})^n P_{max}$, where $(\frac{d_{ij}}{R_T^{\max}})^n P_{max}$ is the minimum transmission power to ensure the received power exceeds α_1,

and $(\frac{d_{kj}}{R_I^{\max}})^n P_{max}$ is the maximum transmission power to avoid interrupting transmission from RS i to RS j, which can be derived from (1); Otherwise, $p_{ij}^m(t) = 0$ and $p_{kh}^m(t) \le P_{max}$. These constraints can be re-written as:

$$\left(\frac{d_{ij}}{R_T^{\max}}\right)^n P_{max} x_{ij}^m(t) \le p_{ij}^m(t) \le P_{max} x_{ij}^m(t), \tag{6}$$

$$p_{kh}^m(t) \le P_{max} - \left[1 - \left(\frac{d_{kj}}{R_I^{\max}}\right)^n\right] P_{max} x_{ij}^m(t). \tag{7}$$

In addition, for successful scheduling and transmission, the following two constraints must also hold: 1) RS i cannot transmit and receive simultaneously over the same band $m \in \mathcal{M}_i(t)$; 2) Similar to (5), RS i cannot use the same band $m \in \mathcal{M}_i(t)$ to receive data from two different RSs at the same time. Interestingly, it turns out that the above two constraints are mathematically embedded in (6) and (7) as shown in [18].

Flow Routing and Link Capacity Constraints. We define the arrival processes in terms of the source and destination of data flow: $A_i^c(t)$ is the amount of data exogenously arriving at RS i and destined for RS c at time slot t. For simplicity of explanation, we assume the arrival matrix $\mathbf{A}(t)$ ($N \times N$ matrix, its element is $A_i^c(t)$) are i.i.d. over time with arrival rate $\mathbb{E}\{\mathbf{A}(t)\} = \boldsymbol{\lambda}$. ($\boldsymbol{\lambda}$ is $N \times N$ matrix, its element is $\lambda_i^c = \mathbb{E}\{A_i^c(t)\}$.) All data transmitted from any source to a particular RS $c \in \mathcal{N}$ is defined as commodity (or data item) c. Data is stored at each RS according to its destination, and we let $U_i^c(t)$ represent the current backlog of commodity c in RS i. The resulting 1-step queuing equation for backlog $U_i^c(t)$ satisfies (for $c \ne i$)

$$U_i^c(t+1) = \max\left\{U_i^c(t) - \sum_{j \in \bigcup_{m \in \mathcal{M}_i(t)} \mathcal{T}_i^m(t)} u_{ij}^c(t), 0\right\} + \sum_{j \in \bigcup_{m \in \mathcal{M}_i(t)} \mathcal{T}_i^m(t)} u_{ji}^c(t) + A_i^c(t), \tag{8}$$

where $u_{ij}^c(t)$ is the commodity c routed from RS i to RS j at time slot t. It must satisfy the link rate constraint which depends on the available bandwidth $\overrightarrow{\mathcal{M}}(t)$ ($1 \times N$ vector, its element is $\mathcal{M}_i(t)$) and power allocation scheme $\mathbf{p}(t)$ ($N \times N \times |\mathcal{B}|$ matrix, its element is $p_{ij}^m(t)$) as follows:

$$\sum_{c \in \mathcal{N}} u_{ij}^c(t) \le u_{ij}\left(\overrightarrow{\mathcal{M}}(t), \mathbf{p}(t)\right) = \sum_{m \in \mathcal{M}_i(t) \cap \mathcal{M}_j(t)} W_m \log\left(1 + \frac{\gamma d_{ij}^{-n} p_{ij}^m(t)}{\eta W_m}\right), \tag{9}$$

where η is the noise power density. Besides, the commodity c transmitted out from RS i cannot exceed its backlog

$$\sum_{j \in \bigcup_{m \in \mathcal{M}_i(t)} \mathcal{T}_i^m(t)} u_{ij}^c(t) \le U_i^c(t). \tag{10}$$

To simplify the expression, we use $I_i^c(\mathbf{u}(t))$ and $O_i^c(\mathbf{u}(t))$ to represent the commodity c routed to and from RS i, which can be expressed as:

$$I_i^c(\mathbf{u}(t)) = \sum_{j \in \bigcup_{m \in \mathcal{M}_i(t)} \mathcal{T}_i^m(t)} u_{ji}^c(t),$$
$$O_i^c(\mathbf{u}(t)) = \sum_{j \in \bigcup_{m \in \mathcal{M}_i(t)} \mathcal{T}_i^m(t)} u_{ij}^c(t), \tag{11}$$

where $\mathbf{u}(t)$ is an $N \times N \times N$ matrix and its element is $u_{ij}^c(t)$. Thus, Eq. (8) and Eq. (10) can be simplified as (12) and (13), respectively:

$$U_i^c(t+1) = \max\left\{U_i^c(t) - O_i^c(\mathbf{u}(t)), 0\right\} + I_i^c(\mathbf{u}(t)) + A_i^c(t). \tag{12}$$

$$O_i^c(\mathbf{u}(t)) \le U_i^c(t). \tag{13}$$

C. Problem Formulation

In view of the randomness in data arrival and spectrum availability, we aim to develop a dynamic scheduling strategy to minimize the average power consumption while stabilizing the queues of the system, by jointly considering power control $\mathbf{p}(t)$, link scheduling $\mathbf{x}(t)$ ($N \times N \times |\mathcal{B}|$ matrix, its element is $x_{ij}^m(t)$), and flow routing $\mathbf{u}(t)$. The average power consumption can be expressed as $\lim_{T\to\infty}(1/T)\sum_{t=0}^{T-1}\mathbb{E}\{f(\mathbf{p}(t))\}$, where the expectation is taken over the potential randomness of data arrival, spectrum availability as well as scheduling decision. Here,

$$f(\mathbf{p}(t)) = \sum_{i \in \mathcal{N}} \sum_{m \in \mathcal{M}_i(t)} \sum_{j \in \mathcal{T}_i^m(t)} p_{ij}^m(\mathrm{t}) \tag{14}$$

is the total power consumption at time slot t.

To stabilize the system, we need to ensure the average system backlog is finite [24], i.e.,

$$\overline{\sum_{i,c} U_i^c} \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i \in \mathcal{N}} \sum_{c \in \mathcal{N}} \mathbb{E}\left\{U_i^c(t)\right\} < \infty. \tag{15}$$

Thus, our problem can be formulated as follows:

$$\underset{\mathbf{p}(\mathbf{t}),\mathbf{x}(\mathbf{t}),\mathbf{u}(\mathbf{t})}{\text{minimize}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{f(\mathbf{p}(\mathbf{t}))\} \tag{16}$$

subject to (4), (5), (6), (7), (9), (12), (13) and (15), which is denoted as problem 0.

The major challenge to solve the above stochastic problem lies in that future data arrival and available spectrum band are uncertain. To overcome this challenge, we design a dynamic scheduling algorithm based on the Lyapunov optimization technique [24] in next section.

IV. Optimal Scheduling Algorithm

A. Lyapunov Drift Analysis

We first establish the Lyapunov drift technique to ensure the stability and performance optimization to be achieved simultaneously. Let $\mathbf{U}(t)$ ($N \times N$ matrix, its element is $U_i^c(t)$) be the process of queue backlogs that evolves according to a

certain probability distribution. To measure aggregated network congestion, we define a Lyapunov function as $L(\mathbf{U}(t)) = \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}}[U_i^c(t)]^2$. In addition, we define the one-step conditional Lyapunov drift $\Delta(\mathbf{U}(t))$ as:

$$\Delta(\mathbf{U}(t)) = \mathbb{E}\{L(\mathbf{U}(t+1)) - L(\mathbf{U}(t))|\mathbf{U}(t)\}, \tag{17}$$

where the expectation is taken over the potential randomness of the available bandwidth $\overrightarrow{\mathcal{M}}(t)$ and scheduling decision during time slot t, given the current backlog matrix $\mathbf{U}(t)$. The conditional Lyapunov drift has two features as shown in Lemma 1 and Lemma 2, which can guide us to develop our scheduling policy and analyze the system performance. The results can be obtained following similar techniques in [24].

Lemma 1: The conditional Lyapunov drift under any policy satisfies:

$$\begin{aligned}\Delta(\mathbf{U}(t)) + V\mathbb{E}\{f(\mathbf{p}(t))|\mathbf{U}(t)\} &\le CN^2 + 2\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} U_i^c(t)\lambda_i^c \\ &+ \mathbb{E}\Bigg\{V f(\mathbf{p}(t)) - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} \\ &\times\ 2U_i^c(t)\big(O_i^c(\mathbf{u}(t)) - I_i^c(\mathbf{u}(t))\big)\big|\mathbf{U}(t)\Bigg\},\end{aligned} \tag{18}$$

where

$$\begin{aligned} C &\triangleq \left(A_{\max} + \mu_{\max}^{in}\right)^2 + \left(\mu_{\max}^{out}\right)^2, \\ A_{\max} &\triangleq \max_{i,c,t}\mathbb{E}\left\{A_i^c(t)\right\}, \\ \mu_{\max}^{in} &\triangleq \max_{\{i,c,t,\mathbf{p}(t),\mathbf{x}(t),\overrightarrow{\mathcal{M}}(t)\}} I_i^c(\mathbf{u}(t)), \\ \mu_{\max}^{out} &\triangleq \max_{\{i,c,t,\mathbf{p}(t),\mathbf{x}(t),\overrightarrow{\mathcal{M}}(t)\}} O_i^c(\mathbf{u}(t)). \end{aligned} \tag{19}$$

Proof: By squaring both sides of (8), we can obtain the following results:

$$\begin{aligned}\left[U_i^c(t+1)\right]^2 &\le \left[U_i^c(t) - O_i^c(\mathbf{u}(t))\right]^2 + \left[I_i^c(\mathbf{u}(t)) + A_i^c(t)\right]^2 \\ &+ 2U_i^c(t)\left[I_i^c(\mathbf{u}(t)) + A_i^c(t)\right] \\ &= \left[U_i^c(t)\right]^2 + \left[O_i^c(\mathbf{u}(t))\right]^2 - 2U_i^c(t)O_i^c(\mathbf{u}(\mathbf{t})) \\ &+ \left[I_i^c(\mathbf{u}(t)) + A_i^c(t)\right]^2 \\ &+ 2U_i^c(t)\left[I_i^c(\mathbf{u}(t)) + A_i^c(t)\right] \\ &\le \left[U_i^c(t)\right]^2 + C + 2U_i^c(t)A_i^c(t) \\ &- 2U_i^c(t)\left[O_i^c(\mathbf{u}(t)) - I_i^c(\mathbf{u}(t))\right].\end{aligned} \tag{20}$$

Given the current backlog matrix $\mathbf{U}(\mathbf{t})$, taking an expectation of (20) and summing over $c, i \in \mathcal{N}$, we derive the following results:

$$\begin{aligned}\Delta(\mathbf{U}(t)) &\le CN^2 + 2\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} U_i^c(t)\lambda_i^c \\ &- \sum_{i\in\mathcal{N}}\mathbb{E}\Bigg\{\sum_{c\in\mathcal{N}} 2U_i^c(t)\big(O_i^c(\mathbf{u}(t)) - I_i^c(\mathbf{u}(t))\big)|\mathbf{U}(t)\Bigg\}.\end{aligned} \tag{21}$$

After adding $V\mathbb{E}\{f(\mathbf{p}(t))|\mathbf{U}(t)\}$ to both sides of the inequality, we prove Lemma 1. ■

Lemma 2: (Lyapunov Drift with Performance Optimization) If there exist positive constants V, B, ε, f^*, such that for all time slots t and all matrices $\mathbf{U}(t)$, the one-step conditional Lyapunov drift satisfies

$$\Delta(\mathbf{U}(t)) + V\mathbb{E}\{f(\mathbf{p}(t))|\mathbf{U}(t)\} \le B - 2\varepsilon\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} U_i^c(t) + Vf^*, \tag{22}$$

then the system is stable and the time average backlog satisfies

$$\overline{\sum_{i,c} U_i^c} \triangleq \limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}}\mathbb{E}\left\{U_i^c(t)\right\} \le \frac{B + Vf^*}{2\varepsilon}, \tag{23}$$

while the time average power consumption satisfies

$$\overline{f(\mathbf{p}(t))} \triangleq \limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{f(\mathbf{p}(t))\} \le f^* + B/V. \tag{24}$$

The proof is similar to that in [24] which is omitted here for brevity.

B. Energy-Efficient Scheduling Algorithm

Based on Lemma 1, we now develop a practical scheduling algorithm that stabilizes the system and consumes an average power consumption that is arbitrarily close to the minimum value f^*, (i.e., *Energy-Efficient Scheduling Algorithm (EESA)*): At every time slot, each RS observes the current level of queue backlog $\mathbf{U}(\mathbf{t})$, senses available spectrum bands $\overrightarrow{\mathcal{M}}(t)$, and submits them to the SSP over the common control channels. After that, the SSP determines the link scheduling matrix $\mathbf{x}(t)$, power matrix $\mathbf{p}(t)$ and flow routing matrix $\mathbf{u}(t)$ to minimize the right hand side of inequality (18) under the constraints (4), (5), (6), (7), (9), (12) and (13). The corresponding optimization problem can be expressed as:

$$\underset{\mathbf{p}(t),\mathbf{x}(t),\mathbf{u}(t)}{\text{minimize}} \quad Vf(\mathbf{p}(t)) - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\big(O_i^c(\mathbf{u}(t)) - I_i^c(\mathbf{u}(t))\big) \tag{25}$$

subject to (4), (5), (6), (7), (9), (12) and (13), which is denoted as problem 1.

In problem 1, V represents an arbitrary positive control parameter. Here, V is a parameter like Lagrange multiplier and affects a tradeoff between energy consumption and average queueing delay as shown in next subsection. After problem 1 is solved, the backlogs can be updated as Eq. (8). From problem 1, it can be observed that EESA does not need any knowledge of future data arrival and available spectrum information.

C. Performance Analysis of EESA

In this subsection, we analyze the performance of EESA based on the features of conditional Lyapunov drfit in Lemma 1 and Lemma 2. Before that, we first give the following definition for network capacity region.

Definition 1: The network capacity region Λ is the closure of the set of all rate matrices $\boldsymbol{\lambda}$ that can be stably supported over the network, considering all possible algorithms (possibly those with full knowledge events).

Theorem 1: If the rate matrix $\boldsymbol{\lambda}$ is interior to the network capacity region Λ, and the data arrivals as well as available band states are i.i.d. over time slots, then EESA ($\mathbf{p}^{\mathbf{EESA}}(t)$, $\mathbf{u}^{\mathbf{EESA}}(t)$) stabilizes the network and yields a time average congestion bound as

$$\overline{\sum_{i,c} U_i^{cEESA}} \le \frac{CN^2 + Vf(\mathbf{P_{max}})}{2\varepsilon_{\max}}, \tag{26}$$

where $\mathbf{P_{max}}$ is a constant matrix (each element is P_{max}) and $\varepsilon_{\max}$ is the largest ε such that $\boldsymbol{\lambda} + \varepsilon\mathbf{1} \in \Lambda$ ($\mathbf{1}$ is a $N \times N$ matrix, its element is 1_i^c being an indicator function equal to 0 if $U_i^c(t) \triangleq 0$, and 1 otherwise). Further, the time average power consumption satisfies

$$\begin{aligned}&\overline{f\left(\mathbf{p}^{\mathbf{EESA}}(t)\right)}\\ &\quad\triangleq \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{N}}\mathbb{E}\left\{\sum_{m\in\mathcal{M}_i(t)}\sum_{j\in\mathcal{T}_i^m} p_{ij}^{mEESA}(t)\right\}\\ &\quad\le f^* + CN^2/V.\end{aligned} \tag{27}$$

Proof: In [24], it is shown that if $\boldsymbol{\lambda}$ is interior to the network capacity region Λ, i.e., there is an ε such that $\boldsymbol{\lambda}+\varepsilon\mathbf{1} \in \Lambda$, and the data arrivals as well as available band states are i.i.d. over time slots, a stationary randomized policy (STAT) that takes scheduling decisions $\mathbf{p}^{\mathbf{STAT}}(t)$, $\mathbf{x}^{\mathbf{STAT}}(t)$ and $\mathbf{u}^{\mathbf{STAT}}(t)$ can be developed to satisfy $\mathbb{E}\{u_{ij}^{cSTAT}(t)\} = f_{ij}^c$, where the variables f_{ij}^c are the flows such that

$$\sum_{j\in\bigcup_{m\in\mathcal{M}_i(t)}\mathcal{T}_i^m} f_{ij}^c - \sum_{j\in\bigcup_{m\in\mathcal{M}_i(t)}\mathcal{T}_i^m} f_{ji}^c = \lambda_i^c + \varepsilon. \tag{28}$$

Besides, the stationary policy also satisfies

$$\mathbb{E}\left\{f(\mathbf{p}^{\mathbf{STAT}}(t))\right\} = f^*(\varepsilon), \tag{29}$$

where $f^*(\varepsilon)$ is the minimum cost for stabilizing rates $\boldsymbol{\lambda} + \varepsilon\mathbf{1}$ and it satisfies $f^*(\varepsilon) \to f^*$ as $\varepsilon \to 0$.

From (28), under this stationary policy we have

$$\begin{aligned}&\sum_{i\in\mathcal{N}}\mathbb{E}\left\{\sum_{c\in\mathcal{N}} 2U_i^c(t)\left(O_i^c\left(\mathbf{u}^{\mathbf{STAT}}(t)\right) - I_i^c\left(\mathbf{u}^{\mathbf{STAT}}(t)\right)\right)\middle|\mathbf{U}(t)\right\}\\ &\qquad = \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\left(\lambda_i^c + \varepsilon\right).\end{aligned} \tag{30}$$

Based on Lemma 1 (step a), and the fact that EESA achieves a smaller object value than any policy including STAT in problem 1 (step b), we plug (29) as well as (30) (step c) and obtain

$$\begin{aligned}&\Delta(\mathbf{U}(t)) + V\mathbb{E}\left\{f\left(\mathbf{p}^{\mathbf{EESA}}(t)\right)\middle|\mathbf{U}(t)\right\}\\ &\overset{(a)}{\le} CN^2 + 2\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} U_i^c(t)\lambda_i^c + \mathbb{E}\left\{Vf\left(\mathbf{p}^{\mathbf{EESA}}(t)\right)\right.\\ &\quad\left. - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\left(O_i^c\left(\mathbf{u}^{\mathbf{EESA}}(t)\right) - I_i^c\left(\mathbf{u}^{\mathbf{EESA}}(t)\right)\right)\middle|\mathbf{U}(t)\right\}\\ &\overset{(b)}{\le} CN^2 + 2\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} U_i^c(t)\lambda_i^c + \mathbb{E}\left\{Vf\left(\mathbf{p}^{\mathbf{STAT}}(t)\right)\right.\\ &\quad\left. - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\left(O_i^c\left(\mathbf{u}^{\mathbf{STAT}}(t)\right) - I_i^c\left(\mathbf{u}^{\mathbf{STAT}}(t)\right)\right)\middle|\mathbf{U}(t)\right\}\\ &\overset{(c)}{\le} CN^2 - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\varepsilon + Vf^*(\varepsilon).\end{aligned} \tag{31}$$

Thus, from Lemma 2, we attain the following results from (31):

$$\overline{\sum_{i,c} U_i^{cEESA}} \le \frac{CN^2 + Vf^*(\varepsilon)}{2\varepsilon}, \tag{32}$$

$$\overline{f\left(\mathbf{p}^{\mathbf{EESA}}(t)\right)} \le f^*(\varepsilon) + CN^2/V. \tag{33}$$

The performance bounds in (32) and (33) hold for any value $\varepsilon > 0$ such that $\boldsymbol{\lambda} + \varepsilon\mathbf{1} \in \Lambda$. The particular choice of ε only affects the bound calculation and does not affect the EESA policy or change any sample path of system dynamics. We can thus optimize the bounds in (32) and (33) separately over all possible ε values. The bound in (33) is clearly minimized by taking a limit as $\varepsilon \to 0$, yielding

$$\overline{f\left(\mathbf{p}^{\mathbf{EESA}}(\mathbf{t})\right)} \le f^* + CN^2/V. \tag{34}$$

Conversely, the bound in (32) is minimized by considering the largest feasible $\varepsilon > 0$ such that $\boldsymbol{\lambda} + \varepsilon\mathbf{1} \in \Lambda$ (defined as ε_{max}), yielding

$$\overline{\sum_{i,c} U_i^{cEESA}} \le \frac{CN^2 + Vf(\mathbf{P_{max}})}{2\varepsilon_{\max}}. \tag{35}$$

This proves Theorem 1. ■

Theorem 1 shows that in addition to stabilizing the system, EESA can achieve the average power consumption deviated no more than $O(1/V)$ from the optimal result where V is a control parameter. Furthermore, (26) reveals that with a larger V value, EESA will cause a longer queue length and thus suffer larger delay. However, Inequality (27) shows that a larger V value can narrow the gap between average power consumption performance and the optimal value. Thus, V has an influence on the tradeoff between the delay and energy consumption performance.

V. Approximate Algorithm Based on Branch-and-Bound for Problem 1

The EESA proposed in the previous section involves solving a sub-problem at every time slot, i.e., problem 1. Problem 1 is in the form of *mixed-integer non-linear program* (MINLP), which is NP-hard in general [39]. In this section, we develop an approximate algorithm based on the branch-and-bound framework [40] (ABB), to solve this problem with low computing complexity. Moreover, we analyze the effect of implementation of ABB at each time slot on the system performance. In order to simplify the presentation, we omit the script t in the rest of this section.

A. Algorithm Based on Branch-and-Bound

The challenges to solve problem 1 lie in the binary variables $\mathbf{x}$ and the non-linear constraint (9), i.e., the log term. As far as we know, the so-called branch-and-bound framework [40], [41] is commonly used to solve this kind of problem [17]. The core idea of branch-and-bound are to progressively reduce the searching space to finally determine the values for binary variables and to reduce the computing complexity by relaxing the non-linear term. Under this framework with some problem specific components (e.g., linear relaxation technique, local search algorithm and branching as we will show), we develop an algorithm ABB to provide a θ-optimal solution, which is formally defined as.

Definition 2: For a minimization problem, denote S^* the objective value of the optimal solution. A feasible solution with objective value S is called a θ-optimal solution if we have $S^* \geq S - \theta$, where $\theta \geq 0$.

Note that the above definition is different from that in [17] which is in the form of $S^* \geq \theta S$. In the prior work, the objective value is obviously positive, thus they can define the θ-optimal solution as $S^* \geq \theta S$. With $\theta < 1$, they can ensure the existence of θ-optimal solution. However, using the definition $S^* \geq \theta S$, it is hard to set the value of θ for our problem since the objective value may be negative or positive. For example, when the feasible solution and the optimal solution are positive, we need to set $\theta \leq 1$ to ensure that there exists a θ-optimal solution. Similarly, if the feasible solution and the optimal solution are negative, then we need to set $\theta \geq 1$ to ensure the existence of θ-optimal solution. For problem 1, it is difficult to know whether the solution is negative or positive, thus it is hard to set the θ value. With Definition 2 which is more general, we do not have the above problem and we can ensure the existence of θ-optimal solution only if $\theta \geq 0$. Moreover, we can set $\theta = (1-\theta_1)S$ to achieve $S^* \geq \theta_1 S$ as that in prior work.

Algorithm 1 shows the pseudocode of ABB. Initially, for problem 1, we set its lower bound $LB = -\infty$ and upper bound $UB = \infty$. Using ***linear relaxation technique***, we obtain a relaxation problem for problem 1 which can be solved in polynomial time. The solution of this relaxation problem, which may not be feasible, provides a lower bound LB_1 for problem 1. Then, based on this relaxation solution, we use a ***local search algorithm*** to find a feasible solution, which provides an upper bound UB_1 for problem 1. We update $UB = UB_1$ and $LB = LB_1$. If $LB \geq UB - \theta$, we can easily conclude that UB is a θ-optimal solution, i.e., we can stop with UB. Otherwise, the θ-optimal solution has not been found and we should narrow the gap between UB and LB with a tighter relaxation. We use ***branching*** to reduce the feasible space of problem 1 by fixing some $\mathbf{x}$ variables, which are called as partition variables. In particular, we select a partition variable x_{ij}^m based on the solution of the relaxation problem and fix $x_{ij}^m = 1$ and 0 to split the whole feasible space into two parts. After that, problem 1 is partitioned into two new sub-problems (denoted as problem 2 and problem 3) with different x_{ij}^m values. Again, we perform ***linear relaxation technique*** to get lower bounds LB_2 and LB_3 for problems 2 and 3, respectively.

Algorithm 1 Algorithm Based on Branch-and-Bound

1: Initialize the solution $\psi_\theta = \infty$, upper bound $UB = \infty$, lower bound $LB = -\infty$, and a problem list $List = \emptyset$.
2: Build a relaxation problem for problem 1 and obtain its solution $\hat{\psi}_1$ with object value LB_1 using *linear relaxation technique*.
3: Insert problem 1 into *List*.
4: **while** $List \neq \emptyset$ **do**
5: In *List*, select a problem z that has the minimum lower bound (denoted as LB_z). Remove problem z from *List*.
6: Update $LB = LB_z$. If $LB \geq UB - \theta$, stop with ψ_θ.
7: Get a feasible solution ψ_z with object value UB_z from $\hat{\psi}_z$ via *local search algorithm*.
8: **if** $UB_z < UB$ **then**
9: Update $\psi_\theta = \psi_z$ and $UB = UB_z$. If $LB \geq UB - \theta$, stop with ψ_θ.
10: **end if**
11: *Branching*: select a variable x_{ij}^m with the largest relaxation error, and build two new sub-problems $z1$ and $z2$ by setting $x_{ij}^m = 1$ and $x_{ij}^m = 0$ respectively.
12: Obtain LB_{z1}, $\hat{\psi}_{z1}$ and LB_{z2}, $\hat{\psi}_{z2}$ for sub-problems $z1$ and $z2$ via *linear relaxation technique*, then insert sub-problems $z1$ and $z2$ into *List*.
13: *Cut Branches:* remove all the problem z' with $LB_{z'} \geq UB - \theta$ from *List*.
14: **end while**

Obviously, the relaxations in problems 2 and 3 are both tighter than that in problem 1 due to the fixed x_{ij}^m, which leads to $\min\{LB_2, LB_3\} \geq LB_1$. Thus, we can update LB from LB_1 to $\min\{LB_2, LB_3\}$. We also perform the ***local search algorithm*** to find feasible solutions that provide upper bounds UB_2 and UB_3 for problems 2 and 3, respectively. If $\min\{UB_2, UB_3\} \leq UB_1$, we update UB from UB_1 to $\min\{UB_2, UB_3\}$. As a result, we now have a smaller gap between LB and UB. We test again and if $LB \geq UB - \theta$, we stop with UB. Otherwise, we choose a problem and perform ***branching*** in the same way. This process is repeated till we find the θ-optimal solution. During this process, we ***cut branches*** which do not affect the final result to accelerate traversal. It should be noted that we can set $\theta = (1-\theta_1)UB$ to achieve $LB \geq \theta_1 UB$ as that in prior work. There are four key problem specific components in ABB, which are described and designed as follows:

1) Linear Relaxation Technique. The goal of this linear relaxation technique (see line 2 in Algorithm 1) is to cut down the complexity of our original problem, which comes from the binary variables $\mathbf{x}$ in (4) and the non-linear constraint in (11). For the binary constraint (4), we relax the binary requirement and replace it with $0 \leq x_{ij}^m \leq 1$. For the log term in (9), similarly to [19], we introduce a variable $c_{ij}^m = \log(1 + \gamma d_{ij}^{-n} p_{ij}^m)$ and try to get a linear relaxation for $y = \log x$ over $x_L \leq x \leq x_U$ which is the interval we are concerned with. As shown in Fig. 3, the function $y = \log x$ can be bounded by four segments, i.e., segments I, II, III and IV. Segments I, II, and III are tangential lines at

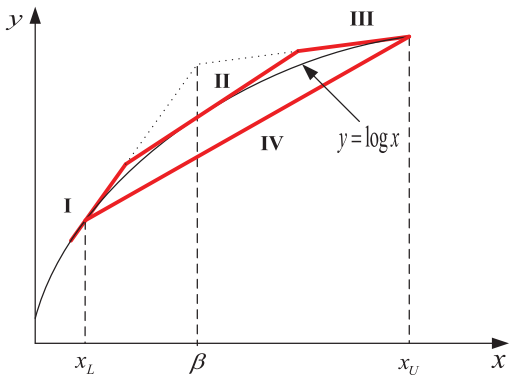


Fig. 3. The linear relaxation for function $y = \log x$.

$(x_L, \log x_L)$, $(\beta, \log \beta)$, and $(x_U, \log x_U)$ respectively, where $\beta = \frac{x_L \cdot x_U \cdot (\log x_U - \log x_L)}{(x_U - x_L)\ln 2}$ is the horizontal location of the point intersecting extended tangent segments I and III. Segment IV is the chord that joins points $(x_L, \log x_L)$ and $(x_U, \log x_U)$. These segments can be described by the following four linear constraints:

$$\begin{aligned}
&\frac{x_L}{\ln 2} y - x \le \frac{x_L}{\ln 2}(\log x_L - 1),\\
&\frac{\beta}{\ln 2} y - x \le \frac{\beta}{\ln 2}(\log \beta - 1),\\
&\frac{x_U}{\ln 2} y - x \le \frac{x_U}{\ln 2}(\log x_U - 1),\\
&(x_U - x_L) y + (\log x_L - \log x_U) x \ge x_U \log x_L - x_L \log x_U.
\end{aligned} \tag{36}$$

As a result, the log term has been relaxed into linear constraints. Now, we can relax our original problem from MINLP to a linear program (LP), which can be solved in polynomial time and provides a lower bound for the objective function.

2) Local Search Algorithm. Although the solution of a relaxation problem z (denoted as $\hat{\psi}_z$) provides a lower bound, it may not be feasible. The goal of local search algorithm (see line 7 in Algorithm 1) is to obtain a feasible solution ψ_z from $\hat{\psi}_z$, which provides a upper bound. We use the same routing solution $\mathbf{u}$ to that in $\hat{\psi}_z$ as the feasible solution ψ_z. Then we try to determine the values for $\mathbf{x}$ and $\mathbf{p}$ in ψ_z such that (9) holds for each link as shown in Algorithm 2: Initially, with the fixed x_{ij}^m in problem z, each p_{ij}^m is set to the smallest value in its value space, i.e., $(p_{ij}^m)_L = 0$ for unused bands and $(p_{ij}^m)_L = (\frac{d_{ij}}{R_T^{\max}})^n P_{max}$ for active bands. Based on these p_{ij}^m, we compute the capacity $\sum_{m \in \mathcal{M}_i \cap \mathcal{M}_j} W_m \log(1 + \gamma d_{ij}^{-n} p_{ij}^m)$ and the requirement $\sum_{c \in \mathcal{N}} u_{ij}^c$ for each link. If there is no link whose requirement is larger than its capacity, then the feasible solution is found. Otherwise, we find a link and try to satisfy (9) by enlarging its capacity through increasing the transmission power under its value limitation and using some other available but unused bands. It should be noted that once one band is used (i.e., $x_{ij}^m = 1$), it has the impact on the available bands and transmission power limitation of other links (see line 6 in Algorithm 2). After we do this for all links with requirements larger than their capacities, if all the requirements have been satisfied, then we have found the feasible ψ_z. Otherwise, we fail to find a feasible solution and we set the objective value to ∞.

Algorithm 2 Local Search Algorithm

1: Set $p_{ij}^m = (p_{ij}^m)_L$ based on the fixed x_{ij}^m. Compute the capacity and requirement for each link.
2: **for** each link $i \to j$ whose requirement is larger than its capacity **do**
3: **while** its requirement is larger than its capacity **do**
4: Increase p_{ij}^m among active bands with Δ each time in the non-increasing order of $\hat{p}_{ij}^m$ and under the limitation that $p_{ij}^m \le (p_{ij}^m)_U$.
5: If its capacity is still insufficient, then try to use an available but unused band m in the non-increasing order of $\hat{p}_{ij}^m$. Also set $x_{ij}^m = 1$, $x_{ih}^m = 0$ for $h \in \mathcal{T}_i^m \backslash \{j\}$ based on (5), and let $(p_{kh}^m)_U = (\frac{d_{kj}}{R_I^{\max}})^n P_{max}$ for $k \in \mathcal{I}_j^m \backslash \{i\}$, $h \in \mathcal{T}_k^m$ based on (7).
6: If there is not any available band and the power can not be increased any more, **break**.
7: **end while**
8: If link $i \to j$ cannot be satisfied, set the objective value to ∞ and stop.
9: **end for**
10: **for** each link whose requirement is larger than 0 **do**
11: Back up the found $\mathbf{x}$ and $\mathbf{p}$.
12: Use additional available and unused bands, and adjust the transmission power on each its used band to a same value under the limitations.
13: If its requirement can be still satisfied, use the adjusted $\mathbf{x}$ and $\mathbf{p}$. Otherwise, use the backup $\mathbf{x}$ and $\mathbf{p}$.
14: **end for**

After the feasible $\mathbf{x}$ and $\mathbf{p}$ are found, we can try to adjust $\mathbf{x}$ and $\mathbf{p}$ to further reduce the object value (see line 10∼14 in Algorithm 2). From the Shannon's capacity formula, we find that given the capacity and channel condition, the more bandwidths the link uses, the less power it consumes. Besides, based on the convexity of log function, we know that given capacity and bandwidth, the optimal transmission power on each band is equivalent. Thus, for each link, we can use the additionally available and unused bands. Except for that, we check if the link's requirement can be still satisfied after we adjust the transmission power on each its available and unused band under the limitation to a same value. If the answer is yes, then we adjust the power and treat the adjusted $\mathbf{x}$ and $\mathbf{p}$ as solution, otherwise, we still use the originally found $\mathbf{x}$ and $\mathbf{p}$ as solution.

3) Branching. After updating LB to the lower bound LB_z of problem z, if it is still smaller than $UB - \theta$, we should further close the gap between LB and UB to get the θ-optimal solution, that is branching (see line 11 in Algorithm 1). To achieve this, we can scale down the feasible space by fixing some $\mathbf{x}$ variables, which are called partition variables. How to choose these partition variables is important as it affects the computing complexity. We propose to choose the partition variable based

on the solution of the relaxation problem $\hat{\psi}_z$, that is, we choose an x_{ij}^m with the largest relaxation error $\min\{\hat{x}_{ij}^m, 1-\hat{x}_{ij}^m\}$ among all $\mathbf{x}$ variables, and fix its value in two new sub-problems $z1$ and $z2$ as 0 and 1, respectively. It should be noted that the fixed x_{ij}^m imposes constraints on other variables: if $x_{ij}^m = 0$, then we have $p_{ij}^m = 0$ based on (6); if $x_{ij}^m = 1$, then we have $x_{ih}^m = 0$ for $h \in \mathcal{I}_i^m, h \neq j$ based on (5), $p_{ij}^m \geq (\frac{d_{ij}}{R_T^{\max}})^n P_{max}$ based on (8), and $p_{kh}^m \leq (\frac{d_{kj}}{R_I^{\max}})^n P_{max}$ for $k \in \mathcal{I}_j^m, k \neq i, h \in \mathcal{T}_k^m$ based on (7).

4) Cut Branches. We can reduce the computing complexity through cutting some branches (see line 13 in Algorithm 1). As shown in [18], during the process of ABB, if we find a problem z with $LB_z \geq UB - \theta$, then we can conclude that this problem cannot contribute to find a θ-optimal solution and thus we can remove this problem from the list for further consideration.

B. Performance Analysis of ABB

As shown in the previous subsection, only if $\theta \geq 0$, the worst case is that we traverse all binary variables and ABB can eventually find the θ-optimal solution. Although the worst-case complexity of ABB is exponential, the actual running time could be fast when all partition variables are binary. Firstly, we can reduce the computing complexity through cutting branches. Besides, once one x_{ij}^m variable is fixed as 1, then some related x variables can be fixed as 0 simultaneously based on (5), which further cuts down the actual running time. Especially, we can control computing complexity by setting a fit θ value. With a larger θ, ABB will achieve the θ-optimal solution faster. Except for that, we can set a maximum iteration number (after these iterations, we terminate the ABB even if we have not found the θ-optimal solution) to make sure the computing complexity is acceptable. However, it may not find the θ-optimal solution with the maximum iteration number constraint.

Obviously, ABB can just guarantee a θ-optimal solution at each time slot. It is unclear whether the θ-optimal solution achieved at each time slot have any influence on the energy consumption and stability performance of the whole system. In this subsection, we analyze the impact of implementation of ABB at each time slot on the system performance, which is shown in the following Theorem.

Theorem 2: If the rate matrix $\boldsymbol{\lambda}$ is interior to the network capacity region Λ, and the data arrivals as well as available band states are i.i.d. over time slots, then ABB ($\mathbf{p}^{\mathbf{ABB}}(t)$, $\mathbf{u}^{\mathbf{ABB}}(t)$) stabilizes the network and yields a time average congestion bound of

$$\overline{\sum_{i,c} U_i^{cABB}} \leq \frac{CN^2 + \theta + Vf(\mathbf{P_{max}})}{2\varepsilon_{\max}}. \tag{37}$$

Further, the time average cost satisfies

$$\overline{f(\mathbf{p}^{\mathbf{ABB}}(t))} \triangleq \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}\sum_{i\in\mathcal{N}}\mathbb{E}\left\{\sum_{m\in\mathcal{M}_i(t)}\sum_{j\in\mathcal{T}_i^m} p_{ij}^{mABB}(t)\right\} \leq f^* + \frac{CN^2+\theta}{V}. \tag{38}$$

Proof: Define $S(\mathbf{p}(t), \mathbf{u}(t)) = Vf(\mathbf{p}(t)) - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)(O_i^c(\mathbf{u}(t)) - O_i^c(\mathbf{u}(t))$. Suppose the objective value of the optimal solution to problem 1 is $f^* = S(\mathbf{p}^*(\mathbf{t}), \mathbf{u}^*(\mathbf{t}))$ where $\mathbf{p}^*(t)$ and $\mathbf{u}^*(t)$ are the corresponding optimal solutions. The object value achieved by ABB is $S(\mathbf{p}^{\mathbf{ABB}}(t), \mathbf{u}^{\mathbf{ABB}}(t))$ where $\mathbf{p}^{\mathbf{ABB}}(t)$ and $\mathbf{u}^{\mathbf{ABB}}(t)$ are the corresponding solutions. Then, based on Definition 2, we can have

$$S\big(\mathbf{p}^*(t), \mathbf{u}^*(t)\big) \geq S\Big(\mathbf{p}^{\mathbf{ABB}}(t), \mathbf{u}^{\mathbf{ABB}}(t)\Big) - \theta. \tag{39}$$

Besides, we have

$$S\Big(\mathbf{p}^{\mathbf{STAT}}(t), \mathbf{u}^{\mathbf{STAT}}(t)\Big) \geq S\big(\mathbf{p}^*(t), \mathbf{u}^*(t)\big). \tag{40}$$

Thus, combining (39) and (40), we obtain

$$S\Big(\mathbf{p}^{\mathbf{STAT}}(t), \mathbf{u}^{\mathbf{STAT}}(t)\Big) \geq S\Big(\mathbf{p}^{\mathbf{ABB}}(t), \mathbf{u}^{\mathbf{ABB}}(t)\Big) - \theta. \tag{41}$$

Plugging (41) into (18) under ABB, we obtain

$$\begin{aligned}
&\Delta(\mathbf{U}(t)) + V\mathbb{E}\Big\{f\Big(\mathbf{p}^{\mathbf{ABB}}(t)\Big)|\mathbf{U}(t)\Big\} \\
&\leq CN^2 + 2\sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} U_i^c(t)\lambda_i^c \\
&\quad + \mathbb{E}\Bigg\{Vf\Big(\mathbf{p}^{\mathbf{STAT}}(t)\Big) - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\Big(O_i^c\Big(\mathbf{u}^{\mathbf{STAT}}(t)\Big) \\
&\qquad - I_i^c\Big(\mathbf{u}^{\mathbf{STAT}}(t)\Big)\Big)\Big|\mathbf{U}(t)\Bigg\} + \theta.
\end{aligned} \tag{42}$$

With (29) and (30), we have

$$\begin{aligned}
&\Delta(\mathbf{U}(t)) + V\mathbb{E}\Big\{f\Big(\mathbf{p}^{\mathbf{ABB}}(t)\Big)|\mathbf{U}(t)\Big\} \\
&\qquad \leq CN^2 - \sum_{i\in\mathcal{N}}\sum_{c\in\mathcal{N}} 2U_i^c(t)\varepsilon + Vf^*(\varepsilon) + \theta.
\end{aligned} \tag{43}$$

Similarly to the proof of Theorem 1, we can easily derive the results (37) and (38). ■

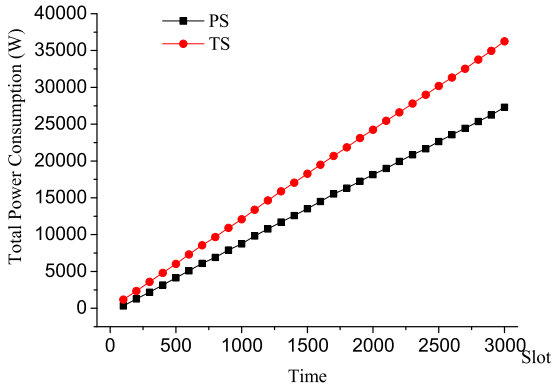
Theorem 2 shows that although ABB just achieves the θ-optimal solution at each time slot, it has the same performance features as EESA: 1) achieves the average power consumption deviated no more than $O(1/V)$ from the optimal solution; 2) ensures the system is stable; 3) offers an explicit tradeoff between energy consumption and delay performance.

VI. Performance Evaluation

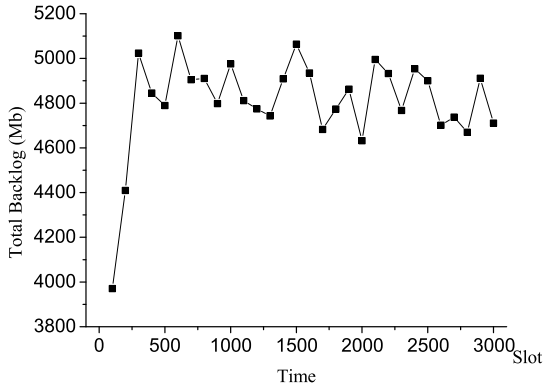
In this section, we present numerical results for the proposed algorithm. Our goals are to demonstrate the effectiveness of our algorithm and study the impact of control parameter V on the system performance.

A. Simulation Setup

We consider a multi-hop CRN consisting of $|\mathcal{N}| = 10$ RSs randomly distributed in a 800×800 m^2 area. There is a session with rate uniformly distributed within [0,85] Mb/slot, whose source RS and destination RS are selected randomly in $\mathcal{R}$. The total number of spectrum bands $|\mathcal{B}|$ is 8. For illustrative purposes, we assume all the bands have identical bandwidth,

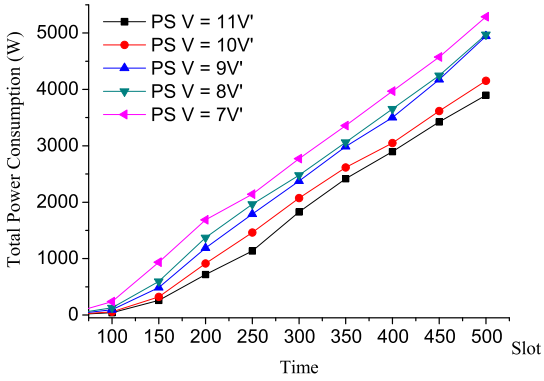


(a) Total power consumption performance of PS and TS.

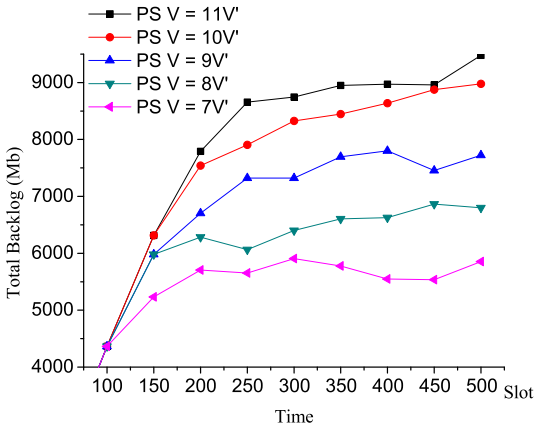


(b) Total backlog performance of PS.

Fig. 4. Performance comparison between PS and TS ($V = 5V'$).



(a) Total power consumption performance of PS.



(b) Total backlog performance of PS.

Fig. 5. Performance of PS with different control parameter values.

which is set to be 10 MHz, i.e., $W_m = 10$ MHz for all $m \in \mathcal{B}$. All the RSs have the same maximum transmission power $P_{max} = 10$ W on each band. Considering the AWGN channel, we assume the noise power $\eta W_m = 10^{-10}$ W [20]. Moreover, we suppose the path loss factor $n = 4$, the antenna parameter $\gamma = 3.90625$, the receiver sensitivity $\alpha_1 = 10^{-8}$ W and the interference threshold $\alpha_1 = 6.25 \times 10^{-10}$ W [20]. According to the illustration in Section III-B, we can calculate the maximum transmission range $R_T^{\max}$ and the maximum interference range $R_I^{\max}$, which are equal to 250 m and 500 m, respectively. Besides, for the simplicity of computation, we set the maximum iteration number in ABB to 1000. For ABB, we let $\Delta = 0.25$ W and $\theta = 0.25NV$. At each time slot, the available bands of each RS and the data rate are randomly generated.

B. Results and Analysis

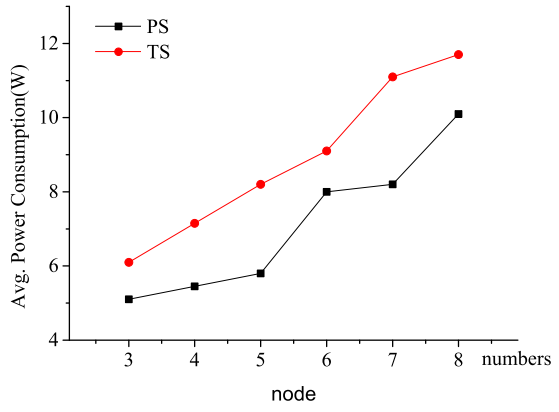
Based on the simulation settings above, we conduct simulations to study the optimal power consumption problem in CCH with the following two parts: 1) Through comparing our proposed strategy (denoted as PS) with the traditional real-time transmission strategy (denoted as TS, refer to Section I for the definition), we try to demonstrate the effectiveness of PS; 2) Through setting different values of control parameter $\mathcal{N}$, $\mathcal{B}$, and V, we try to investigate their impact on the system energy and delay performance.

Fig. 4 depicts the results of performance comparison between PS and TS. The control parameter is set to $V = 5V'$, where $V' = 7300$. From Fig. 4, we have the following two observations:

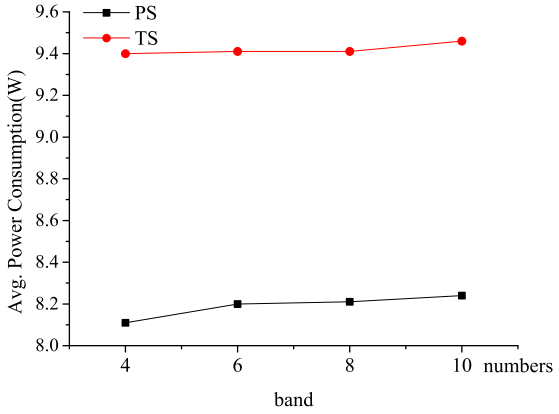
1) As shown in Fig. 4(a), the power consumption of PS is less than that of TS. The average power consumption of PS and TS are 9.125 W and 12.1 W, respectively, which means that PS can save 24.5% energy as compared to TS. The reason is that instead of transmitting the data timely, PS can intelligently delay some delay-tolerant data and transmit it when the system has more available spectrum bands. In this way, we can save some energy as depicted in the capacity formula, where the capacity increases linearly with bandwidth, but only logarithmically with transmission power.

2) As shown in Fig. 4(b), the total system backlog firstly increases, and finally reaches a steady value which is close to 4900 Mb. The result illustrates that PS can ensure the system is stable. Using Little's law, we can also calculate the average delay from the above mean backlog length.

Next, we study the performance of PS under different values of the control parameter V. The results are illustrated in Fig. 5. From Fig. 5(a), we find that the power consumption decreases with the increase of V. However, from Fig. 5(b), we find that the system backlog increases with the increase of V. All these observations demonstrate that we can achieve the tradeoff



(a) Average power consumption under different number of nodes



(b) Average power consumption under different number of bands

Fig. 6. Average power consumption comparison between PS and TS.

between the delay and energy saving performance through controlling the values of V, which corroborates the accuracy of our theoretical analysis in Theorem 2. Furthermore, whatever the value of V is, all the curves in Fig. 5(b) indicate that ABB can ensure the stability of system.

In addition, we compare the average power consumption of PS and TS under different number of nodes $|\mathcal{N}|$ and different number of available bands $|\mathcal{B}|$. The results are shown in Fig. 6. It can be observed that the average power consumption of PS is always lower than that of TS. Finally, we evaluate the transmission delay and convergence speed of PS with various number of nodes and available bands, and the results are illustrated in Table I. As Table I shows, the growth rate of delay is highly relative to the number of nodes and bands, but it is also proportional to the total backlog. Thus, the backlog could truly reflect the delay incurred by a data stream under different circumstance. Besides, the convergence speed of PS is very fast when the number of nodes and bands is small. When the number of nodes and bands increases, we can see that the PS algorithm can still achieve 25% energy efficiency gain at the expense of delay performance with the tolerable computing complexity through setting a maximum number of iterations. In a word, TS is suit for delay-sensitive services and PS is appropriate for delay-tolerant services such as data and streaming mobile apps.

TABLE I
THE PERFORMANCE OF DELAY AND NUMBER OF ITERATIONS OF PS

(a) Performance under different number of nodes

node number	delay (slot)	total backlog (Mb)	iteration times
3	49.05	2087	64
4	132.05	5898	198
5	161.19	7301	287
6	190.67	8705	710
7	253.33	11182	730
8	316.33	13659	750

(b) Performance under different number of bands

band number	delay (slot)	total backlog (Mb)	iteration times
4	249	10616	62
6	190.67	8705	710
8	173.2	7956	738
10	126.86	5303	761

VII. CONCLUSION

In this paper, based on a novel architecture of CRNs we have proposed for spectrum harvesting and sharing, we address the energy saving problem considering both temporal and spatial feature of varying spectrum. A cross-layer stochastic optimization framework which minimizes the time-average expected power consumption while stabilizing the network is formulated. Based on the Lyapunov optimization technique and branch-and-bound framework, we design an online algorithm to obtain an approximate solution. Theoretical analysis and simulation results show that our algorithm offers an explicit tradeoff between energy consumption and delay performance. Besides, numerical results illustrate that thanks to the specific characteristic in CRNs, i.e., both time and space varying features of spectrum availability, our proposed strategy is superior to the traditional real-time transmission strategy in term of energy consumption performance.

REFERENCES

[1] D. Chen, S. Yin, Q. Zhang, M. Liu, and S. Li, "Mining spectrum usage data: A large-scale spectrum measurement study," in *Proc. Int. Conf. Mobile Comput. Netw. (Mobicom)*, Beijing, China, Sep. 2009, pp. 13–24.

[2] M. A. McHenry, P. A. Tenhula, D. McCloskey, D. A. Roberson, and C. S. Hood, "Chicago spectrum occupancy measurements and analysis and a long-term studies proposal," in *Proc. TAPAS*, Boston, MA, USA, Aug. 2006, pp. 1–3.

[3] G. Gýr and F. Alagýz, "Green wireless communications via cognitive dimension: An overview," *IEEE Netw.*, vol. 25, no. 2, pp. 50–56, Mar./Apr. 2011.

[4] A. El Gamal, C. Nair, B. Prabhakar, E. Uysal-Biyikoglu, and S. Zahedi, "Energy-efficient scheduling of packet transmissions over wireless networks," in *Proc. IEEE INFOCOM*, New York, NY, USA, Jun. 2002, pp. 1773–1782.

[5] Y. Sun, S. Du, O. Gurewitz, and D. B. Johnson, "DW-MAC: A low latency, energy efficient demand-wakeup MAC protocol for wireless sensor networks," in *Proc. ACM MobiHoc*, Hong Kong, May 2008, pp. 53–62.

[6] Z. Ding, S. M. Perlaza, I. Esnaola, and H. V. Poor, "Power allocation strategies in energy harvesting wireless cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 846–860, Feb. 2014.

[7] Q. Dong, S. Banerjee, M. Adler, and A. Misra, "Minimum energy reliable paths using unreliable wireless links," in *Proc. ACM MobiHoc*, Urbana-Champaign, IL, USA, May 2005, pp. 449–459.

[8] S. Kwon and N. B. Shroff, "Unified energy-efficient routing for multi-hop wireless networks," in *Proc. IEEE INFOCOM*, Phoenix, AZ, USA, Apr. 2008, pp. 430–438.

[9] D. Zhang, G. Li, K. Zheng, X. Ming, and Z.-H. Pan, "An energy-balanced routing method based on forward-aware factor for wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 766–773, Feb. 2014.

[10] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, USA, Apr. 2003, pp. 702–711.

[11] R. Bhatia and M. Kodialam, "On power efficient communication over multi-hop wireless networks: Joint routing, scheduling and power control," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004, pp. 1457–1466.

[12] L. Lin, X. Lin, and N. B. Shroff, "Low-complexity and distributed energy minimization in multihop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 501–514, Apr. 2010.

[13] W. Liao, M. Li, S. Salinas, P. Li, and M. Pan, "Energy-source-aware cost optimization for green cellular networks with strong stability, " *IEEE Trans. Emerg. Topics Comput.*, to be published. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TETC.2014.2386612

[14] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[15] J. Tang, S. Misra, and G. Xue, "Joint spectrum allocation and scheduling for fair spectrum sharing in cognitive radio wireless networks," *Comput. Netw.*, vol. 52, no. 11, pp. 2148–2158, Aug. 2008.

[16] N. Michelusi and U. Mitra, "Cross-layer estimation and control for cognitive radio: Exploiting sparse network dynamics," *IEEE Trans. Cogn. Commun. Netw.*, vol. 1, no. 1, pp. 128–145, Mar. 2015.

[17] Y. T. Hou, Y. Shi, and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 146–155, Jan. 2008.

[18] Y. Shi, Y. T. Hou, and H. Zhou, "Per-node based optimal power control for multi-hop cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5290–5299, Oct. 2009.

[19] Y. Shi and Y. T. Hou, "A distributed optimization algorithm for multi-hop cognitive radio networks," in *Proc. IEEE INFOCOM*, Phoenix, AZ, USA, Apr. 2008, pp. 1292–1300.

[20] M. Pan, C. Zhang, P. Li, and Y. Fang, "Spectrum harvesting and sharing in multi-hop CRNs under uncertain spectrum supply," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 369–378, Feb. 2012.

[21] M. Pan, H. Yue, Y. Fang, and H. Li, "The X loss: Band-mix selection for opportunistic spectrum accessing with uncertain spectrum supply from primary service providers," *IEEE Trans. Mobile Comput.*, vol. 11, no. 12, pp. 2133–2144, Dec. 2012.

[22] M. Pan, C. Zhang, P. Li, and Y. Fang, "Joint routing and scheduling for cognitive radio networks under uncertain spectrum supply," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 2237–2245.

[23] M. Li, P. Li, X. Huang, Y. Fang, and S. Glisic, "Energy consumption optimization for multihop cognitive cellular networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 2, pp. 358–372, Feb. 2015.

[24] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[25] C.-P. Li and M. J. Neely, "Energy-optimal scheduling with dynamic channel acquisition in wireless downlinks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 4, pp. 527–539, Apr. 2010.

[26] L. Huang and M. J. Neely, "Delay efficient scheduling via redundant constraints in multihop networks," *Perform. Eval.*, vol. 68, no. 8, pp. 670–689, Aug. 2011.

[27] R. Urgaonkar and M. J. Neely, "Opportunistic cooperation in cognitive femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 607–616, Apr. 2012.

[28] C. Jiang, H. Zhang, Y. Ren, and H.-H. Chen, "Energy-efficient non-cooperative cognitive radio networks: Micro, meso, and macro views," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 14–20, Jul. 2014.

[29] S. Bayhan and F. Alagoz, "Scheduling in centralized cognitive radio networks for energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 582–595, Feb. 2013.

[30] Y. Liu, L. X. Cai, X. Shen, and H. Luo, "Deploying cognitive cellular networks under dynamic resource management," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 82–88, Apr. 2013.

[31] W.-Y. Lee and I. F. Akyildiz, "Spectrum-aware mobility management in cognitive radio cellular networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 529–542, Apr. 2012.

[32] K.-C. Chen, Y.-J. Peng, N. Prasad, Y.-C. Liang, and S. Sun, "Cognitive radio network architecture: Part I—General structure," in *Proc. ACM Int. Conf. Ubiquitous Inf. Manag. Commun. (ICUIMC)*, Suwon, South Korea, Jan. 2008, pp. 114–119.

[33] I. F. Akyildiz, W.-Y. Lee, and K. R. Chowdhury, "CRAHNs: Cognitive radio ad hoc networks," *Ad Hoc Netw.*, vol. 7, no. 5, pp. 810–836, Jul. 2009.

[34] J. Sachs, I. Maric, and A. Goldsmith, "Cognitive cellular systems within the TV spectrum," in *Proc. IEEE DySPAN*, Singapore, Apr. 2010, pp. 1–12.

[35] H.-Y. Hsieh and R. Sivakumar, "Performance comparison of cellular and multi-hop wireless networks: A quantitative study," in *Proc. ACM SIGMETRICS*, Cambridge, MA, USA, Jun. 2001, pp. 113–122.

[36] X. Li, B.-C. Seet, and P. H. J. Chong, "Multihop cellular networks: Technology and economics," *Comput. Netw.*, vol. 52, no. 9, pp. 1825–1837, 2008.

[37] Defense Advanced Research Projects Agency (DARPA). *The NeXt Generation Program (XG) Official Website*. Accessed on Sep. 27, 2016. [Online]. Available: http://www.darpa.mil/sto/smallunitops/xg.html

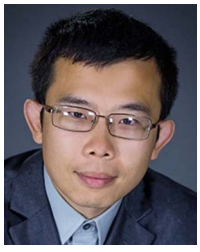
[38] H. Yue, M. Pan, Y. Fang, and S. Glisic, "Spectrum and energy efficient relay station placement in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 883–893, May 2013.

[39] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: Freeman, 1979, pp. 245–248.

[40] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*. New York, NY, USA: Wiley, 1999.

[41] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*, vol. 3. San Rafael, CA, USA: Morgan & Claypool, 2010, pp. 1–122.

Jinlin Peng received the B.S. and Ph.D. degrees from the Department of Electrical Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2010 and 2015, respectively. He is currently a Research Engineer with Huawei Shanghai Research Institute, Shanghai. His research interests include green communication and cooperative communication.

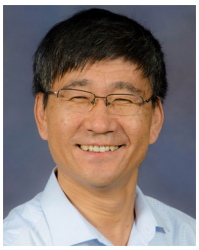


Hao Yue received the B.Eng. degree in telecommunication engineering from Xidian University, Xi'an, China, in 2009, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2015. He is currently an Assistant Processor with the Department of Computer Science, San Francisco State University, San Francisco, CA, USA. His research interests include cyber-physical systems, cybersecurity, wireless networking, and mobile computing.

Kaiping Xue (M'09–SM'15) received the B.S. degree from the Department of Information Security, University of Science and Technology of China (USTC), in 2003, and the Ph.D. degree from the Department of Electronic Engineering and Information Science, USTC, in 2007, where he is currently an Associate Professor with the Department of Information Security and Department. His research interests include next-generation Internet, distributed networks, and network security.

Ying Luo received the B.S. and M.S. degrees from the Department of Information Engineering, South West University of Science and Technology in 2012 and 2015, respectively, where she is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering and Information Science. Her research interests have focused on the analysis of heterogeneous cellular networks.

Peilin Hong received the B.S. and M.S. degrees from the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China (USTC), in 1983 and 1986, respectively. She is currently a Professor and an Advisor for Ph.D. candidates with the Department of EEIS, USTC. She has published two books and over 150 academic papers in several journals and conference proceedings. Her research interests include next-generation Internet, policy control, IP QoS, and information security.



Yuguang Fang (F'08) received the M.S. degree from Qufu Normal University, Shandong, China, in 1987, the Ph.D degree from Case Western Reserve University in 1994, and the Ph.D. degree from Boston University in 1997. He joined the Department of Electrical and Computer Engineering, University of Florida in 2000, where he has been a Full Professor since 2005. He held a University of Florida Research Foundation Professorship from 2006 to 2009, a Changjiang Scholar Chair Professorship with Xidian University, China, from 2008 to 2011 and with Dalian Maritime University, China from 2015 to present, a Guest Chair Professorship with Tsinghua University, China, from 2009.

He was a recipient of the U.S. National Science Foundation Career Award in 2001, the Office of Naval Research Young Investigator Award in 2002, the 2015 IEEE Communications Society CISTC Technical Recognition Award, the 2014 IEEE Communications Society WTC Recognition Award, the Best Paper Award from IEEE ICNP in 2006, the 2010–2011 UF Doctoral Dissertation Advisor/Mentoring Award, the 2011 Florida Blue Key/UF Homecoming Distinguished Faculty Award, and the 2009 UF College of Engineering Faculty Mentoring Award. He was the Editor-in-Chief of the IEEE WIRELESS COMMUNICATIONS from 2009 to 2012, and serves/served on several editorial boards of journals, including the IEEE TRANSACTIONS ON MOBILE COMPUTING from 2003 to 2008 and from 2011, the IEEE TRANSACTIONS ON COMMUNICATIONS from 2000 to 2011, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2002 to 2009. He is the Editor-in-Chief of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He actively participates in conference organizations such as serving as the Technical Program Co-Chair for the IEEE INOFOCOM'2014 and the Technical Program Vice-Chair for the IEEE INFOCOM'2005. He is a fellow of IEEE and a fellow of the American Association for the Advancement of Science.