This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2721107, IEEE Transactions on Green Communications and Networking

Energy Efficient Scheduling for Delay-Constrained Spectrum Aggregation: A Differentiated Water-Filling Approach

Yitu Wang, *Student Member*, Wei Wang, *Senior Member, IEEE*, Lin Chen, *Member, IEEE*, Pan Zhou, *Member, IEEE*, Zhaoyang Zhang, *Member, IEEE*

***Abstract*—Spectrum aggregation (SA) enables wireless devices to utilize heterogeneous resources, which can potentially fulfill the requirement of broadband services. In this paper, we study the delay-constrained spectrum aggregation, where the characteristics of SA bring various technical challenges. Specifically, the SA capability limitation induces a complicated coupling among the data rate, power and channel allocation, and the total power consumption varies according to the channel aggregation due to the SA circuit structure. Moreover, with these practical considerations, the water-filling power allocation cannot be adopted over all the channels. To overcome these challenges, we design the ESSA scheduling scheme in two steps. First, with given the sum data rate and the channel allocation, we minimize the total power consumption for SA, including both the transmit power and the circuit power. Due to the properties of delay-constrained SA, we divide the scheduled users into the conforming and nonconforming user sets, and design their water-filling power allocation strategies differentially. Second, based on the differentiated water-filling power control, we optimize the channel allocation and rate control iteratively via Lyapunov optimization to minimize the power consumption under average delay constraint. The proposed ESSA scheme is finally evaluated by simulation.**

I. Introduction

Spectrum aggregation (SA) [2], [3] has its distinctive value in wireless communication systems, which enables the devices to provide homogeneous broadband service by bonding heterogeneous fragmentary spectrum resources. It is proved by theoretical analysis and experimental results that SA can significantly enhance the system capability and reduce the power consumption [4]. Recently, SA became one of the key features of enhanced mobile broadband (eMBB) for 5G standardization.

Future wireless systems have critical requirements to support higher data rate and more real-time services [5]. Wireless transmission with high data rate costs enormous energy, and real-time services have a strict delay requirement. The power consumption and the delay performance are crucial towards reliability and stability of wireless systems [6]. The tradeoff between power and delay is discussed in [7] and [8], where the former solves the problem by Markov decision process in a single user case, which incurs a very high complexity considering multiple users, while the latter one addresses the problem using Lyapunov method which can be applied to multi-user but single-channel cases in wireless systems.

With the SA capability, a device can adjust the number of channels adaptively according to the service demands, which provides an extra degree of freedom to achieve energy efficiency. Due to the practical considerations, there are technical challenges involved as follows:

- **Delay performance under SA capability limitation**: Due to the practical hardware limitation [9], [10], the aggregation range of SA is restricted, i.e., a limited number of channels can be aggregated. The SA capability limitation leads to the complicated coupling between rate, power and channel allocation, whose effects to the delay performance are not straightforward. As a result, the conventional scheduling based on Lyapunov method [8] cannot be adopted directly for delay-constrained SA.
- **Energy consumption with SA circuit structure**: The process of SA requires the support of specific circuit structure [11], where a part of energy is consumed by the transmission over each channel and the other part of energy is consumed by each device which aggregates multiple channels. As a result, the total power consumption varies according to the combination of channel allocation and needs to be taken into consideration for designing the energy efficient scheduling.

In spite of the above challenges, SA makes it possible that one user can support the simultaneous transmission over multiple channels. It is obvious that the power consumption is reduced by balancing the water-filling levels across multiple channels used by the same user. If the water-filling levels of more channels are balanced together, the power consumption will be reduced further. In delay-constrained SA systems, due to the limitation of the SA capability, we handle the water-filling scheme for the users differentially, i.e., the water-filling levels are balanced across a part of users and are set individually for the other users. The number of users with the

Part of this work has been presented at IEEE Globecom 2016, Washington DC, U.S.A. [1].

This work is supported in part by National Natural Science Foundation of China (Nos. 61571396, 61401169), Zhejiang Provincial Natural Science Foundation of China (No. LR17F010001), Young Elite Scientist Sponsorship Program by CAST (No. 2016QNRC001), and Fundamental Research Funds for the Central Universities (No. 2015XZZX001-03).

Yitu Wang, Wei Wang and Zhaoyang Zhang are with College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China.

Lin Chen is with Laboratoire de Recherche en Informatique (LRI), University of Paris-Sud and CNRS, Orsay 91405, France.

Pan Zhou is with School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China.

same water-filling level depends on the aggregation capability of SA.

In this paper, we develop an analytical framework for energy efficient scheduling for delay-constrained SA (ESSA). Taking the SA capability and the circuit power consumption into consideration, ESSA determines the data rate, transmit power and channel allocation to minimize the energy consumption with average delay constraint. The scheduling decisions are made according to both the channel quality and the queue backlogs. Specifically, we design the ESSA scheduling algorithm in two steps. First, with given the sum data rate and the channel allocation, we minimize the total power consumption for SA, including both the transmit power as well as the circuit power. The water-filling levels are balanced differentially by partitioning the users into the conforming/nonconforming user sets. Second, based on the results of power control, we propose a suboptimal scheduling algorithm based on Lyapunov optimization to determine the sum data rate and the channel allocation in an iterative manner. The simulation results show that ESSA reduces the power consumption significantly for delay-constrained SA.

The rest of this paper is organized as follows. Section II discusses the related works. Section III presents the system model. In Section IV, we propose the ESSA algorithm to minimize the power consumption by balancing the water-filling levels across users. Following this, the performance of ESSA is evaluated by simulation results in Sections V. Finally, this paper is concluded in Section VI.

II. Related Works

This paper develops an analytical framework for energy efficient scheduling for delay-constrained SA. In this section, we briefly review existing works on the SA resource optimization and delay-aware considerations.

A. SA Resource Optimization

There are several existing works on the resource optimization for SA. In [12], a heuristic suboptimal algorithm considering both efficiency and fairness is proposed for SA by optimizing two metrics separately to lower the complexity. An optimal one is proposed later in [13] in a two-carrier case. In [14], a utility optimal resource allocation scheme for SA is proposed with the log utility proportional fairness by adopting primal-dual Lagrangian method. In [15], the spectrum sharing is studied between two groups of users, i.e., public safety and commercial LTE users, and a resource allocation algorithm is proposed with providing priority to the public safety users whose minimum quality of service should be ensured. A joint carrier selection and power control strategy is proposed in [16] to improve the average throughput using an estimation-based method. An energy efficient dynamic carrier aggregation scheduling scheme is proposed in [17], in which an energy efficient metric based on bits-per-joule is derived for elastic traffic. In [18], the capacity and delay trade-off is studied for cognitive radio networks with SA and characterizes the delay distribution using approximation, but the performance degradation can only be shown through simulation. In [19], a survey of radio resource management is given for SA in LTE-A, where the scheduling achieves the performance gain from multi-user frequency domain scheduling diversity by prioritizing the allocation of resource blocks to the users that experience good channel quality. Since the resource blocks scheduling delay is an important design constraint, scheduling structures are proposed to minimize the scheduling delay [20], [21]. Different to the above mentioned works, we take the practical issues including the aggregation capability limitation and the circuit structure of SA into consideration, which bring new technical challenges to the SA resource allocation. The challenges are even more pronounced when considering the aggregation capability limitation for delay-constrained systems, because we can allocate only a limited number of channels to the users with urgent demands but their delay-constrained requirements need to be met. Such a limitation leads to the complicated coupling between the data rate, power and channel allocation, whose effects to the delay performance are not straightforward.

B. Delay-Aware Considerations

There are a lot of research efforts made to delay-aware considerations. A systematic approach to the delay-aware optimization problem is the Markov Decision Process (MDP). Generally, the optimal control policy can be obtained by solving the Bellman equation. However, conventional solutions, such as brute-force iteration or policy iteration [22], incur huge complexity. To alleviate the computational complexity, some works use the stochastic approximation approach with distributed online learning algorithm [23] to tackle the problem, which has desirable linear complexity. However, the stochastic learning approach can only give a numerical solution and may suffer from slow convergence and lack of insight. To bypass the difficulty of characterizing the delay, the blocking probability is adopted instead to represent the delay indirectly. In [24], stochastic delay is discussed by using discrete Markov process, and a scheduling policy is proposed to minimize the delay of the scheduled packets, which also incurs very high complexity.

We treat this issue and provide insight into the problem by stochastic optimization [8], i.e., Lyapunov method is adopted to balance power and delay with low complexity. The channel bonding in SA leads to complicated rate, power and channel allocation, which are coupled with each other and their effects to the delay are not straightforward, especially for the cases with the SA capability limitation and the circuit structure. As a result, the conventional Lyapunov method cannot be adopted directly for delay-constrained SA, and our paper tries to address this problem by optimizing the data rate and the channel allocation with differentiated water-filling power allocation in an iterative manner.

III. System Model

In this section, we first introduce the physical layer model of SA systems. Considering the specified SA circuit structure, we present the circuit power consumption model for SA. Finally, we formulate the power minimization problem with delay constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2721107, IEEE Transactions on Green Communications and Networking

A. SA System

Consider a wireless SA system, which includes N user-receiver pairs sharing K time-varying channels, each of which has the same bandwidth. Denote $\mathcal{N}$ and $\mathcal{K}$ as the set of user and channel indexes respectively, i.e., $\mathcal{N} = \{1, 2, \cdots, N\}$ and $\mathcal{K} = \{1, 2, \cdots, K\}$. Because of the SA capability limitation, a user can transmit over at most M channels simultaneously[1]. The time is slotted and the duration of each time slot is assumed to be a unit of time.

Let $x_i(t)$ denote the information symbol for the i-th pair. The received signal at receiver i using channel j is

$$y_i(t) = h_j^i(t)\sqrt{P_j(t)}x_i(t) + n_i(t), \tag{1}$$

where $h_j^i(t)$ is the complex channel fading coefficient between pair i using channel j, $P_j(t)$ is the transmit power of channel j and $n_i(t)$ is the i.i.d. complex Gaussian channel noise with power N_0. When user i is scheduled for transmission over channel j, the received signal-to-noise ratio is

$$\gamma_i^j(t) = \frac{|h_j^i(t)|^2}{N_0} P_j(t). \tag{2}$$

Let $\mathbf{S}(t)$ be the global channel state in slot t, i.e., $\mathbf{S}(t) = \big(S_j^i(t), j \in \mathcal{K}, i \in \mathcal{N}\big)$, where $S_j^i(t) = \frac{|h_j^i(t)|^2}{N_0}$ represents the state of channel j for user i in slot t, which remains constant within a slot and is i.i.d. over time slots.

At the beginning of each slot, a centralized controller schedules the users to transmit and determines the associated scheduling control variables as follows:

- *Data rate* $\mathbf{r}(t)$: $\mathbf{r}(t) = \big\{r_i(t), \forall i \in \mathcal{N}\big\}$, where $r_i(t)$ is the data rate of user i in slot t.
- *Transmit power* $\mathbf{P}(t)$: $\mathbf{P}(t) = \big\{P_j(t), \forall j \in \mathcal{K}\big\}$, where $P_j(t)$ is the transmit power of channel j in slot t.
- *Channel allocation* $\mathbf{b}(t)$: $\mathbf{b}(t) = \big\{b_j^i(t), \forall i \in \mathcal{N}, j \in \mathcal{K}\big\}$, where $b_j^i(t) \in \{0, 1\}$ and $b_j^i(t) = 1$ represents that user i transmits over channel j in slot t.

The data rate $\mathbf{r}_i(t)$ are determined by channel allocation, power allocation and the corresponding channel quality as well.

$$r_i(t) = \sum_{j=1}^{K} b_j^i(t) R_j\big(t\big), \tag{3}$$

where R_j is the transmission rate of channel j which can be calculated as

$$R_j(t) = \sum_{i=1}^{N} b_j^i(t) \log_2\big(1 + S_j^i(t)P_j(t)\big). \tag{4}$$

Note that $\mathbf{b}(t)$ should satisfy $\sum_{j=1}^{K} b_j^i(t) \leq M, \forall i \in \mathcal{N}$ with the consideration of the SA capability and $\sum_{i=1}^{N} b_j^i(t) \leq 1, \forall j \in \mathcal{K}$ for exclusive channel use.

B. SA Circuit Structure

Although there are different circuit implementations [11], [25] for SA, one of the common characteristics is that the channels aggregated by the same user can share a part of circuit modules. Considering the specified structure for SA circuit as illustrated in Fig. 1, we divide the SA circuit into two parts as follows:

- *Individual modules*: Each set of individual modules provides the processing for a single channel. The individual modules usually include discrete Fourier transform (DFT), mapping, inverse fast Fourier transform (IFFT), clock pulse insertion and multiplier.
- *Shared modules*: The set of shared modules provides the functions which are shared by the channels aggregated by a device. The shared modules usually include digital to analog converter (DAC), mixer, linear power amplifier (LPA) and antennas.

Denote P_1 as the power consumption of a set of individual modules and P_2 as that of a set of shared modules[2]. Based on the specified SA circuit structure, the total circuit power consumption $P_c(t)$ can be modeled as

$$P_c(t) = \sum_{i=1}^{N}\sum_{j=1}^{K} b_j^i(t)P_1 + \sum_{i=1}^{N}\Big(1 - \prod_{j=1}^{K}(1 - b_j^i(t))\Big)P_2. \tag{5}$$

Note that the shared modules consume power if any channel is used. If user i aggregates at least one channel for transmission, i.e., $\exists j \in \mathcal{K}, b_j^i(t) = 1$, then $\prod_{j=1}^{K}(1 - b_j^i(t)) = 0$ and user i consumes power $P2$ for the shared circuit modules.

C. Problem Formulation

Since both power and delay are critical performance metrics in wireless systems, there is an inherent tradeoff between the power consumption and the delay performance. For SA, this tradeoff is more complicated than the conventional systems due to the challenges mentioned before. In this paper, we focus on deriving an energy efficient scheduling for delay-constrained SA to balance the tradeoff between power and delay.

To analyze the average queuing delay, we first discuss the packet queue backlog, since the average queuing delay can be measured by the average queue length according to Little's theorem [27]. Each user possesses a packet queue at its transmitter, whose length is denoted as $U_i(t)$ for user i in slot t. Let $\mathbf{A}(t) = \big\{A_i(t), \forall i \in \mathcal{N}\big\}$ be the random packet arrivals (number of bits) from the application layers to the packet queues, where $A_i(t)$ is the number of arrived bits for user i in slot t. Assume that $\mathbf{A}(t)$ is i.i.d. over time, with $\mathbb{E}[A_i(t)] = \lambda_i$, where λ_i is the average arrival rate for user i. The queue dynamics of $U_i(t)$ is

$$U_i(t+1) = \max\big\{U_i(t) - r_i(t), 0\big\} + A_i(t). \tag{6}$$

[1] The SA capability M is usually smaller than the number of channels K. When the SA capability is larger, the problem is reduced to that with $M = K = \min\{M, K\}$.

[2] SA circuit power consumption can be estimated based on hardware datasheets and time spent by the operations [26].

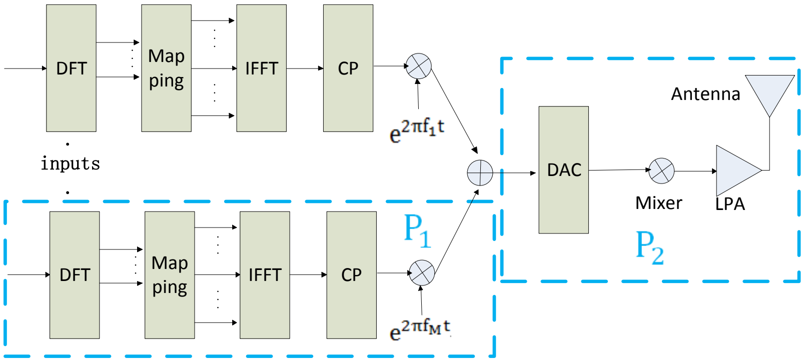


Fig. 1. Illustration of the SA circuit structure

Our goal is to minimize the energy consumption for delay-constrained SA by scheduling, which can be formulated as

$$\min_{\mathbf{b}(t),\mathbf{P}(t)} \quad \sum_{j=1}^{K}\sum_{i=1}^{N} b_j^i(t)P_j(t) + P_c(t), \tag{7}$$

$$s.t. \quad \frac{\mathbb{E}\left[\sum_{i=1}^{N} U_i(t)\right]}{N} \leq Q, \tag{8}$$

$$\mathbb{E}\left[\sum_{j=1}^{K} b_j^i(t)R_j(t)\right] \geq \lambda_i, \forall i \in \mathcal{N}, \tag{9}$$

$$\sum_{j=1}^{K} b_j^i(t) \leq M, \forall i \in \mathcal{N}, \tag{10}$$

$$\sum_{i=1}^{N} b_j^i(t) \leq 1, \forall j \in \mathcal{K}, \tag{11}$$

where the constraint (8) implies the average delay constraint, in which Q denotes the target average queue length corresponding to the average delay, (9) guarantees the system stability, (10) implies the aggregation range due to SA capability and (11) implies a channel can be used by at most one user at a time slot.

IV. ESSA: Energy Efficient Scheduling for Delay Constrained Spectrum Aggregation

The optimization problem in (7) is a mixed integer programming (MIP) problem, which is NP-hard and is usually difficult to solve efficiently. In this section, we solve the problem and design the scheduling scheme in two steps. First, under the sum data rate and the channel allocation matirx $\mathbf{b}(t)$, we minimize the total power consumption of the SA system. Due to the limitations in delay-constrained SA systems, we partition the scheduled users into the conforming/nonconforming user sets, and determine their water-filling power allocation strategies differentially. Second, based on the results of the minimum power consumption under the sum data rate and the channel allocation matirx $\mathbf{b}(t)$, we propose a sub-optimal scheduling algorithm by Lyapunov optimization, to determine the sum data rate and the channel allocation in an iterative manner. Based on the differentiated water-filling power allocation, the data rates are optimized according to the current queue lengths adaptively by adopting Lyapunov optimization.

A. Differentiated Water-Filling Power Allocation

In this subsection, we consider the minimization of power consumption under the sum data rate and the channel allocation matrix $\mathbf{b}(t)$, and obtain the minimum power consumption, which provides a basis for designing ESSA.

In SA systems, one user can support the simultaneous transmission over multiple channels, which makes it possible that the power consumption is reduced by balancing the water-filling levels across multiple channels used by the same user[3]. Moving one step ahead, we balance the water-filling levels across users to reduce the power consumption further. Compared to fixing the individual data rates, the power allocation algorithm with a given sum rate provides more degrees of freedom to minimize the power consumption. Specifically, under the given sum rate, the water-filling levels can be balanced across users, which will further reduce the power consumption due to the exponential relationship between the power consumption and the data rate in Shannon's capacity. Let us consider a simple motivating example that N users transmit over N channels with the same channel gain. With the given individual data rates $r_i, \forall i$, the power consumption is $\sum_{i=1}^{N} 2^{r_i} - N$, while for these users with the sum data rate $\sum_{i=1}^{N} r_i$, the power consumption is $N \cdot 2^{\sum_{i=1}^{N} r_i/N} - N$. It is obvious that $\sum_{i=1}^{N} 2^{r_i} - N \geq N \cdot 2^{\sum_{i=1}^{N} r_i/N} - N$, which implies that balancing the water-filling levels reduces the power consumption.

Under the sum data rate and the channel allocation matrix $\mathbf{b}(t)$, we adopt the rate constraint[4] instead of the queue length constraint in (8), and rewrite the power minimization problem

[3] According to Jensen's inequality [28], the energy consumption of satisfying the sum rate over multiple channels is not more than that of satisfying the rate requirement over each corresponding channel, because the data rate is an increasing concave function of the transmit power [29].

[4] The sum data rate will be given according to the current queue length and the target Q in the next subsection. Thus, the sum data rate constraint can be adopted to guarantee the queue length constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2721107, IEEE Transactions on Green Communications and Networking

as follows:

$$\min_{\mathbf{P}(t)} \quad \sum_{j=1}^{K}\sum_{i=1}^{N} b_j^i(t)P_j(t) + P_c(t), \tag{12}$$

$$s.t. \quad \sum_{j=1}^{K}\sum_{i=1}^{N} b_j^i(t)\log_2\big(1+S_j^i(t)P_j(t)\big) = \sum_{i=1}^{N} r_i(t), \tag{13}$$

$$\mathbb{E}\Big[\sum_{j=1}^{K} b_j^i(t)R_j(t)\Big] = \mathbb{E}[r_i(t)] \geq \lambda_i, \forall i \in \mathcal{N}, \tag{14}$$

$$\sum_{j=1}^{K} b_j^i(t) \leq M, \forall i \in \mathcal{N}, \tag{15}$$

$$\sum_{i=1}^{N} b_j^i(t) \leq 1, \forall j \in \mathcal{K}, \tag{16}$$

where the sum data rate constraint (13) guarantees that K channels provide enough capacity to support transmission and the constraint (14) guarantees the stability of the system. Note that the constraint (9) is a constraint on the average sum queue length rather than the queue length of each user. In such a case, we only need to allocate the power considering the sum data rate constraint[5] in (13) for satisfying (9).

To solve the above optimization problem, we first treat an easier case in which only constraints (13) and (16) are considered, to extract some insight. We establish the Lagrangian function according to (12) and (13) as

$$\begin{aligned} Z\big(P_j(t),\gamma\big) = & \sum_{j=1}^{K}\sum_{i=1}^{N} b_j^i(t)P_j(t) \\ & -\gamma\Big(\sum_{j=1}^{K}\sum_{i=1}^{N} b_j^i(t)\log_2\big(1+S_j^i(t)P_j(t)\big) - \sum_{i=1}^{N} r_i(t)\Big), \end{aligned} \tag{17}$$

where γ is the Lagrangian multiplier.

By Lagrangian method, i.e., letting the partial derivative of $Z\big(P_j(t),\gamma\big)$ with respect to $P_j(t)$ equals to 0, we have

$$\sum_{i=1}^{N} b_j^i(t) - \gamma\Big(\sum_{i=1}^{N} b_j^i(t)\frac{S_j^i(t)\ln 2}{1+S_j^i(t)P_j(t)}\Big) = 0. \tag{18}$$

As for $\sum_{i=1}^{N} b_j^i(t) = 0$, we have $P_j(t) = 0$. As for $\sum_{i=1}^{N} b_j^i(t) = 1$, where $b_j^{i^*}(t) = 1$ and $b_j^i(t) = 0, \forall i \neq i^*$, we have

$$P_j(t) = \gamma\ln 2 - \frac{1}{S_j^{i^*}(t)} = \gamma\ln 2 - \frac{1}{\sum_{i=1}^{N} b_j^i(t)S_j^i(t)}. \tag{19}$$

Letting the partial derivative of γ equals to 0, we have

$$\sum_{j=1}^{K}\sum_{i=1}^{N} b_j^i(t)\log_2\big(1+S_j^i(t)P_j(t)\big) = \sum_{i=1}^{N} r_i(t). \tag{20}$$

To balance the water-filling levels across users, the above equation can be rewritten as

$$\sum_{i=1}^{N} b_j^i(t)\log_2\big(1+S_j^i(t)P_j(t)\big) = \frac{\sum_{i=1}^{N} r_i(t)}{K}. \tag{21}$$

Substituting (19) into (21), we obtain γ as

$$\gamma = \frac{2^{\frac{\sum_{i=1}^{N} r_i(t)}{K}}}{\prod_{k=1}^{K}(\sum_{i=1}^{N} b_k^i(t)S_k^i(t))^{\frac{1}{K}}\ln 2}, \tag{22}$$

and thus the transmit power over channel j is

$$P_j(t) = \frac{2^{\frac{\sum_{i=1}^{N} r_i(t)}{K}}}{\prod_{k=1}^{K}(\sum_{i=1}^{N} b_k^i(t)S_k^i(t))^{\frac{1}{K}}} - \frac{1}{\sum_{i=1}^{N} b_j^i(t)S_j^i(t)}. \tag{23}$$

Essentially, (23) provides a water-filling power allocation over all channels at t, where the first term of $P_j(t)$ is the water-filling level which is the same for all users and the last term of $P_j(t)$ is the sea bed levels.

By substituting (23) into Shannon's formula (4), we obtain the transmission rate of channel j as

$$\begin{aligned} R_j(t) = & \sum_{i=1}^{N} b_j^i(t)\log_2\Bigg(S_j^i(t)\frac{2^{\frac{\sum_{i=1}^{N} r_i(t)}{K}}}{\sum_{k=1}^{K}\sum_{n=1}^{N} b_k^n(t)S_k^n(t)} + 1 \\ & - \frac{S_j^i(t)}{\sum_{n=1}^{N} b_j^n(t)S_j^n(t)}\Bigg). \end{aligned} \tag{24}$$

Since one channel can only be used by one user at t, the last two terms in $\log$ can be reduced to zero. In this case, (24) can be rewritten as

$$\begin{aligned} R_j(t) = & \frac{\sum_{i=1}^{N} r_i(t)}{K} - \sum_{k=1}^{K}\frac{1}{K}\log_2\sum_{i=1}^{N} b_k^i(t)S_k^i(t) \\ & + \log_2\sum_{i=1}^{N} b_j^i(t)S_j^i(t). \end{aligned} \tag{25}$$

Based on the above analysis on the easier case with constaints (13) and (16), we further take constraints on the system stability and SA limitation in (14) and (15) into consideration. If the water-filling levels are balanced for all the users, it is possible that the total transmission rate of M channels cannot support the average arrival rate λ_i, which leads to the instability of the system. Consider a case that even though M channels (the maximum number of channels due to the SA capability) are allocated to the user, constraint (14) cannot be met. Then we should partition the user into the nonconforming user set such that the power can be allocated separately to provide a higher data rate. To enforce the constraint (14) and ensure the stability of the system, we classify the users into two categories in the following definition:

Definition 1 (Conforming/Nonconforming User Sets)**.** *A set of users $\mathcal{U}$ is said to be a conforming user set if it is satisfied that*

$$0 < \lambda_i \leq M\frac{\sum_{l\in\mathcal{U}} r_l(t)}{|\mathcal{C}|}, \forall i \in \mathcal{U}, \tag{26}$$

where $\mathcal{C}$ is the set of channels allocated to the users in $\mathcal{U}$ and $|\mathcal{C}|$ is the cardinality of $\mathcal{C}$ which is the number of channels for

[5] The sum data rate will be determined by Lyapunov optimization in Section IV.B.

the users in $\mathcal{U}$. On the other hand, the scheduled users with a positive rate but not in $\mathcal{U}$ are called nonconforming users. The set of nonconforming users is denoted as $\mathcal{U}^-$ and the set of channels allocated to user $i \in \mathcal{U}^-$ is denoted as $\mathcal{C}_i$. ■

If (26) holds, i.e., the water-filling levels are balanced over the conforming users using the proposed algorithm, then conforming users will be stabilized; Nonconforming users allocate the power separately to provide higher data rates, hence, they can also be stabilized. The conforming user set $\mathcal{U}$ can be determined by executing the following pseudo codes in Algorithm 1, which is launched at the beginning of each time slot.

Algorithm 1 Partition of the Conforming/Nonconforming Users

```
Initialize $\mathcal{U} = \mathcal{N}$ and $\mathcal{C} = \mathcal{K}$
repeat
    for $i \in \mathcal{U}$ do
        if User $i$ does not satisfy (26) then
            Partition user $i$ into the nonconforming user set $\mathcal{U}^-$.
            Partition $M$ channels into the set $\mathcal{C}_i$.
        end if
    end for
until No more user is partitioned into the nonconforming user set
```

We further discuss the uniqueness of the achieved partition results in the following lemma:

Lemma 1 (Unique Partition Property)**.** *The partition method in Algorithm 1 achieves a unique partition result of the conforming/nonconforming user sets.*

Proof: Please refer to Appendix A. ■

Because of the system stability requirement (14) and the limitation of SA capability (15), the water-filling levels cannot be balanced across all the channels but just the channels in $\mathcal{C}$ for the conforming users in $\mathcal{U}$. Note that it is not always to transmit over all the channels, since some channels can be deactivated to save power due to the SA circuit structure.

Theorem 1 (Minimum Power Consumption)**.** *Under the sum data rate for the conforming set/each nonconforming set and the channel allocation matrix $\mathbf{b}(t)$, the minimum power consumption $\phi\big(\sum_i r_i(t), \mathbf{b}(t)\big)$ is*

$$\begin{aligned}\phi\Big(\sum_i r_i(t), \mathbf{b}(t)\Big) &= |\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\prod_{k\in\mathcal{C}} \sum_{i\in\mathcal{U}} b_k^i(t) S_k^i(t)^{\frac{1}{|\mathcal{C}|}}} \\ &+ \sum_{i\in\mathcal{U}^-} |\mathcal{C}_i| \frac{2^{\frac{r_i(t)}{|\mathcal{C}_i|}}}{\prod_{k\in\mathcal{C}_i} \sum_{i\in\mathcal{U}^-} b_k^i(t) S_k^i(t)^{\frac{1}{|\mathcal{C}_i|}}} \\ &+ \big(|\mathcal{C}| + \sum_{i\in\mathcal{U}^-} |\mathcal{C}_i|\big) P_1 + \big(|\mathcal{U}| + |\mathcal{U}^-|\big) P_2 \\ &- \sum_{k\in\mathcal{C}} \frac{1}{\sum_{i\in\mathcal{U}} b_k^i(t) S_k^i(t)} - \sum_{i\in\mathcal{U}^-} \sum_{k\in\mathcal{C}_i} \frac{1}{\sum_{l\in\mathcal{U}^-} b_k^l(t) S_k^l(t)},\end{aligned} \tag{27}$$

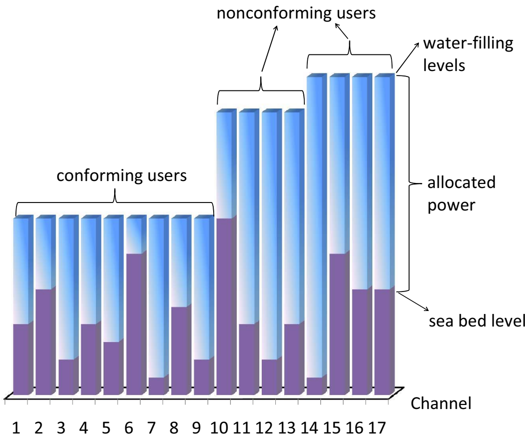


Fig. 2. Balancing the water-filling levels across users

where active channel sets $\mathcal{C}$ and $\mathcal{C}_i$ can be found by one-dimensional numerical searching.

Proof: Please refer to Appendix B. ■

Remark 1 (Balancing the Water-Filling Levels Across Users)**.** *The minimum power consumption in Theorem 1 allocates the power over channels in a differentiated water-filling manner. Specifically, the water-filling levels of the users in the conforming user set $\mathcal{U}$ are balanced together across users as the first line in (27), while those of the users in $\mathcal{U}^-$ are balanced for each individual user as the second line in (27). With balancing the water-filling levels in the conforming user set, the transmission load for each channel is the same, which leads to the decreasing of the power consumption. Moreover, it can be seen that the power can be further saved when more users are balancing the water-filling levels together.* ■

Fig. 2 illustrates the differentiated water-filling power allocation, where the 9 channels at the left hand provide the transmission for the users in $\mathcal{U}$ and the water-filling levels are balanced over all 9 channels. On the contrary, two nonconforming users use the 8 channels at the right hand and allocate their power separately. Obviously, the more channels allocated to the users in $\mathcal{U}$, the less power is consumed because of the benefit achieved by balancing the water-filling levels across more users and channels. Note that a large SA capability M can significantly increase the number of channels with a balanced water-filling level, which contributes to reducing the power consumption.

B. Rate Control and Channel Allocation

For a queueing system, the time average data rate on any link can only exceed the arrival rate due to edge effects. To ensure that the edge effects pose limited influence on the system, we adopt a virtual queue according to [31]. Both the actual and virtual queues should be stabilized for the whole system stability. The actual queue is stabilized to reduce the difference between the current queue and the target queue length Q. The Lyapunov function is composed of two parts as follows:

$$\psi\big(\mathbf{U}(t),\mathbf{X}(t)\big)=L\big(\mathbf{U}(t)\big)+J\big(\mathbf{X}(t)\big). \tag{28}$$

The function $L\big(\mathbf{U}(t)\big)$ is designed to be exponential, which reaches its minimum when $U_i(t)=Q,\forall i\in\mathcal{N}$, and increases exponentially with the difference between $U_i(t)$ and Q.

$$L\big(\mathbf{U}(t)\big)=\sum_{i\in\mathcal{N}}\big(e^{\omega(U_i(t)-Q)}+e^{\omega(Q-U_i(t))}-2\big), \tag{29}$$

where ω is a positive coefficient affecting the rate of exponential increase. This Lyapunov function provides a large enough penalty to push the queue length $U_i(t)$ to the target queue length Q.

The function $J\big(\mathbf{X}(t)\big)$ is designed for the stability of the virtual queue $\mathbf{X}(t)$ as

$$J\big(\mathbf{X}(t)\big)=\sum_{i\in\mathcal{N}}X_i^2(t). \tag{30}$$

The dynamics of the actual queue $U_i(t)$ and the virtual queue $X_i(t)$ are presented respectively as

$$U_i(t+1)=U_i(t)-r_i(t)+A_i(t). \tag{31}$$

$$\begin{aligned}X_i(t)=&\max\big\{X_i(t)-\big(r_i(t)+\epsilon 1_{U_i(t)<Q}(t)\big),0\big\}\\&+A_i(t)+\epsilon 1_{U_i(t)\geq Q}(t),\end{aligned} \tag{32}$$

whose Lyapunov drift can be obtained using a similar method in [31], and is given as follows

$$\begin{aligned}\Delta L(U_i(t))\leq\,& e^{\omega(A_{max}+\nu_{max}-Q)}+\frac{\epsilon\omega}{2}e^{\omega(U_{max}-Q)}\\&-1_{U_i(t)\geq Q}(t)\omega e^{\omega(U_i(t)-Q)}(\delta_i(t)-\frac{\epsilon}{2})\\&-1_{U_i(t)<Q}(t)\omega e^{\omega(Q-U_i(t))}(\delta_i(t)+\frac{\epsilon}{2}).\\\Delta J(X_i(t))\leq\,&(A_{max}+\epsilon)^2+(\nu_{max}+\epsilon)^2\\&-1_{U_i(t)\geq Q}(t)X_i(t)(\delta_i(t)+\epsilon)\\&-1_{U_i(t)<Q}(t)X_i(t)(-\delta_i(t)+\epsilon),\end{aligned} \tag{33}$$

where $\nu_{max}=\max_{i,t}\{r_i(t)\}$, ϵ is a parameter influencing the response rate of the algorithm and $\delta_i(t)=r_i(t)-A_i(t)$.

Using the buffer partitioning technique in [31]: $\delta_i(t)=\epsilon$ when $U_i(t)\geq Q$ and $\delta_i(t)=-\epsilon$ when $U_i(t)<Q$, (33) can be rewritten as

$$\begin{aligned}\Delta L(U_i(t))\leq\,& e^{\omega(A_{max}+\nu_{max}-Q)}+\frac{\epsilon\omega}{2}e^{\omega(U_{max}-Q)}\\&-1_{U_i(t)\geq Q}(t)\omega e^{\omega(U_i(t)-Q)}\frac{\epsilon}{2}\\&+1_{U_i(t)<Q}(t)\omega e^{\omega(Q-U_i(t))}\frac{\epsilon}{2}.\\\Delta J(X_i(t))\leq\,&(A_{max}+\epsilon)^2+(\nu_{max}+\epsilon)^2\\&-1_{U_i(t)\geq Q}(t)X_i(t)(2\epsilon)-1_{U_i(t)<Q}(t)X_i(t)(2\epsilon).\end{aligned} \tag{34}$$

To minimize the Lyapunov drift, we minimize the upper bound instead, which is widely adopted, such as dynamic backpressure algorithm [32]. By removing the constant terms, the part in the objective function for Lyapunov drift can be written as

$$\begin{aligned}&-\sum_{i=1}^{N}1_{U_i(t)\geq Q}(t)\big(\omega e^{\omega(U_i(t)-Q)}+2X_i(t)\big)r_i(t)\\&-\sum_{i=1}^{N}1_{U_i(t)<Q}(t)\big(-\omega e^{\omega(Q-U_i(t))}+2X_i(t)\big)r_i(t).\end{aligned} \tag{35}$$

To minimize the total power consumption with average delay constraint, we adopt V as the weight of the power consumption to balance the tradeoff between the power consumption (27) and the Lyapunov drift (35). The objective is

$$\begin{aligned}\min_{\sum_i r_i(t),\mathbf{b}(t)}\; & Y(t)=V\Bigg(|\mathcal{C}|\frac{2^{\frac{\sum_{i\in\mathcal{U}}r_i(t)}{|\mathcal{C}|}}}{\prod_{j\in\mathcal{C}}\sum_{i\in\mathcal{U}}b_j^i(t)S_j^i(t)^{\frac{1}{|\mathcal{C}|}}}\\&+\sum_{i\in\mathcal{U}^-}|\mathcal{C}_i|\frac{2^{\frac{r_i(t)}{|\mathcal{C}_i|}}}{\prod_{j\in\mathcal{C}_i}b_j^i(t)S_j^i(t)^{\frac{1}{|\mathcal{C}_i|}}}\\&-\sum_{j\in\mathcal{C}}\frac{1}{\sum_{i\in\mathcal{U}}b_j^i(t)S_j^i(t)}-\sum_{j\in\mathcal{C}_i}\frac{1}{\sum_{i\in\mathcal{U}^-}b_j^i(t)S_j^i(t)}\\&+\big(|\mathcal{C}|+\sum_{i\in\mathcal{U}^-}|\mathcal{C}_i|\big)P_1+\big(|\mathcal{U}|+|\mathcal{U}^-|\big)P_2\Bigg)\\&-\sum_{i=1}^{N}1_{U_i(t)\geq Q}(t)\big(\omega e^{\omega(U_i(t)-Q)}+2X_i(t)\big)r_i(t)\\&-\sum_{i=1}^{N}1_{U_i(t)<Q}(t)\big(-\omega e^{\omega(Q-U_i(t))}+2X_i(t)\big)r_i(t).\end{aligned} \tag{36}$$

A large V provides a high weight to the metric on the power consumption, e.g., the unit price of power is expensive, which leads to a small power consumption and a large average queue backlog. Similarly, a small V achieves a large power consumption and a small average queue backlog.

The SA capability limitation has to be considered for stabilizing the system, which induces a more complicated coupling among the data rate, power and channel allocation. As a result, the conventional Lyapunov method cannot be adopted directly for delay-constrained SA, and we address this problem by optimizing the data rate and the channel allocation in an iterative manner as follows.

1) Rate Vector Optimization:

For a given channel allocation matrix $\mathbf{b}(t)$, we optimize the sum data rate to minimize $Y(t)$ in (36).

Considering the same water-filling level of the power allocation for the users in $\mathcal{U}$, the scheduled rate of each user $i\in\mathcal{U}$ is a function of $\mathbf{b}(t)$ as

$$\begin{aligned}r_i(t)=&\sum_{k\in\mathcal{C}}b_k^i(t)\frac{\sum_{l\in\mathcal{U}}r_l(t)}{|\mathcal{C}|}\\&+\sum_{k\in\mathcal{C}}b_k^i(t)\cdot\Big(\log_2\big(\sum_{l\in\mathcal{U}}b_k^l(t)S_k^l(t)\big)-\sum_{j\in\mathcal{C}}\frac{1}{N}\log_2\big(\sum_{l\in\mathcal{U}}b_j^l(t)S_j^l(t)\big)\Big).\end{aligned} \tag{37}$$

Substituting (37) into (36), and using Lagrangian method, we obtain

$$\begin{aligned}\sum_{l\in\mathcal{U}} r_l(t) =&|\mathcal{C}|\Bigg(\log_2(\sum_{i\in\mathcal{U}}\big(1_{U_i(t)\geq Q}(t)\big(\omega e^{\omega(U_i(t)-Q)}+2X_i(t)\big)\\&+1_{U_i(t)<Q}(t)\big(-\omega e^{\omega(Q-U_i(t))}+2X_i(t)\big)\big)\frac{\sum_{j\in\mathcal{C}} b_j^i(t)}{|\mathcal{C}|}\\&+\sum_{j\in\mathcal{C}}\log_2(\sum_{i\in\mathcal{U}}(b_j^i(t)S_j^i(t))^{\frac{1}{|\mathcal{C}|}}))-\log_2(V)\Bigg).\end{aligned} \tag{38}$$

According to (26), if $\exists i\in\mathcal{U},\lambda_i > M\frac{\sum_{l\in\mathcal{U}} r_l(t)}{|\mathcal{C}|}$, user i is partitioned into the nonconforming set $\mathcal{U}^-$. The procedure is iterated until $\forall i\in\mathcal{U},\lambda_i < M\frac{\sum_{l\in\mathcal{U}} r_l(t)}{|\mathcal{C}|}$, and one dimensional search is adopted to obtain the optimal $\mathcal{C}$.

As for the users in $\mathcal{U}^-$, the scheduled rate can be obtained as

$$\begin{aligned}r_i =&|\mathcal{C}_i|\Bigg(\log_2(\sum_{l\in\mathcal{U}^-}\big(1_{U_l(t)\geq Q}(t)\big(\omega e^{\omega(U_l(t)-Q)}+2X_l(t)\big)\\&+1_{U_l(t)<Q}(t)\big(-\omega e^{\omega(Q-U_l(t))}+2X_l(t)\big)\big)\frac{\sum_{j\in\mathcal{C}_i} b_j^i(t)}{|\mathcal{C}_i|}\\&+\sum_{j\in\mathcal{C}_i}\log_2(\sum_{l\in\mathcal{U}^-}(b_j^l(t)S_j^l(t))^{\frac{1}{|\mathcal{C}_l|}}))-\log_2(V)\Bigg).\end{aligned} \tag{39}$$

2) Channel Allocation Matrix Optimization:

For given the sum data rate, we optimize the channel allocation matrix $\mathbf{b}(t)$ to minimize $Y(t)$ in (36).

For the cases without the SA capability limitation, the channel allocation does not affect the power consumption performance. However, for the delay-constrained SA in this paper, the channel allocation is a 0-1 integer optimization problem which is usually NP-hard. We propose a heuristic scheme to find a sub-optimal allocation matrix $\mathbf{b}(t)$.

After a feasible solution of $\mathbf{b}(t)$ is obtained by greedy method, we propose a switching algorithm to improve the value of the objective function. The channels are switched from user k to user i according to the ascending order of the priority values $f(i)$ unidirectionally, which indicates that channel j can not switch from user k to user i if $f(k) > f(i)$. Define $g(i,k,j)$ as the improvement value which represents the performance improvement which channel j is allocated to another user i instead of the original user k. The priority value and improvement value for the users and the channels are

$$\begin{aligned}f(i) =& 1_{U_i(t)\geq Q}(t)\big(\omega e^{\omega(U_i(t)-Q)}+2X_i(t)\big)\\&+1_{U_i(t)<Q}(t)\big(-\omega e^{\omega(Q-U_i(t))}+2X_i(t)\big).\\ g(i,k,j) =& Y(t)|_{b_j^k(t)=1}-Y(t)|_{b_j^i(t)=1}.\end{aligned} \tag{40}$$

Only one channel j with the largest improvement value $g(i,k,j)$ is switched from user k to user i, where

$$j=\arg\max_l g(i,k,l),\forall i,k. \tag{41}$$

This procedure is repeated until the value of the objective function cannot be further improved.

Based on the discussions in the above two subsections, we provide the details of ESSA using pseudo codes in Algorithm 2, which is launched at the beginning of each time slot.

Algorithm 2 ESSA

```
Initialize $\mathcal{U}=\{1,2,\cdots,N\}$ and $\mathcal{C}=\{1,2,\cdots,K\}$.
repeat
    (Rate Vector Optimization)
    Determine $\mathcal{U}$ and $\mathcal{U}^-$ according to Algorithm 1.
    Obtain the rate of user $i\in\mathcal{U}$ according to (37) and (38).
    Obtain the rate of user $i\in\mathcal{U}^-$ according to (39).
    (Channel Allocation Matrix Optimization)
    Allocate $b_j^i=0, r_i=0,\forall j$ to user $i$ with $U_i(t)<Q$, and initialize $\mathbf{b}(t)$ that each channel in $\mathcal{K}$ is allocated to the user who has the best channel quality among the rest users in $\mathcal{N}$.
    repeat
        Determine $\mathcal{U}$ and $\mathcal{U}^-$ according to Algorithm 1.
        Calculate $f(i)$ for each user, and calculate $g(i,k,j)$ for each channel and users according to (40).
        Find the channel $j$ according to (41), set $b_k^j(t)=0$ and $b_i^j(t)=1$ if $f(i)>f(k)$.
    until The largest $g(i,k,j)$ is less than 0.
until Both the Rate Vector and the Channel Allocation Matrix are steady.
Update the actual queue and the virtual queue according to (31) and (32), respectively.
```

To analyze the convergence behavior of ESSA, we first evaluate the value of the objective function in each iteration round. 1). Given the sum rate, the proposed switching algorithm decreases the value of the objective function by improving the channel quality. 2). The value of the first two lines in (38) and (39) does not change during the iteration, while the value of the last line is decreased compared to that of the previous round because of the improved channel quality. Thus, the value of the objective function decreases monotonically in each iteration round. Then, it is obvious that the monotonic objective function is lower bounded by the minimum value 0 when $r_i=0,\forall i$. Therefore, we can draw the conclusion that the proposed iterative algorithm converges. The iteration stops when the changes of both the data rate and the channel allocation matrix between two iteration rounds are small enough.

C. Performance Analysis

In this subsection, we discuss the performance of the proposed ESSA algorithm by theoretic analysis. Here, we consider two performance metrics including the average delay T_{ave} and the average per-user power consumption E_{ave}. Especially, we provide comparison between the performance of proposed algorithm ESSA and the baseline algorithm TOCA [8], since TOCA captures the case without SA which can be treated as a special case of ESSA when $M=1$.

By setting the parameters as the same as those in [8], i.e., $N=K$, $\omega=\frac{\epsilon}{\delta_{\max}^2}e^{\frac{-\epsilon}{\delta_{\max}}}$, $\epsilon=1/V$, $Q=(6/\omega)\log_2\big(1/\epsilon\big)$, and $\nu=\max_{i,t}\{r_i(t)\}$, where $\delta_{\max}=\max_{i,t}\{\lambda_i-r_i(t)\}$, we obtain the following theorem to provide T_{ave} and E_{ave} of the proposed ESSA algorithm.

Theorem 2 (T_{ave} and E_{ave} of ESSA)**.** *If ESSA is a* $1+\gamma$ *approximation of the optimal algorithm. For any* $V > \nu$*, ESSA algorithm yields*

$$T_{ave} \leq \frac{(1+\gamma)}{\omega}\log_2\left(2\frac{D+\frac{V}{N}h}{\omega\epsilon}\right). \tag{42}$$

$$E_{ave} - \phi_{min}(\boldsymbol{\lambda}) \leq \frac{D}{V} + \gamma\phi_{min}(\boldsymbol{\lambda}) + \frac{(1+\gamma)}{N}\sum_{i=1}^{N}\frac{\partial\phi(\boldsymbol{\lambda})}{\partial\lambda_i}\epsilon\boldsymbol{\Delta}_i + (1+\gamma)\epsilon^2\kappa, \tag{43}$$

where

$$\kappa = \max_{\boldsymbol{\sigma}\in(-\epsilon,+\epsilon)^N}\frac{||\nabla^2\phi(\boldsymbol{\lambda}+\boldsymbol{\sigma})||}{2}, \kappa > 0, \tag{44}$$

$$\phi(\boldsymbol{\lambda}) = \min_{\mathbf{b}(t)}\phi(\boldsymbol{\lambda}, \mathbf{b}(t)), \tag{45}$$

$$\Delta_i = \begin{cases} \frac{\sum_{i\in\mathcal{U}}\left(\alpha_i^R - \alpha_i^L\right)}{|\mathcal{C}|} & i \in \mathcal{U} \\ \frac{\alpha_i^R - \alpha_i^L}{|\mathcal{C}_i|} & i \in \mathcal{U}^-, \end{cases} \tag{46}$$

where $\alpha_i^R = \Pr[U_i(t) \geq Q]$*,* $\alpha_i^L = \Pr[U_i(t) < Q]$*.*

Proof: Please refer to Appendix D. ■

Remark 2. (Performance Comparison between ESSA and TOCA): *From [8], the performance metrics of TOCA are provided as follows:*

$$T^*_{ave} \leq \frac{1}{\omega}\log_2\left(2\frac{D+\frac{V}{N}h}{\omega\epsilon}\right). \tag{47}$$

$$E^*_{ave} - \phi^*_{min}(\boldsymbol{\lambda}) \leq \frac{D}{V} + \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\phi^*(\boldsymbol{\lambda})}{\partial\lambda_i}\epsilon\left(\alpha_i^R - \alpha_i^L\right) + \epsilon^2\kappa, \tag{48}$$

where

$$\phi^*(\boldsymbol{\lambda}) = \sum_{i\in\mathcal{N}}\frac{2^{\lambda_i(t)}-1}{S_i^i(t)}. \tag{49}$$

Comparing the above two equations with the results in Theorem 2, we can find that the optimal ESSA $(\gamma = 0)$ *and TOCA achieve the same average delay by choosing a smaller* $V' < V$ *for ESSA, which satisfies*

$$\frac{(1+\gamma)}{\omega'}\log\left(2\frac{D+\frac{V'}{N}h}{\omega'\epsilon'}\right) = \frac{1}{\omega}\log_2\left(2\frac{D+\frac{V}{N}h}{\omega\epsilon}\right), \tag{50}$$

where $\omega' = \frac{\epsilon'}{\delta^2_{\max}}e^{\frac{-\epsilon'}{\delta_{\max}}}$ $\epsilon' = 1/V'$*, ESSA and TOCA achieve the same average delay.*

According to the convexity of $\phi(\boldsymbol{\lambda})$ *since the partial derivative grows exponentially with* λ*, the proposed algorithm ESSA allocates more channels to the users with large enough* λ*, which significantly reduce the power consumption of the users in* $\mathbf{U}^-$ *at the cost of slightly increasing (or even decreasing due to SA) the power consumption of the users in* $\mathbf{U}$*. Therefore, ESSA achieves a lower power consumption compared with TOCA, despite of choosing a slightly smaller* $V' < V$ *in ESSA to achieve the same average delay with TOCA.* ■

We further discuss the complexity of the proposed algorithm ESSA, which is mainly brought by the switching procedure in channel allocation matrix optimization.

Lemma 2 (Asymptotic Analysis)**.** *When SA capability* M *is large enough such that all the users are in the conforming user set* $\mathcal{U}$ *and* $\frac{\sum_{i\in\mathcal{U}}r_i(t)}{K} \gg \max_{i,j}\{S_j^i(t)\}$*, a channel can only be switched once during the switching procedure.*

Proof: Please refer to Appendix C. ■

Remark 3 (Computational Complexity)**.** *The complexity is mainly brought by channel allocation optimization whose complexity is* $O(N^3K^2)$ *in each iteration. Therefore, when* M *is large, the complexity of ESSA is* $O(cN^3K^3)$*, where* c *denotes the iteration rounds. When the SA capability* M *is small, the complexity of ESSA is* $O(cN^4K^3)$ *due to the unidirectional channel switch for at most* N *times per channel.*

V. Simulation

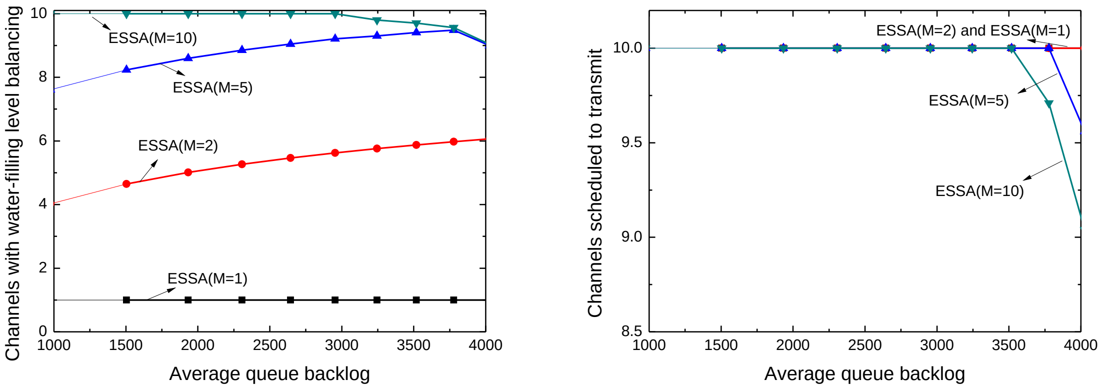
In this section, we evaluate the performance of the proposed ESSA algorithm by simulation. The performance evaluation includes two aspects. First, the characteristics of the proposed schemes are analyzed, including the channel utilization and the influence of key parameters. Second, the performance of the proposed scheme is compared with those of the conventional schemes. For the performance comparison, we adopt three baseline schemes:

- *Baseline 1 (Tradeoff optimal control algorithm (TOCA))*: Lyapunov method is adopted for single-channel case without the water-filling power allocation [8], which is a special case for ESSA ($M = 1$).
- *Baseline 2 (Throughput-based scheme)*: Throughput performance is optimized according to the current channel quality [33] with considering SA.
- *Baseline 3 (Queue-based scheme)*: Lyapunov method is adopted for time invariant links [34] with considering SA.

In this simulation, there are 10 users in two categories, including the heavy users[6] and the light users. With different ratios of users with heavy traffic, we consider three scenarios [35] that reflect the expected share of mobile broadband subscribers, i.e.,

- **Scenario 1**: 20 percent of the subscribers are classified as the users with heavy traffic. The data arrival follows Bernoulli distribution with the probabilities 1, 0, 0, 0, 0, 0, 1, 0, 0, 0 for each user respectively, and the data amount of each time of arrival is set to $A_i(t) = 12$. This scenario serves as an upper bound on the traffic for 2015.
- **Scenario 2**: 10 percent of the subscribers are classified as the users with heavy traffic. The data arrival follows Bernoulli distribution with the probabilities 0.9, 0.15, 0.25, 0.2, 0.1, 0.15, 0.35 0.15, 0.3, 0.45 for each user respectively, and the data amount of each time of arrival is set to $A_i(t) = 8$. This scenario is the most relevant European scenario for 2015.
- **Scenario 3**: No user with heavy traffic. The data arrival follows Bernoulli distribution with the probabilities 0.8, 0.6, 0.9, 0.4, 0.2, 0.6, 0.7, 0.3, 0.6, 0.9 for each user respectively, and the data amount of each time of arrival

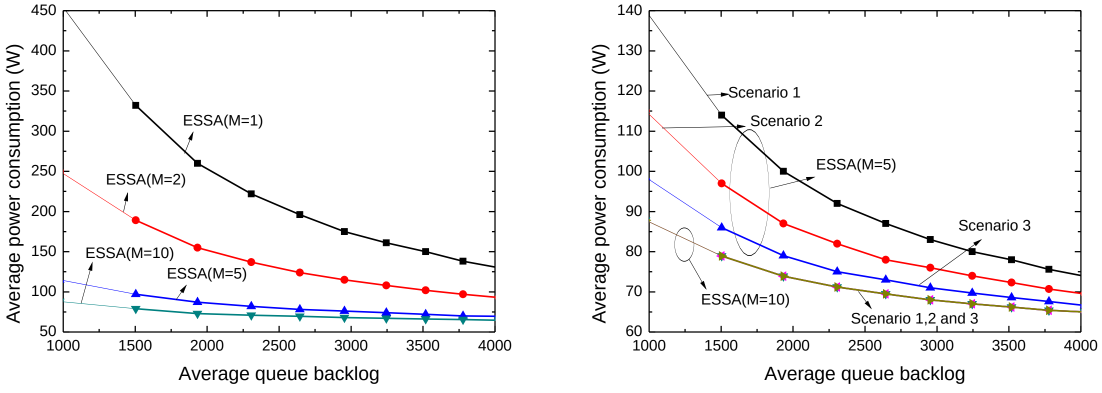
[6] A user is with heavy traffic if it satisfies $\lambda_i \geq \sum_{j=1}^{10}\lambda_j/3$ in our simulation.



(a) Channels with water-filling level balancing

(b) Channels scheduled to transmit

Fig. 3. Channel utilization



(a) Influence brought by M

(b) Performance of ESSA in different scenario

Fig. 4. Influence of the aggregation capability M

is set to $A_i(t) = 4$. This scenario serves as an ideal traffic mode.

These users share 10 time-varying channels, which obey the Rayleigh distribution with the fading coefficient 6.5 and are i.i.d. over time slots. The values of the circuit power are set as $P_1 = 2.04W$ and $P_2 = 4.06W$ according to the practical base stations [35]. We adopt the iteration number $c = 50$ in the simulation, where the proposed algorithm ESSA is seen converged after 50 iteration rounds in all three scenarios.

Fig. 3 discusses the channel utilization of the proposed ESSA algorithm in Scenario 2. Fig. 3(a) demonstrates the number of the channels in $\mathcal{C}$ which have the water-filling levels balanced across the users in $\mathcal{U}$. It can be found from the simulation results that when the SA capability M or the average queue backlog increases, the number of the channels with water-filling level balancing increases, which further reduces the power consumption. There is a drop for the performances of ESSA ($M = 10$) and ESSA ($M = 5$) with high average queue backlog, because a high average queue backlog is achieved with a large V. In this case, some channels may be turned off for saving energy due to the circuit power consumption. Fig. 3(b) demonstrates the number of the channels scheduled to transmit. When the average target delay is large, it is not necessary to use all channels for transmission, because the saved circuit power by using fewer channels dominates the increased transmit power, which verifies the performance drop in Fig. 3(a).

Fig. 4 illustrates the influence brought by the aggregation capability M and the average arrival rates. It is seen from Fig. 4(a) that the average power consumption is small with a large aggregation capability M. According to (27) in Theorem 1, the more channels participate in water-filling level balancing, the more power can be saved, which is consistent with the results in Fig. 3(a) that the number of the channels with water-filling level balancing is large for a large M. In Fig. 4(a), the average power consumption is small for a large M. The average power consumption decreases as the average queue backlog increases, because a large average queue backlog represents that the unit power cost is high.

The performance comparison between ESSA and baseline TOCA is performed in Fig. 4(a). It is seen that ESSA with $M \geq 2$ outperforms TOCA (which is just ESSA ($M = 1$)),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2721107, IEEE Transactions on Green Communications and Networking

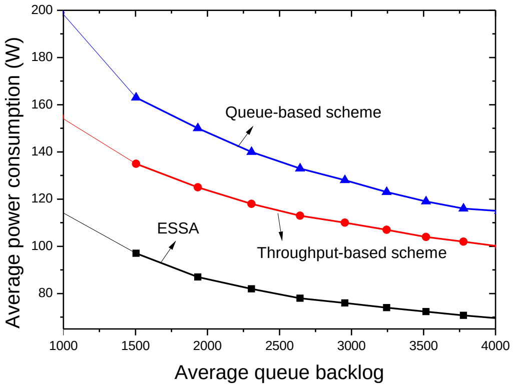


Fig. 5. The performance comparison

the performance gain is achieved by two reasons. First, the water-filling levels of the channels are balanced across users in ESSA. The larger number of the channels with the balanced water-filling levels leads to further reduction of the power consumption. Second, the circuit power is considered in ESSA to schedule an appropriate number of channels to minimize the total power consumption. In addition, the achieved performance gain increases as the SA capability M increases.

Fig. 4(b) evaluates the influence of different scenarios to ESSA algorithm. As for fixed aggregation capability $M = 5$, the power consumption of Scenario 1 is larger than than that of Scenario 2 and Scenario 3, which implies that the more concentrated the traffic flow is, the larger power cost it brings. Moreover, a large SA capability can potentially overcome this issue, e.g., the average power consumptions of ESSA ($M = 10$) in three scenarios are nearly the same. With a large M, the power consumption in different scenarios get close, which verifies that the SA capability brings robustness to the variation of incoming data rates.

Fig. 5 provides the performance comparison among the proposed scheme ESSA ($M = 5$) and the two baseline schemes, i.e., throughput-based scheme ($M = 5$) and queue-based scheme ($M = 5$) in Scenario 2. It is shown in the simulation results that ESSA outperforms the two baseline schemes. The throughput-based scheme does not care about the queue lengths of users. As a result, some users under good channel condition will receive excessive resources than needed, and the users under poor channel condition cannot be allocated enough resources. Therefore, the throughput-based scheme consumes more power to achieve the same average queue backlog as ESSA. The queue-based scheme is originally designed for the system with time invariant links, and hence only considers queue backlog, which consumes more power than the other two schemes. ESSA dynamically allocates the power according to both queue backlog and channel conditions, and the throughput-based scheme dynamically allocates the power according to channel conditions, so they are more energy efficient than the queue-based scheme. Note that the queue-based scheme may not be stable with a small average power consumption, e.g., the power is less than $105W$ in the simulation, which emphasizes the importance of ESSA.

VI. Conclusions

In this paper, we develop an analytical scheduling framework for delay-constrained energy efficient SA. Due to the practical hardware limitation, the conventional scheduling based on Lyapunov method cannot be adopted directly for delay-constrained SA. Also, the total power consumption varies according to the combination of channel allocation and needs to be taken into consideration for designing the energy efficient scheduling. To address the above two problems, we design the ESSA algorithm in two steps. First, we minimize the total power consumption for SA, including both the transmit power as well as the circuit power, by differentiated water-filling. Second, we propose an iterative Lyapunov optimization method to adjust the data rate and the channel allocation to minimize the power consumption with delay constraint. ESSA is proved to achieve a lower power consumption compared with the baseline scheme TOCA both theoretically and by simulation. Furthermore, we show the performance improvements of ESSA compared to the other two existing baseline algorithms.

Appendix A
Proof of Lemma 1

To prove that the partition method in Algorithm 1 always partitions the users into the conforming/nonconforming user sets uniquely, we can prove alternatively that the users in the nonconforming set cannot be further partitioned into another conforming set in which multiple uses adopt the same water-filling level.

Assume the users in the nonconforming set have different average arrival rate. There always exist a user $i \in \mathcal{U}^-$ whose rate satisfies $\lambda_i(t) > \lambda_j(t), i \neq j, j \in \mathcal{U}^-$. According to [8], we obtain $\mathbb{E}[r_i] = \lambda_i + \mathbb{E}[\epsilon]$ and $\mathbb{E}[\epsilon] = 0$. We can easily see that $\lambda_i > \mathbb{E}[M\frac{\sum_{l\in\mathcal{U}^-} r_l}{kM}] = M\frac{\sum_{l\in\mathcal{U}^-}\lambda_l}{kM}$. Therefore, user i cannot be in another conforming set. Similarly, we can find that all the users in the nonconforming set cannot be further partitioned into another conforming set.

Therefore, the users in the nonconforming set cannot form another conforming set and Lemma 1 holds.

Appendix B
Proof of Theorem 1

To analyze the minimum power consumption, we discuss the power consumption of the conforming users and the nonconforming users respectively.

A. Minimum Power Consumption of the Conforming Users

The power consumption of the conforming users can be obtained according to Lagrangian method. From (25), the transmission rate of channel j is

$$\begin{aligned} R_j(t) =& \frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|} - \sum_{k\in\mathcal{C}} \frac{1}{|\mathcal{C}|} \log_2 \sum_{i\in\mathcal{U}} b_k^i(t) S_k^i(t) \\ &+ \log_2 \sum_{i\in\mathcal{U}} b_j^i(t) S_j^i(t). \end{aligned} \quad (51)$$

With the channel set $\mathcal{C}$, the total SA circuit power consumption of the conforming users is

$$P_c(t) = |\mathcal{C}|P_1 + |\mathcal{U}|P_2. \tag{52}$$

Considering both the transmit power and the circuit power, we obtain the total power consumption of the conforming users as

$$\begin{aligned} P_{\mathcal{U}}(t) &= \sum_{j\in\mathcal{C}} P_j(t) + P_c(t) \\ &= |\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\prod_{k\in\mathcal{C}} \sum_{i\in\mathcal{U}} b_k^i(t) S_k^i(t)^{\frac{1}{|\mathcal{C}|}}} - \sum_{k\in\mathcal{C}} \frac{1}{\sum_{i\in\mathcal{U}} b_k^i(t) S_k^i(t)} \\ &\quad + |\mathcal{C}|P_1 + |\mathcal{U}|P_2. \end{aligned} \tag{53}$$

With a smaller $|\mathcal{C}|$, the system consumes more transmit power and less circuit power. Therefore, there is a tradeoff between the transmit power and the circuit power. It is necessary to optimize the transmitting channels $\mathcal{C}$.

The inactive channels are selected from the channel sets for the conforming users. In such a channel set, deactivating the channels with smaller gains always achieves a lower power consumption than those with larger gains, so we only need to figure out the optimal number of active channels $|\mathcal{C}|$ rather than which channels should be active. The active channels can be found easily by one-dimensional searching as follows:

1) Initialize a channel set in which all channels are active.
2) Deactivate the channel with the smallest $S_j^i(t)$ from all active channels if the total power consumption can be reduced.
3) Repeat Step 2 until the power consumption cannot be further reduced.

B. Minimum Power Consumption of the Nonconforming Users

For the nonconforming users, the main difference to the conforming users is that the power are allocated over the channels for a single user by water-filling approach. The power consumption of the nonconforming users is obtained by Lagrangian method.

With channel set $\mathcal{C}_i$, the SA circuit power consumption of the nonconforming user $i \in \mathcal{U}^-$ is

$$P_{c,i}(t) = |\mathcal{C}_i|P_1 + P_2. \tag{54}$$

Similar to the case with conforming users, there is a tradeoff between the transmit power and the circuit power, and it is necessary to optimize the transmitting channels $\mathcal{C}_i$. Similarly, the optimal number of active channels for each nonconforming user can be found easily by one-dimensional searching, where the inactive channels are selected from the channel sets for each nonconforming user separately.

Considering the power consumption of both the conforming users and the nonconforming users, we calculate the total minimum power consumption as

$$P(t) = P_{\mathcal{U}}(t) + \sum_{i\in\mathcal{U}^-} P_{\mathcal{U}^-,i}(t). \tag{55}$$

Substituting (53) and $P_{\mathcal{U}^-,i}(t)$ into (55), (27) can be obtained and Theorem 1 is proved.

Appendix C
Proof of Lemma 2

The objective function can be reduced to

$$\begin{aligned} \min_{\mathbf{b}(t)} Y(t) &= V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\prod_{j\in\mathcal{C}} \sum_{i\in\mathcal{U}} b_j^i(t) S_j^i(t)^{\frac{1}{|\mathcal{C}|}}} \\ &- \Bigg(\sum_{i=1}^{N} 1_{U_i(t)\geq Q}(t) \big(\omega e^{\omega(U_i(t)-Q)} + 2X_i(t) \big) \\ &+ \sum_{i=1}^{N} 1_{U_i(t)<Q}(t) \big(-\omega e^{\omega(Q-U_i(t))} + 2X_i(t) \big) \Bigg) \cdot \sum_{k\in\mathcal{C}} b_k^i(t) \frac{\sum_{l\in\mathcal{U}} r_l(t)}{|\mathcal{C}|}. \end{aligned} \tag{56}$$

Consider a channel j allocated to user k_1 initially, and three other users k_0, k_2, k_3 satisfying $f(k_3) > f(k_2) > f(k_1) > f(k_0)$. Channel j is firstly switched from user k_1 to user k_2, and there are the following 3 cases might happen.

- Case 1: Channel j is switched from user k_1 to user k_2.
- Case 2: Channel j is firstly switched from user k_1 to user k_2, and then to user k_3.
- Case 3: Channel j is firstly switched from user k_1 to user k_2, and then to user k_0.

After the switching, the average value of channel condition has changed. For simplicity, we denote $\prod_{j\in\mathcal{C}} \sum_{i\in\mathcal{U}} b_j^i(t) S_j^i(t)^{\frac{1}{|\mathcal{C}|}} = \overline{s}(t) S_j^{k_1}(t)^{1/\mathcal{C}}$ before switching and $\prod_{j\in\mathcal{C}} \sum_{i\in\mathcal{U}} b_j^i(t) S_j^i(t)^{\frac{1}{|\mathcal{C}|}} = \overline{s}^*(t)(S_j^{k_2}(t))^{1/\mathcal{C}}$ after switching. According to the initialization of the allocation vector $\mathbf{b}(t)$, we observe that $\prod_{j\in\mathcal{C}} \sum_{i\in\mathcal{U}} b_j^i(t) S_j^i(t)^{\frac{1}{|\mathcal{C}|}}$ reaches its maximal before switching, and decreases during the switching process. Thus, it is obtained that

$$\overline{s}(t) \geq \overline{s}^*(t). \tag{57}$$

Channel j is switched from user k_1 to user k_2 in Case 1, which is optimal. According to the switching rule $Y(t)|_{b_j^{k_1}(t)=1} - Y(t)|_{b_j^{k_2}(t)=1} > Y(t)|_{b_j^{k_1}(t)=1} - Y(t)|_{b_j^{k_3}(t)=1}$, we further obtain

$$V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\overline{s}(t) S_j^{k_2}(t)^{1/\mathcal{C}}} - V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\overline{s}(t) S_j^{k_3}(t)^{1/\mathcal{C}}} + \Delta U_{23}(t) < 0, \tag{58}$$

where $\Delta U_{23}(t)$ represents the difference between k_2 and k_3 of the last term of $Y(t)$ in Eq. (56).

Case 2 happens if and only if

$$Y(t)|_{b_j^{k_2}=1} - Y(t)|_{b_j^{k_3}=1} > 0, \tag{59}$$

where

$$\begin{aligned} &Y(t)|_{b_j^{k_2}(t)=1} - Y(t)|_{b_j^{k_3}(t)=1} = \\ &V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\overline{s}^*(t) S_j^{k_2}(t)^{1/\mathcal{C}}} - V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\overline{s}^*(t) S_j^{k_3}(t)^{1/\mathcal{C}}} + \Delta U_{23}(t). \end{aligned} \tag{60}$$

According to (57) and (58),

$$\begin{aligned} 0 &> V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\overline{s}^*(t) S_j^{k_2}(t)^{1/\mathcal{C}}} - V|\mathcal{C}| \frac{2^{\frac{\sum_{i\in\mathcal{U}} r_i(t)}{|\mathcal{C}|}}}{\overline{s}^*(t) S_j^{k_3}(t)^{1/\mathcal{C}}} + \frac{\overline{s}(t)}{\overline{s}^*(t)} \Delta U_{23}(t) \\ &\geq Y(t)|_{b_j^{k_2}(t)=1} - Y(t)|_{b_j^{k_3}(t)=1}. \end{aligned} \tag{61}$$

Therefore, Case 2 does not occur. Adopting a similar technique, neither does Case 3.

Appendix D
Proof of Theorem 2

We prove this theorem using a similar method to [8]. Since the average delay is guaranteed by the buffer partitioning approach, we obtain T_{ave} as [8]

$$T_{ave} \leq \frac{1+\gamma}{\omega}\log_2\left(2\frac{D+\frac{V}{N}h}{\omega\epsilon}\right), \tag{62}$$

where D and h are constants.

The minimum power function $\phi(\mathbf{r}(t))$ has the convexity property for all $\boldsymbol{\lambda}$ in the capacity region, it follows by the multi-dimensional Taylor theorem [30] that

$$\phi\big(\boldsymbol{\lambda}+\epsilon\boldsymbol{\Delta}\big) \leq \phi(\boldsymbol{\lambda}) + \sum_{i=1}^{N}\frac{\partial\phi(\boldsymbol{\lambda})}{\partial\lambda_i}\epsilon\boldsymbol{\Delta}_i + N\epsilon^2\kappa. \tag{63}$$

ESSA is an $1+\gamma$ approximation of the optimal algorithm, therefore, the power consumption is $1+\gamma$ times larger. It can be obtained that

$$E_{ave} \leq (1+\gamma)\phi(\boldsymbol{\lambda})+\frac{D}{V}+\frac{(1+\gamma)}{N}\sum_{i=1}^{N}\frac{\partial\phi(\boldsymbol{\lambda})}{\partial\lambda_i}\epsilon\boldsymbol{\Delta}_i+(1+\gamma)\epsilon^2\kappa. \tag{64}$$

As the users in the conforming set $\mathcal{U}$ balance their water-filling levels together, the transmission rate R_j for each channel $j\in\mathcal{C}$ follows that

$$\mathbb{E}[R_j] = \frac{\sum_{i\in\mathcal{U}}(\lambda_i)+\sum_{i\in\mathcal{U}}\left(\alpha_i^R-\alpha_i^L\right)}{|\mathcal{C}|}. \tag{65}$$

For user i in the nonconforming set $\mathcal{U}^-$, the transmission rate R_j for each channel $j\in\mathcal{C}_i$ follows that

$$\mathbb{E}[R_j] = \frac{\lambda_i+\alpha_i^R-\alpha_i^L}{|\mathcal{C}_i|}. \tag{66}$$

It can be obtained that

$$\Delta_i = \begin{cases} \frac{\sum_{i\in\mathcal{U}}\left(\alpha_i^R-\alpha_i^L\right)}{|\mathcal{C}|} & i\in\mathcal{U} \\ \frac{\alpha_i^R-\alpha_i^L}{|\mathcal{C}_i|} & i\in\mathcal{U}^-, \end{cases} \tag{67}$$

where $\alpha_i^R = \Pr[U_i(t)\geq Q]$, $\alpha_i^L = \Pr[U_i(t)<Q]$.

References

[1] Y. Wang, W. Wang, L. Chen and Z. Zhang, "Energy efficient scheduling for delay-constrained spectrum aggregation," *Proc. of IEEE Globecom 2016*, Dec. 2016.

[2] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger and F. Xu, "Overview of 3GPP LTE-advanced carrier aggregation for 4G wireless communications," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 122-130, Feb. 2012.

[3] W. Wang, Z. Zhang, A. Huang, "Spectrum aggregation: Overview and challenges," *Network Protocols and Algorithms*, vol. 2, no. 1, pp. 184 - 196, May 2010.

[4] R. Zhang, Z. Zheng, M. Wang, X. Shen and L. Xie, "Equivalent capacity analysis of LTE-advanced systems with carrier aggregation," *Proc. of IEEE ICC 2013*, pp. 6118-6122, Jun. 2013.

[5] M. Iwamura, K. Etemad, M. H. Fong, R. Nory and R. Love, "Carrier aggregation framework in 3GPP LTE-advanced," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 60-67, Jan. 2010.

[6] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Commu.*, vol. 9, no. 5, pp. 76-83, Mar. 2012.

[7] X. He and A. Yener, "On the energy-delay trade-off of a two-way relay network," *Inf. Sci. Syst.*, pp. 865-870, Apr. 2008.

[8] M. J. Neely , "Optimal energy and delay tradeoffs for multiuser wireless downlinks," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3095-3113, Jun. 2007.

[9] L. Wu, W. Wang, Z. Zhang and L. Chen, "A rollout-based joint spectrum sensing and access policy for cognitive radio networks with hardware limitations," *Proc. of IEEE Globecom 2012*, pp. 1277-1282, Dec. 2012.

[10] W. Wang, L. Wu, Z. Zhang and L. Chen, "Joint spectrum sensing and access for stable spectrum aggregation," *EURASIP J. Wireless Commun. Netw.*, 2015:130, pp. 1-14, May 2015.

[11] R. Ratasuk, D. Tolli and A. Ghosh, "Carrier aggregation in LTE-advanced," *Proc. of IEEE VTC 2010-Spring*, pp. 1-5, May 2010.

[12] F. Wu, Y. Mao, S. Leng and X. Huang, "A carrier aggregation based resource allocation scheme for pervasive wireless networks," *Proc. of IEEE DASC 2011*, pp. 196-201, Dec. 2011.

[13] H. Shajaiah, A. Abdel-Hadi and C. Clancy, "Utility proportional fairness resource allocation with carrier aggregation in 4G-LTE," *Proc. of IEEE Milcom 2013*, pp. 412-417, Jun. 2013.

[14] A. Abdelhadi and C. Clancy, "An optimal resource allocation with joint carrier aggregation in 4G-LTE," *Proc. of IEEE ICNC 2014*, pp. 138-142), Feb. 2015.

[15] H. Shajaiah, A. Abdel-Hadi and C. Clancy, "Spectrum sharing between public safety and commercial users in 4G-LTE," *Proc. of IEEE ICNC 2014*, pp. 674-679, Feb. 2014.

[16] R. Zhang, M. Wang, Z. Zheng, X. Shen and L. Xie, "Cross-layer carrier selection and power control for LTE-A uplink with carrier aggregation," *Proc. of IEEE GLOBECOM 2013*, pp. 4668-4673, Dec. 2013.

[17] F. Liu, K. Zheng, W. Xiang and H. Zhao, "Design and performance analysis of an energy-efficient uplink carrier aggregation scheme," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 197-207, Feb. 2014.

[18] L. Chen, C. Liu, X. Hong, C. X. Wang, J. Thompson and J. Shi, "Capacity and delay tradeoff of secondary cellular networks with spectrum aggregation," arXiv preprint, arXiv: 1612.08778, 2016.

[19] H. Lee, S. Vahid and K. Moessner, "A survey of radio resource management for spectrum aggregation in LTE-advanced," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 745-760, May. 2014.

[20] Y. L. Chung, L. J. Jang and Z. Tsai, "An efficient downlink packet scheduling algorithm in LTE-advanced systems with carrier aggregation," *Proc. of IEEE CCNC 2011*, pp. 632-636, Jan. 2011.

[21] G. Galaviz, D. Covarrubias and A. Andrade, "On a spectrum resource organization strategy for scheduling time reduction in carrier aggregated systems," *IEEE Commun. Lett.*, vol. 15, pp. 1202-1204, Nov. 2011.

[22] D. P. Bertsekas, *Dynamic Programming and Optimal Control, 3rd editon* Massachusetts: Athena Scientific, 2007.

[23] Y. Cui, Q. Huang and V. K. N. Lau, "Queue-aware dynamic clustering and power allocation for network MIMO systems via distributed stochastic learning," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1229C1238, Mar. 2011.

[24] B. Ji, G. R. Gupta, M. Sharma, X. Lin and N. B. Shroff "Achieving optimal throughput and near-optimal asymptotic delay performance in multichannel wireless networks with low complexity: a practical greedy scheduling policy," *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 880-893, Jun. 2015.

[25] K. I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. G. U. Garcia and Y. Wang, "Carrier aggregation for LTE-advanced: functionality and performance aspects," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 89-95, Jun. 2011.

[26] M. Calle and J. Kabara, "Measuring energy consumption in wireless sensor networks using GSP," *Proc. of IEEE PIMRC 2006*, pp. 1-5, Aug. 2006.

[27] W. A. Rosenkrantz, "Little's theorem, A stochastic integral approach," *Queueing systems*, vol. 12, no. 3, pp. 319-324, Jan. 1992.

[28] M. Kuczma, *An introduction to the theory of functional equations and inequalities: Cauchy equation and Jensen's inequality*, Springer, Oct. 2008.

[29] L. Lovasz, "On the Shannon capacity of a graph," *IEEE Trans. Inf. Theory*, vol. 25, no. 1, pp. 1-7, Dec. 1979.

[30] L. M. Graves, "Riemann integration and *Taylor* theorem in general analysis," *Amer. Math. Soc*, vol. 1, no. 29, pp.163-177, Mar. 1927.

[31] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lec. Commun. Netw.*, vol. 3, no. 1, pp. 1-211, May 2010.

[32] L. Georgiadis, M. J. Neely and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1-144, Apr. 2006.

[33] H. Wang, C. Rosa and K. Pedersen, "Performance of uplink carrier aggregation in LTE-advanced systems," *Proc. of IEEE VTC 2010*, pp. 1-5, Sep. 2010.

[34] H. C. Cho, M. S. Fadali, J. W. Lee, Y. J. Lee and K. S. Lee, "Lyapunov-based fuzzy queue scheduling for Internet routers," *International J. Control Auto. Syst.*, vol. 5, no. 3, pp. 317-323, Jun. 2007.

[35] G. Auer, O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading, E. Katranaras, M. Olsson, D. Sabella, P. Skillermark and W. Wajda, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," *EARTH*, pp. 1-69, Nov. 2010.

Yitu Wang (S'16) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2013. From August to November 2014, he was a visiting student with the University of Paris-Sud, Orsay, France. He is currently a Ph.D candidate at Zhejiang University, Hangzhou, China. His research interests mainly focus on stochastic optimization for cross-layer resource allocation in wireless networks and cache-enabled networks.

Wei Wang (S'08-M'10-SM'15) received the B.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, China, in 2004 and 2009, respectively. From 2007 to 2008, he was a Visiting Student with the University of Michigan, Ann Arbor, USA. From 2013 to 2015, he was a Hong Kong Scholar with the Hong Kong University of Science and Technology, Hong Kong. He is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, China. His research interests mainly focus on stochastic optimization for cross-layer resource allocation in wireless networks, and caching and computing in wireless networks. He is the Editor of the book entitled Cognitive Radio Systems, and serves as an Editor of the IEEE Access, Transactions on Emerging Telecommunications Technologies, and KSII Transactions on Internet and Information Systems.

Lin Chen (S'07-M'10) received his B.E. degree in Radio Engineering from Southeast University, China in 2002 and the Engineer Diploma from Telecom ParisTech, Paris in 2005. He also holds a M.S. degree of Networking from the University of Paris 6. He currently works as associate professor in the department of computer science of the University of Paris-Sud. He serves as Chair of IEEE Special Interest Group on Green and Sustainable Networking and Computing with Cognition and Cooperation, IEEE Technical Committee on Green Communications and Computing. His main research interests include modeling and control for wireless networks, distributed algorithm design and game theory.

Pan Zhou (S'07-M'14) is currently an associate professor with School of Electronic Information and Communications, Wuhan, P.R. China. He received his Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology (Georgia Tech) in 2011, Atlanta, USA. He received his B.S. degree in the Advanced Class of HUST, and a M.S. degree in the Department of Electronics and Information Engineering from HUST, Wuhan, China, in 2006 and 2008, respectively. He held honorary degree in his bachelor and merit research award of HUST in his master study. He was a senior technical member at Oracle Inc, America during 2011 to 2013, and worked on Hadoop and distributed storage system for big data analytics at Oralce could Platform. His current research interest includes: big data analytics and machine learning, security and privacy, and information networks.

Zhaoyang Zhang (M'10) received the Ph.D. degree in communication and information systems from Zhejiang University, Hangzhou, China, in 1998. He is currently a Full Professor with the College of Information Science and Electronic Engineering, Zhejiang University. He has coauthored more than 150 refereed international journal and conference papers, as well as two books in his areas of interest. His research interests are mainly focused on information theory and coding theory, signal processing techniques, and their applications in wireless communications and networking.

Dr. Zhang coreceived three conference Best Paper Awards/Best Student Paper Award. He is currently serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, IET Communications, and several other international journals. He has served as a Technical Program Committee Cochair or a Symposium Cochair for many international conferences, such as the 2013 International Conference on Wireless Communications and Signal Processing and the 2014 IEEE Global Communications Conference Wireless Communications Symposium.