On Effective Capacity and Effective Energy Efficiency in Relay-Assisted Wireless Networks

Jiping Li, *Member, IEEE*, Yaoming Ding, *Senior Member, IEEE*, Qiang Ye, *Member, IEEE*, Ning Zhang, *Member, IEEE*, and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—The tremendous proliferation of smart phones and other smart devices has spurred the explosive growth of delay-sensitive high-rate multimedia services, which significantly increases energy consumption in wireless cellular networks. Therefore, it is an important yet challenging issue to carry out resource allocation for cellular networks to meet the quality of service requirement in terms of high data rate and, at the same time, achieve high energy efficiency. In this paper, two resource allocation approaches are proposed, where user selection and power allocation are jointly considered, to maximize effective capacity (EC) and effective energy efficiency (EEE), respectively, for a relay-assisted downlink of cellular networks. Channel estimation errors are taken into account to make the proposed approaches more practical. Moreover, to solve the EC maximization problem efficiently, we develop an optimal algorithm based on Lagrange method and Karush–Kuhn–Tucker conditions, and derive closed-form optimal solutions for user selection and power allocation at both the base station and relays. For the EEE maximization problem, an iterative algorithm based on the Dinkelbach method is developed to find the optimal solutions. Theoretical analysis and simulation results are presented to demonstrate the performance of the proposed approaches in terms of EC, EEE, and delay bound violation probability.**

***Index Terms*—Effective capacity, effective energy efficiency, downlink transmission with relays, channel estimation error, resource allocation, QoS.**

I. Introduction

THE tremendous popularity of smart phones and electronic tablets has spurred the explosive growth of high-rate multimedia services for next generation wireless cellular networks. It is reported that the data traffic volume of mobile broadband networks is expected to increase 89 times in year 2020 as compared to the traffic volume in 2010 [1]. Therefore, to provide better service coverage for an enlarged network scale with more mobile users, a relay-assisted cellular network is imperative especially for non line-of-sight communications, where a set of relay nodes are placed between base stations (BSs) and end users to amplify and forward signals for a better received signal-to-noise (SNR) ratio for end-to-end packet transmissions. For a cellular network, the relay mechanism not only enlarges the network service coverage, but also reduces signal attenuation and improves performance for cell-edge users [9]. However, the considerable increase of both the number of mobile users and the data traffic volume on each user continues to pose technical challenges on resource allocation in the relay-assisted cellular network: First, with limited radio resources, how to consistently satisfy the quality of service (QoS) requirements, in terms of data rate and packet transmission delay, for an increasing number of users should be considered in allocating resources; Second, BSs and relays are the main sources of energy depletion for the whole network, which contribute to over 70% of network operator electricity consumption [2], [3]. Therefore, how to achieve effective energy conservation for both BSs and relays is of paramount importance from the perspective of service operators, which becomes more challenging when the aggregate traffic load gets higher in next-generation cellular networks [4], [5].

In literature, extensive research works focus on how to achieve energy efficiency for wireless cellular networks, which can be mainly classified into two categories based on the adopted performance metrics: Shannon capacity per unit energy consumption [6]–[12] and effective capacity (EC) per unit energy consumption. The Shannon capacity achieves a theoretical performance upper bound, which cannot be obtained in practical systems. Moreover, the classic Shannon capacity has no consideration of user's delay QoS requirements in practical applications. Therefore, it is mostly used for evaluating the performance of delay insensitive services [13], [14]. For supporting delay-sensitive applications, such as online gaming, video conferencing and autonomous driving, the network capacity analysis using Shannon formulation may not be accurate. Moreover, due to the time-varying nature of wireless channels, it is difficult to provide a deterministic delay guarantee for services over wireless networks. Therefore, to facilitate design for delay-sensitive applications, effective capacity, defined as the maximum constant arrival rate supported by a given wireless channel to maintain a satisfied QoS, has been proposed and gained more and more attentions, providing a statistical delay provisioning by guaranteeing a small delay violation probability for packet transmissions [15].

Manuscript received June 21, 2017; revised October 3, 2017; accepted December 3, 2017. Date of publication December 25, 2017; date of current version May 14, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61370223 and in part by the Natural Science Foundation of Hubei Province of China under Grant 2014CFB577. The review of this paper was coordinated by Dr. P. Lin. *(Corresponding author: Qiang Ye.)*

J. Li and Y. Ding are with the School of Computer and Information Science, Hubei Engineering University, Xiaogan 432000, China (e-mail: oucljp@aliyun.com; xgdym2015@aliyun.com).

Q. Ye and W. Zhuang are with the University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: q6ye@uwaterloo.ca; wzhuang@uwaterloo.ca).

N. Zhang is with the Department of Computing Science, Texas A&M University-Corpus Christi, Corpus Christi, TX 78412 USA (e-mail: Ning.Zhang@tamucc.edu).

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

Digital Object Identifier 10.1109/TVT.2017.2787114

0018-9545 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

By using the EC model, many resource allocation schemes are proposed recently under delay violation probability constraints [14], [16]–[22]. In [14], a power allocation scheme for EC maximization is investigated in a point-to-point communication system with a flat-fading channel and with statistical delay and EEE requirements. In the proposed scheme, maximizing EC can be achieved at an average input power level where the EEE constraint is satisfied at equality. In [16], a power allocation scheme is proposed to jointly optimize link-layer energy efficiency (EE) and EC of a point-to-point Rayleigh flat-fading channel with a delay-outage probability constraint. To solve the multi-objective optimization problem, an importance weight is introduced to change the priority level of EE and EC, and to convert the multi-objective optimization problem into a single-objective optimization problem which can be solved by using fractional programming. In [17], an energy-efficient and delay-aware cross-layer resource allocation scheme is proposed to maximize the EEE of a wireless point-to-point link. Different from previous schemes, an average power consumption model allows for the probability of emptying the buffer during the transmission timeframe for the data link layer. In [18], a closed-form expression of EC is presented for a flat fading channel between a source-destination transmission pair with the help of multiple relays. Besides, the effects of some relaying schemes such as best relay selection (BRS) and distributed space-time coding (DSTC) on the EC performance are investigated. In [19], an energy-efficient design for a downlink orthogonal frequency division multiple access (OFDMA) network with EC-based delay provisioning is investigated, in which the tradeoff between EE and delay, the relationship between spectral-efficient design and energy-efficient design, and the impact of system parameters such as circuit power and delay exponents on the overall performance are discussed. In [20], an optimal energy-efficient power allocation scheme for a point-to-point multi-carrier link over a frequency-selective fading channel under a target delay-outage probability constraint is investigated, and the global optimal solution is derived by using fractional programming due to the quasi-concavity of the objective function. In [21], a joint optimization of link-layer EE and EC in a Nakagami-m point-to-point fading channel under a delay-outage probability constraint and an average transmit power constraint is investigated. A normalized multi-objective optimization problem is formulated and transformed into a single-objective optimization problem by applying the weighted sum method, and is then solved by using Charnes-Cooper transformation and Karush-Kuhn-Tucker (KKT) conditions. In [22], the authors investigate resource allocation for LTE-A relay networks under statistical QoS constraints, and derive the optimal subcarrier and power allocation strategies to maximize the EC of the underlying LTE-A relay systems by dual decomposition. Moreover, a low-complexity suboptimal scheme is developed through optimizing the subcarrier and power allocation individually. In [23], fixed and adaptive source and relay power allocation is investigated for a 3-node buffer-aided relaying network with statistical QoS constraint in terms of maximum acceptable end-to-end queue-length bound outage probability. In [24], a novel approach for exact EC performance analysis of a CSI-assisted AF multihop system over arbitrary and correlated fading channels is presented, and single integral expression for the EC is derived by using a moment generating function.

Although extensive research works evaluate the EE using the EC model, most of them focus on the network model of a single source-destination transmission pair with multiple relays placed in between for packet forwarding in a cooperative mode. Limited works study the resource allocation for a multi-user multi-relay downlink in cellular networks. In addition, delay requirements for users near the cell edge are rarely considered, and perfect channel state information is assumed in most of the existing works. However, perfect channel state information is difficult to acquire in a real system due to the highly unpredictable nature of wireless channels and inevitable channel estimation errors. Therefore, channel estimation errors should be taken into account in conducting resource allocation to avoid an undesired increase of outage probability and transmit power wastage. In this paper, we investigate the resource allocation for a single-cell multi-relay multi-user downlink of independent frequency-flat Rayleigh fading channels. Amplify-and-forward (AF) relays are adopted due to its low complexity in implementation. Specifically, we focus on deriving the optimal power allocation and user selection for maximizing EC and EEE, respectively, with the relay link power and delay bound constraints. The main contributions of this paper are summarized as follows:

1) An effective end-to-end signal-to-noise ratio (SNR) for a two-hop transmission link via an AF relay node is derived, considering channel estimation errors, which is the first-step towards deriving a more complex end-to-end SNR for a multi-hop transmission link via multiple relays;
2) We formulate an EC maximization problem with statistical QoS constraints and power constraints. By using Lagrangian function and KKT conditions, we solve the original problem and derive a closed-form optimal power allocation solution for both BS and relays and a closed-form user selection policy for near-cell-edge users;
3) An EEE maximization problem is formulated. We transform the original quasi-concave fractional optimization problem into a concave subtractive optimization problem. Then, the Dinkelbach method is used to solve the problem and derive the optimal power allocation and user selection. The Lagrange multipliers related to the delay and power constraints are also optimized via a sub-gradient method;
4) We investigate the effect of channel estimation errors on both EC and EEE, and the effect of circuit power consumption on the EEE.

The remainder of the paper is organized as follows. The system model and problem formulations are presented in Section II. In Section III, we discuss how to solve the EC maximization problem to derive the optimal power allocation and user selection under the statistical QoS constraint and the transmit power constraints on the BS and relays. In Section IV, we propose mutually cooperative algorithms to obtain optimal power allocation and user selection for maximizing the EEE under the relay power constraint and the delay requirement. Simulation results are presented in Section V to evaluate the performance

TABLE I
SUMMARY OF NOTATIONS

Notations	Description
h_{sr_i}	channel gains from BS to the ith relay
$h_{r_iu_k}$	channel gains form the ith relay to the kth user
$\hat{h}_{r_iu_k}$	channel gain based on MMSE estimation
$e_{r_iu_k}$	estimation error
σ_e^2	variance of the estimation error
p_{sr_i}	transmit power at the BS
$p_{r_iu_k}$	transmit power for the kth user at the ith relay
$n_{sr_i}, n_{r_iu_k}$	receiving noise at the ith relay and the kth user
N_0	variance of the random variables n_{sr_i}, $n_{r_iu_k}$
β	amplification factor for an AF relay
ρ_{ik}	effective receiving SNR of kth user via the ith relay
r_{ik}	Shannon capacity of the two-hop transmission link
$D_{ik}, D_{ik}^{\max}$	end-to-end transmission delay, maximum delay bound
p_{ik}^o	delay bound violation probability
ε_{ik}	probability of a nonempty queue
θ_{ik}	QoS exponent
A_{ik}	the packet arrival rate
$EC_{ik}(\theta_{ik})$	effective capacity with the QoS exponent θ_{ik}
M_i	number of users in the ith cluster
ς_s, ς_r	reciprocal of drain efficiency of power amplifiers
p_s^C, p_r^C	power consumption on circuit blocks at BS and relay
P_i^t	total power consumption for the ith cluster
x_{ik}	binary variable indexing user selection
$\tilde{x}_{ik}$	a real variable within the interval $[0, 1]$
η_i^{EEE}	effective energy efficiency in the ith cluster
$p_{\max}$	allowable source and relay transmit power limit

of the proposed solutions. Section VI concludes this work and discusses further research directions. Table I summarizes important notations.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we first present the system model, and then formulate the EC maximization problem and the EEE maximization problem, respectively, with the power constraint on the relay link and the delay requirement on transmission to each mobile user.

A. System Model

Consider a relay-assisted downlink in a single-cell cellular network, shown in Fig. 1, where the BS is placed at the center of the cell to provide the communication coverage for the entire cell, and a group of end users associated with the BS are randomly distributed within the cell. Some users located near the cell edge area, depicted by the shaded area in Fig. 1, experience more attenuated SNR for the downlink packet reception than other users due to a longer communication distance from the BS, and are thus denoted as *near-cell-edge users*. To improve the receiving SNR especially for near-cell-edge users, the network coverage area is logically divided into several virtual clusters, with an AF relay placed at a fixed location in each cluster. Each AF relay is responsible for amplifying and forwarding the received data to one of the end users in its cluster at any given time instant [25]. Downlink transmissions in separate clusters are assumed to take place simultaneously using orthogonal channels. Therefore, the transmission interference among different clusters is neglected.

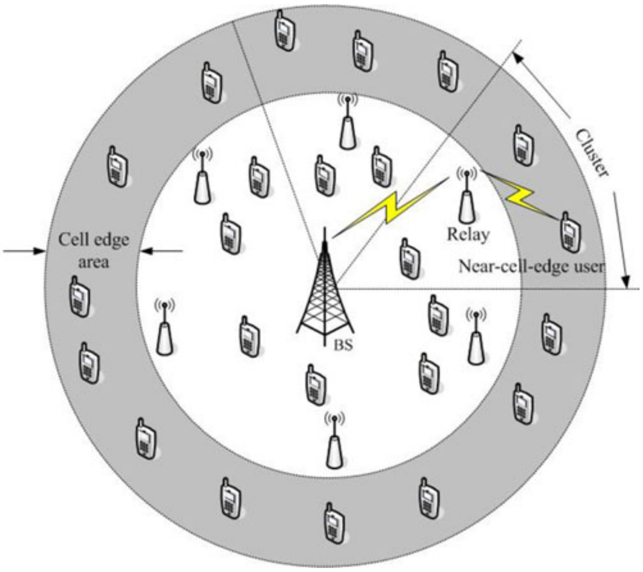


Fig. 1. A relay-assisted downlink cellular network.

Consider the worst case SNR guarantee for near-cell-edge users by assuming that the receiving SNR for a direct link transmission from the BS to a near-cell-edge user is unacceptable. Therefore, the end-to-end SNR can be enhanced by a two-hop transmission, with an AF relay node amplifying the received signal from the BS in the first hop and retransmitting the signal to the destination near-cell-edge user in the second hop.We assume that downlink wireless channels from the BS to a relay and from a relay to an end user are independent frequency-flat Rayleigh fading channels. Since the BS and relays are fixed in locations and the heights of their transmitting antennas are normally high, shadowing effect is not considered in the channel modeling for the first hop packet transmissions. In the second transmission hop, because of the user mobility and possible communication obstacles in between relays and end users, a combined path loss and shadowing model should be used [26].

Suppose that a minimum mean square error (MMSE) based channel estimation method [27] is adopted by near-cell-edge users to estimate relay-to-user channel gains. Channel gains from the BS to the ith relay in the first hop, and from the ith relay to the kth user in the second hop are denoted by h_{sr_i} and $h_{r_iu_k}$, respectively. When channel estimation errors are taken into account, the channel gain from the ith relay to the kth user can be expressed as [29], [30]

$$h_{r_iu_k} = \hat{h}_{r_iu_k} + e_{r_iu_k} \tag{1}$$

where $\hat{h}_{r_iu_k}$ is the channel gain based on MMSE estimation, and $e_{r_iu_k}$ is the estimation error which is uncorrelated with the estimated channel gain [28]. We assume that $e_{r_iu_k}$ behaves as additive white Gaussian noise (AWGN) with variance denoted by σ_e^2. The received signal strength by the ith relay from the BS for the first hop, denoted by y_{sr_i}, and the received signal strength at the kth user from the ith relay for the second hop are given, respectively, by [29], [30]

$$y_{sr_i} = \sqrt{p_{sr_i}} h_{sr_i} s + n_{sr_i} \tag{2}$$

and

$$\begin{aligned} y_{r_iu_k} &= \beta y_{sr_i} h_{r_iu_k} + n_{r_iu_k} \\ &= \beta(\sqrt{p_{sr_i}} h_{sr_i} s + n_{sr_i}) h_{r_iu_k} + n_{r_iu_k} \end{aligned} \tag{3}$$

where p_{sr_i} is the transmit power at the BS, s denotes the signal symbol transmitted from the BS, n_{sr_i} and $n_{r_iu_k}$ are the receiving noises at the ith relay and the kth user, respectively. The noise components n_{sr_i} and $n_{r_iu_k}$ are zero-mean complex Gaussian random variables with the same variance N_0, i.e., $n_{sr_i}, n_{r_iu_k} \sim CN(0, N_0)$, $\beta = \sqrt{\frac{p_{r_iu_k}}{p_{sr_i}|h_{sr_i}|^2 + N_0}}$ is the amplification factor for an AF relay, where $p_{r_iu_k}$ is the transmit power at the ith relay.

By substituting (2) into (3), the received signal strength for the kth user in the ith cluster (indicating the cluster associated with the ith relay) can be expressed as

$$\begin{aligned} y_{r_iu_k} &= \beta\sqrt{p_{sr_i}} h_{sr_i} \hat{h}_{r_iu_k} s + \beta\sqrt{p_{sr_i}} h_{sr_i} e_{r_iu_k} s \\ &\quad + \beta \hat{h}_{r_iu_k} n_{sr_i} + \beta e_{r_iu_k} n_{sr_i} + n_{r_iu_k}. \end{aligned} \tag{4}$$

By setting $s = 1$ in (4), the message part, error part and the noise part of the received signal can be extracted. Due to the mutual independence among $\hat{h}_{r_iu_k}$, $e_{r_iu_k}$, n_{sr_i} and $n_{r_iu_k}$, the effective end-to-end receiving SNR for the two-hop transmission link from the BS to the kth user via the ith relay is given by

$$\rho_{ik} = \frac{p_{sr_i}\phi_1 p_{r_iu_k}\phi_2}{1 + p_{sr_i}\phi_1 p_{r_iu_k}\phi_3 + p_{sr_i}\phi_1 + p_{r_iu_k}(\phi_2 + \phi_3)} \tag{5}$$

where $\phi_1 = \frac{|h_{sr_i}|^2}{N_0}$, $\phi_2 = \frac{|\hat{h}_{r_iu_k}|^2}{N_0}$ and $\phi_3 = \frac{\sigma_e^2}{N_0}$. If no channel estimation error is considered, i.e. $\phi_3 = 0$, (5) can be simplified as

$$\rho_{ik} = \frac{p_{sr_i}\phi_1 p_{r_iu_k}\phi_2}{1 + p_{sr_i}\phi_1 + p_{r_iu_k}\phi_2}. \tag{6}$$

Assuming a sufficiently high SNR value at the receiver, ρ_{ik} in (5) can be approximated by

$$\rho_{ik} \approx \frac{p_{sr_i}\phi_1 p_{r_iu_k}\phi_2}{p_{sr_i}\phi_1 p_{r_iu_k}\phi_3 + p_{sr_i}\phi_1 + p_{r_iu_k}(\phi_2 + \phi_3)}. \tag{7}$$

Therefore, the Shannon capacity of the two-hop transmission link from the BS to the kth user via the ith relay can be express as [31]

$$r_{ik} = \frac{1}{2}\log_2(1 + \rho_{ik}). \tag{8}$$

where the factor $\frac{1}{2}$ accounts for the fact that two time slots are required for a two-hop AF transmission.

B. QoS Provisioning

Due to the time-varying nature of wireless channels, it is difficult to guarantee the deterministic QoS for each packet transmission. Alternatively, we choose to provide statistical QoS satisfaction for each individual user in terms of guaranteeing certain delay-bound violation probability for higher resource utilization. To do so, we model the end-to-end communication system as a queueing system, where packets are generated constantly at the BS, and then transmitted via two-hop links to the kth user in the ith cluster. For a dynamic queueing system with stationary and ergodic arrival and service processes, the delay bound violation probability, i.e., the probability that the end-to-end transmission delay, D_{ik}, exceeding a maximum delay bound $D_{ik}^{\max}$, can be computed as [16]

$$p_{ik}^o = Pr\left\{D_{ik} \geq D_{ik}^{\max}\right\} \approx \varepsilon_{ik} e^{-\theta_{ik} A_{ik} D_{ik}^{\max}} \tag{9}$$

where $\varepsilon_{ik} \approx \frac{A_{ik}}{\lim_{\theta_{ik}\to 0} EC_{ik}(\theta_{ik})}$ denotes the probability of a non-empty queue and can be approximated by the ratio of the constant arrival rate to the average service rate. θ_{ik} is the QoS exponent, representing the exponentially decaying rate of the QoS violation probability. In (9), A_{ik} denotes the constant packet arrival rate destined for the kth user in the ith cluster, and $EC_{ik}(\theta_{ik})$ is the effective capacity of the wireless channels from the BS to the kth user via the ith relay with the QoS exponent θ_{ik}. Equation (9) shows that if a source requires a delay-bound violation probability of at most p_{ik}^o, it needs to limit its data rate to a maximum of A_{ik}.

C. Effective Capacity and Effective Energy Efficiency

The effective capacity is defined as the maximum constant arrival rate supported by a wireless channel so that the QoS requirements are guaranteed. The EC from the BS to the kth user via the ith relay can be expressed as [16]

$$EC_{ik}(p_{sr_i}, p_{r_iu_k}) = -\frac{1}{\theta_{ik}} \ln\left[E\left(e^{-\frac{1}{2}\theta_{ik}\log_2(1+\rho_{ik})}\right)\right] \tag{10}$$

where $E(\cdot)$ denotes the expectation operator, and $\theta_{ik}(>0)$ indicates the exponential decay rate of the QoS violation probability. Large and small values of θ_{ik} correspond to fast and slow decaying rates, respectively, indicating stringent and loose QoS requirements.

On the other hand, total energy consumption for the downlink transmission consists of transmit power from the BS and relays and the energy consumed on circuit block at the transceivers, such as analog-to-digital converter, digital-to-analog converter, synthesizer and mixer [32]. Therefore, the total power consumption, P_i^t, for the ith cluster is expressed as

$$P_i^t = \sum_{k=1}^{M_i}\left((\varsigma_s p_{sr_i} + \varsigma_r p_{r_iu_k}) + \left(p_s^C + p_r^C\right)\right) \tag{11}$$

where M_i is the number of users in the ith cluster, ς_s and ς_r are the reciprocal of drain efficiency of power amplifiers in the BS and relays, p_s^C and p_r^C are power consumption on circuit blocks in all the transmitters and receivers, respectively.

The effective energy efficiency in the ith cluster, η_i^{EEE}, for the downlink relay-assisted system is defined as the ratio of the sum of selected users' EC over the total energy consumption in the cluster, given by

$$\eta_i^{EEE} = \frac{\sum_{k=1}^{M_i} x_{ik} EC_{ik}}{\sum_{k=1}^{M_i} \left((\varsigma_s p_{sr_i} + \varsigma_r p_{r_i u_k}) + (p_s^C + p_r^C) \right)} \tag{12}$$

where x_{ik} is a binary variable for user selection which equals to 1 if the kth user is selected and 0 otherwise.

D. Problem Formulation

1) Effective Capacity Maximization: For the downlink multi-relay multi-user cellular system, we first formulate an EC maximization problem with statistical delay guarantee for the ith virtual cluster as P1.

$$\mathbf{P1}: \max_{p_{sr_i}, p_{r_i u_k}, x_{ik}} \sum_{k=1}^{M_i} EC_{ik}$$

$$\text{s.t.} \begin{cases} C1: x_{ik} \in \{0,1\}, & k=1,2,\ldots,M_i \\ C2: \sum_{k=1}^{M_i} x_{ik} = 1, & \forall i \\ C3: EC_{ik}(p_{sr_i}, p_{r_i u_k}, x_{ik}) \geq A_{ik}, & k=1,2,\ldots,M_i \\ C4: p_{sr_i} + p_{r_i u_k} \leq p_{\max}, & k=1,2,\ldots,M_i \\ C5: p_{sr_i} \geq 0,\ p_{r_i u_k} \geq 0, & k=1,2,\ldots,M_i. \end{cases}$$

In P1, constraints $C1$ and $C2$ indicate that the relay node can only relay traffic for one user in the cluster at a time; $C3$ guarantees the delay bound violation probability as p_{ik}^o; $C4$ is to limit the transmit power consumption on both the BS and the relay, where $p_{\max}$ denotes allowable source and relay transmit power limit.

2) Effective Energy Efficiency Maximization: In P1, we aim at maximizing the EC without taking into account the power consumption by the BS and the relay. Next, we aim at designing a resource allocation scheme by maximizing the effective energy efficiency as in P2, where the set of decision variables and constraints are the same as those in P1.

$$\mathbf{P2} \max_{p_{sr_i},\, p_{r_i u_k},\, x_{ik}} \eta_i^{EEE}$$

$$\text{s.t.} \begin{cases} C1: x_{ik} \in \{0,1\}, & k=1,2,\ldots,M_i \\ C2: \sum_{k=1}^{M_i} x_{ik} = 1, & \forall i \\ C3: EC_{ik}(p_{sr_i},\, p_{r_i u_k}, x_{ik}) \geq A_{ik}, & k=1,2,\ldots,M_i \\ C4: p_{sr_i} + p_{r_i u_k} \leq p_{\max}, & k=1,2,\ldots,M_i \\ C5: p_{sr_i} \geq 0,\ \ p_{r_i u_k} \geq 0, & k=1,2,\ldots,M_i. \end{cases}$$

III. Optimal Power Allocation and User Selection for Maximizing EC

In this section, we present how to solve P1. Since P1 is a combinatorial integer programming problem, the branch-and-bound method can be used. However, the computational complexity increases exponentially with the problem size. To make the problem tractable, we relax the binary variable x_{ik} in P1 as a real number $\tilde{x}_{ik}$ within the interval $[0,1]$. Then, we formulate a transformed problem of P1 as P1$'$.

$$\mathbf{P1'} \min_{p_{sr_i}, p_{r_i u_k}, \tilde{x}_{ik}} \sum_{k=1}^{M_i} -EC_{ik}$$

$$\text{s.t.} \begin{cases} C1: \tilde{x}_{ik} \in [0,1], & k=1,2,\ldots,M_i \\ C2: \sum_{k=1}^{M_i} \tilde{x}_{ik} = 1, & \forall i \\ C3: A_{ik} \leq EC_{ik}(p_{sr_i}, p_{r_i u_k}, \tilde{x}_{ik}), & k=1,2,\ldots,M_i \\ C4: p_{sr_i} + p_{r_i u_k} \leq p_{\max}, & k=1,2,\ldots,M_i \\ C5: p_{sr_i} \geq 0,\ p_{r_i u_k} \geq 0, & k=1,2,\ldots,M_i. \end{cases}$$

Since P1$'$ can be proved as a convex optimization problem with respect to p_{sr_i}, $p_{r_i u_k}$ and $\tilde{x}_{ik}$ (see Appendix), we solve the problem to get the optimal power allocation and user selection in two steps: In the first step, we suppose that the kth user is selected as the destination for the ith relay. Thus, the Lagrangian function for P1$'$ can be expressed as

$$\begin{aligned} &L(p_{sr_i}, p_{r_i u_k}, \lambda_{1k}, \lambda_{2k}, \lambda_3, \lambda_{4k}) \\ &= -\sum_{k=1}^{M_i} EC_{ik}(p_{sr_i}, p_{r_i u_k}) \\ &\quad + \sum_{k=1}^{M_i} \lambda_{1k} \left(A_{ik} - EC_{ik}(p_{sr_i}, p_{r_i u_k}) \right) \\ &\quad + \sum_{k=1}^{M_i} \lambda_{2k} (p_{sr_i} + p_{r_i u_k} - p_{\max}) \\ &\quad - \lambda_3 p_{sr_i} - \sum_{k=1}^{M_i} \lambda_{4k} p_{r_i u_k} \end{aligned} \tag{13}$$

where $\lambda_{1k}, \lambda_{2k}, \lambda_3, \lambda_{4k}$ are Lagrangian multipliers. The optimal transmit power at the BS and the ith relay can be achieved by differentiating $L(p_{sr_i}, p_{r_i u_k}, \lambda_{1k}, \lambda_{2k}, \lambda_3, \lambda_{4k})$ with respect to p_{sr_i} and $p_{r_i u_k}$. In the second step, based on the obtained optimal power allocation $p_{sr_i}^*$ and $p_{r_i u_k}^*$, the optimal user selection can be determined as

$$x_{ik} = \begin{cases} 1, & \text{if } k = \arg\max_m EC_{im}\left(p_{sr_i}^*, p_{r_i u_k}^*\right) \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The KKT conditions for P1$'$ are given by

$$\lambda_{1k}^* \left(A_{ik} - EC_{ik}\left(p_{sr_i}^*, p_{r_i u_k}^*\right) \right) = 0 \tag{15a}$$

$$\lambda_{2k}^* \left(p_{sr_i}^* + p_{r_i u_k}^* - p_{\max} \right) = 0 \tag{15b}$$

$$\lambda_3^* p_{sr_i}^* = 0, \quad \lambda_{4k}^* p_{r_i u_k}^* = 0 \tag{15c}$$

$$A_{ik} \leq EC_{ik}(p_{sr_i}^*, p_{r_i u_k}^*) \tag{15d}$$

$$p_{sr_i}^* + p_{r_i u_k}^* \leq p_{\max} \tag{15e}$$

$$p_{sr_i}^* \geq 0, \quad p_{r_i u_k}^* \geq 0 \tag{15f}$$

$$\lambda_{1k}^* \geq 0,\ \lambda_{2k}^* \geq 0\, \lambda_3^* \geq 0,\ \lambda_{4k}^* \geq 0 \tag{15g}$$

$$\frac{\partial L}{\partial p_{sri}^{*}} = -\frac{E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\frac{\theta_{ik}\rho_{ik}^{'}|_{p_{sr_i}^{*}}}{2\ln 2\left(1+\rho_{ik}^{*}\right)}\right)}{\theta_{ik}E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\right)} - \frac{\lambda_{1k}^{*}E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\frac{\theta_{ik}\rho_{ik}^{'}|_{p_{sr_i^*}}}{2\ln 2\left(1+rho_{ik}^{*}\right)}\right)}{\theta_{ik}E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\right)} + \lambda_{2k}^{*} - \lambda_{3}^{*} = 0 \tag{16}$$

and

$$\frac{\partial L}{\partial p_{r_i u_k}^{*}} = -\frac{E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\frac{\theta_{ik}\rho_{ik}^{'}|_{p_{r_i u_k}^{*}}}{2\ln 2\left(1+\rho_{ik}^{*}\right)}\right)}{\theta_{ik}E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\right)} - \frac{\lambda_{1k}^{*}E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\frac{\theta_{ik}\rho_{ik}^{'}|_{p_{r_i u_k}^{*}}}{2\ln 2\left(1+\rho_{ik}^{*}\right)}\right)}{\theta_{ik}E_{\gamma}\left(e^{-\frac{1}{2}\theta_{ik}\log 2\left(1+\rho_{ik}^{*}\right)}\right)} + \lambda_{2k}^{*} - \lambda_{4k}^{*} = 0 \tag{17}$$

where $(\cdot)^*$ represents the value of corresponding variable at the optimal point, (15a)–(15c) are complementary slackness conditions, (15d)–(15g) are feasibility conditions, and (16)–(17) are stationary conditions.

If transmissions from the BS to the near-cell-edge users are successful via relays, the transmit power p_{sr_i} and $p_{r_i u_k}$ should be both positive. Therefore, we have $\lambda_3^* = \lambda_{4k}^* = 0$. From (16), (17) and (15b), we have $\lambda_{2k}^* > 0$, and $p_{sr_i}^* + p_{r_i u_k}^* = p_{\max}$. Then, we derive $p_{sr_i}^*$ as

$$p_{sr_i}^{*} = \sqrt{\frac{\phi_1}{\phi_2+\phi_3}}p_{r_i u_k}^{*}. \tag{18}$$

Finally, we obtain the optimal solutions

$$p_{sr_i}^{*} = \frac{(\phi_2+\phi_3)-\sqrt{\phi_1(\phi_2+\phi_3)}}{\phi_2+\phi_3-\phi_1}p_{\max} \tag{19}$$

and

$$p_{r_i u_k}^{*} = \frac{\sqrt{\phi_1(\phi_2+\phi_3)}-\phi_1}{\phi_2+\phi_3-\phi_1}p_{\max}. \tag{20}$$

Detailed steps in finding optimal transmission power $p_{sr_i}^*$ and $p_{r_i u_k}^*$ at the BS and the ith relay, respectively, are given in Algorithm 1.

IV. Optimal Power Allocation and User Selection for Maximizing EEE

In this section, we solve P2 to obtain the optimal power allocation and user selection for maximizing EEE. As proved in Section III, the numerator of the objective function of P2 is concave. Since the denominator is an affine function with respect to p_{sr_i} and $p_{r_i u_k}$, the objective function of P2 is quasi-concave. The Dinkelbach's method [34] can be used to solve the fractional quasi-concave problem as a sequence of parameterized concave problems.

Algorithm 1: Optimal power allocation and user selection for maximizing effective capacity.

```
Input: $h_{sr_i}, \hat{h}_{r_i u_k} (k = 1, 2, ..., M_i), p_{\max}, A_{ik}, \theta_{ik}, M_i$
Output: $p_{sr_i}^*, p_{r_i u_k}^*, x_{ik}^*$ and $EC_{i\max}^*$
Select the $i$th cluster of the cell
for $k = 1$ to $M_i$ do
    Calculate $\phi_1 = \frac{|h_{sr_i}|^2}{N_0}$, $\phi_2 = \frac{|\hat{h}_{r_i u_k}|^2}{N_0}$, $\phi_3 = \frac{\sigma_e^2}{N_0}$
    Calculate $p_{sr_i}^*$ based on (19) and $p_{r_i u_k}^*$ based on (20)
end for
Calculate $x_{ik}^*$ based on (14) and $\sum_{k=1}^{M_i} EC_{ik}^*$ based on (10).
return $\mathbf{P}^*$, $\mathbf{X}^*$ and $EC_{i\max}^*$
```

To transform the quasi-concave fractional optimization problem P2 into a concave subtractive form, we relax the binary variable, x_{ik}, to a real number $\tilde{x}_{ik}$ within the interval $[0, 1]$. Then, P2 can be reformulated as P2$'$.

$$\mathbf{P2}' \quad \max_{p_{sr_i}, p_{r_i u_k}, \tilde{x}_{ik}} \quad \tilde{\eta}_i^{EEE}$$

$$\text{s.t.} \begin{cases} C1: \tilde{x}_{ik} \in [0, 1], & k = 1,\ 2,\ \ldots, M_i \\ C2: \sum_{k=1}^{M_i} \tilde{x}_{ik} = 1, & \forall i \\ C3: EC_{ik}(p_{sr_i}, p_{r_i u_k}, \tilde{x}_{ik}) \geq A_{ik}, & k = 1,\ 2,\ \ldots, M_i \\ C4: p_{sr_i} + p_{r_i u_k} \leq p_{\max}, & k = 1,\ 2,\ \ldots, M_i \\ C5: p_{sr_i} \geq 0, \ \ p_{r_i u_k} \geq 0, & k = 1,\ 2,\ \ldots, M_i. \end{cases}$$

We denote the set of feasible solutions of P2$'$ as $\mathbb{S}$, and the maximal value of the objective function of P2$'$ as q_i^*, which can be expressed as

$$q_i^* = \max_{\mathbb{S}}\left\{\frac{\sum_{k=1}^{M_i}\tilde{x}_{ik}^{*}EC_{ik}}{\sum_{k=1}^{M_i}\left(\varsigma_s p_{sr_i}^{*} + \tilde{x}_{ik}^{*}\varsigma_r p_{r_i u_k}^{*}\right) + (p_s^C + p_r^C)}\right\}. \tag{21}$$

Based on the solution of the fractional programming, the quasi-concave optimization problem can be converted to the following parametric concave problem:

$$F(q_i^*) = \max_{\mathbb{S}}\left(\sum_{k=1}^{M_i}\tilde{x}_{ik}^{*}EC_{ik} - q_i\left(\sum_{k=1}^{M_i}(\varsigma_s p_{sr_i}^{*} + \tilde{x}_{ik}^{*}\varsigma_r p_{r_i u_k}^{*}) + (p_s^C + p_r^C)\right)\right). \tag{22}$$

The maximal value of q_i in (21) can be derived as the root of equation $F(q_i^*) = 0$. To find the root of the equation, we propose an iterative method to find increasing q_i values by solving the parameterized problem in (22) according to the Dinkelbach method. The detailed steps are given in Algorithm 2.

Algorithm 2: Dinkelbach's method for EEE maximizing under power and delay constraints.

```
Input: $h_{sr_i}$, $\hat{h}_{r_i u_k} (k = 1, 2, ..., M_i)$, $N_0$, $\sigma_e^2$, $p_s^C$, $p_r^C$, $\varsigma_s$,
    $\varsigma_r$, $\theta_{ik}$, $M_i$ and $\epsilon_{\mathrm{out}}$
Output: $q_i^*$, $p_{sr_i}^*$, $p_{r_i u_k}^*$ and $\tilde{x}_{ik}^*$
select the $i$th cluster
$q_0 = 0,\ q_1 = 10^{-2}$ and $\epsilon_{\mathrm{out}} = 10^{-6}$
$i = 1$
while $q_i - q_{i-1} > \epsilon_{\mathrm{out}}$ do
    obtain $p_{sr_i}^*$, $p_{r_i u_k}^*$ and $\tilde{x}_{ik}$ by solving P2$'$ according to
    Algorithm 4.
    $i = i + 1$
    Determine $q_i$ based on (21)
end while
return $q_i^*$, $p_{sr_i}^*$, $p_{r_i u_k}^*$ and $\tilde{x}_{ik}$
```

For a given q_i in (22), we solve the problem by using Lagrangian function as

$$\begin{aligned} & L\left(p_{sr_i}, p_{r_i u_k}, \tilde{x}_{ik}, \lambda_k, v_k, v, w_k\right) \\ & = \sum_{k=1}^{M_i} \left(-\frac{1}{\theta_{ik}} \ln \left(E_\gamma \left((1 + \tilde{x}_{ik} \rho_{ik})^{-\beta_{ik}} \right) \right) \right) \\ & \quad - q_i \sum_{k=1}^{M_i} \left((\varsigma_s p_{sr_i} + \varsigma_r p_{r_i u_k}) + \left(p_s^C + p_r^C \right) \right) \\ & \quad + \sum_{k=1}^{M_i} \lambda_k \left(-\frac{1}{\theta_{ik}} \ln \left(E_\gamma \left((1 + \tilde{x}_{ik} \rho_{ik})^{-\beta_{ik}} \right) \right) - A_{ik} \right) \\ & \quad + \sum_{k=1}^{M_i} v_k \left(p_{\max} - p_{sr_i} - p_{r_i u_k} \right) + v p_{sr_i} \\ & \quad + \varphi_k \left(\sum_{k=1}^{M_i} \tilde{x}_{ik} - 1 \right) + \sum_{k=1}^{M_i} w_k p_{r_i u_k} \end{aligned} \tag{23}$$

where $\beta_{ik} = \frac{\theta_{ik}}{2 \ln 2}$, and λ_k, v_k, v and w_k are non-negative multipliers associated with the constraints $C3$, $C4$ and $C5$ in P2$'$, respectively. With these constraints, the dual problem can be written as

$$\min_{\lambda_k, v_k, v, w_k} \ \max_{\mathbb{S}} \ L\left(p_{sr_i}, p_{r_i u_k}, \lambda_k, v_k, v, w_k\right). \tag{24}$$

Constraints $C1$ and $C2$ in P2$'$ are temporarily omitted, which will be included when deriving the optimal solution in the subsequent section. By using Lagrange dual decomposition, the dual problem in (24) is decomposed into the inner layer problem and the outer layer problem, which can be solved by an iterative method. In the inner layer, for a fixed set of Lagrange multipliers, the Lagrangian function is maximized to derive the optimal values $p_{sr_i}^*$, $p_{r_i u_k}^*$ and $\tilde{x}_{ik}^*$ for a fixed set of Lagrange multipliers λ_k, v_k, v and w_k. In the outer layer, the dual problem uses the solution of inner layer for updating Lagrange multipliers via sub-gradient method.

A. Solving the Inner Layer Problem

The Dinkelbach method is to solve a sequence of maximization problems by means of iterations. For brevity, let $F(q_{i-1})$ denote the parametric concave problem in the ith iteration, which indicates that the problem is dependent on a given parameter q_{i-1} obtained in the $(i-1)$th iteration. Since the numerator and denominator in the objective function of P2$'$ are concave and affine with respect to p_{sr_i}, $p_{r_i u_k}$ and $\tilde{x}_{ik}$, the function $F(q_{i-1})$ is concave. Assuming the existence of an interior point (Slater's condition), there is zero duality gap between the primal problem and its dual problem. Therefore, we solve the inner layer maximization problem using the dual decomposition approach [35]. With a set of fixed values for λ_k, v_k, v, w_k and q_{i-1}, the problem

$$\max_{p_{sr_i}, p_{r_i u_k}} \ L(p_{sr_i}, p_{r_i u_k}, \lambda_k, v_k, v, w_k) \tag{25}$$

is solved to obtain the corresponding transmit power on both the BS and the relay, and then the user selection results are also obtained.

Since the problem in (25) is a standard-form concave problem, the KKT conditions are used to find the optimal solutions [32], given by

$$\begin{aligned} \frac{\partial L}{\partial p_{sr_i}} |p_{sr_i}^* = & \frac{(1 + \lambda_{2k}) (1 + \rho_{ik})^{-(\beta_{ik}+1)} f_\gamma(\gamma) {\rho'}_{ik} |p_{sr_i}^*}{E_\gamma \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2} \\ & - q \varsigma_s - v_k + v = 0 \end{aligned} \tag{26}$$

and

$$\begin{aligned} \frac{\partial L}{\partial p_{r_i u_k}} |p_{r_i u_k}^* = & \frac{(1 + \lambda_{2k}) (1 + \rho_{ik})^{-(\beta_{ik}+1)} f_\gamma(\gamma) {\rho'}_{ik} |p_{r_i u_k}^*}{E_\gamma \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2} \\ & - q \varsigma_r - v_k + w_k = 0. \end{aligned} \tag{27}$$

For a successful transmission from the BS to the kth near-cell-edge user via the ith relay, both p_{sr_i} and $p_{r_i u_k}$ have to be positive. Thus, with the following complementary slackness conditions

$$v \cdot p_{r_i u_k} = 0, \qquad w_k \cdot p_{r_i u_k} = 0 \tag{28}$$

we have $v = 0$ and $w_k = 0$. Based on (26) and (27), we obtain

$$\frac{\partial \rho_{ik}}{\partial p_{sr_i}} |p_{sr_i}^* = \frac{(q \varsigma_s + v_{2k}) E_\gamma \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2}{(1 + \lambda_{2k}) (1 + \rho_{ik})^{-(\beta_{ik}+1)} f_\gamma(\gamma)} \tag{29}$$

and

$$\frac{\partial \rho_{ik}}{\partial p_{r_i u_k}} |p_{r_i u_k}^* = \frac{(q \varsigma_r + v_{2k}) E_\gamma \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2}{(1 + \lambda_{2k}) (1 + \rho_{ik})^{-(\beta_{ik}+1)} f_\gamma(\gamma)}. \tag{30}$$

Then, (29)–(30) can be further simplified as

$$\frac{\frac{\partial \rho_{ik}}{\partial p_{sr_i}} |p_{sr_i}^*}{\frac{\partial \rho_{ik}}{\partial p_{r_i u_k}} |p_{r_i u_k}^*} = \frac{q \varsigma_s + v_k}{q \varsigma_r + v_k}. \tag{31}$$

Algorithm 3: Search for the optimal $p^*_{sr_i}$ and $p^*_{r_iu_k}$ for given q_{i-1} and v_k.

```
Input: $q_{i-1}, v_k, \phi_1, \phi_2, \phi_3, \varsigma_s, \varsigma_r, p_{\max}, A_{ik}, \theta_{ik}, M_i$
Output: $p^*_{sr_i}, p^*_{r_iu_k}$
Select the $i$th cluster and let $m = \sqrt{\frac{(q_{i-1}\varsigma_s + v_k)\phi_1}{(q_{i-1}\varsigma_r + v_k)(\phi_2+\phi_3)}}$.
$p^*_{sr_i} = \frac{1}{3}p_{\max}/(1+m), p^*_{r_iu_k} = mp^*_{sr_i}$.
$t = 1$
while $\{\frac{-1}{\theta_{ik}} \ln\left[E_\gamma(1+\rho_{ik})^{\beta_{ik}}\right] - A_{ik} < 0\}$ do
    $p^*_{sr_i} = p^*_{sr_i} + (\frac{p_{\max}}{m+1} - p^*_{sr_i})/(2\sqrt{t})$
    $p^*_{r_iu_k} = mp^*_{sr_i}$
    $t = t + 1$
end while
return $p^*_{sr_i}, p^*_{r_iu_k}$
```

From (8), the ratio of partial derivatives of ρ_{ik} with respective to p_{sr_i} and $p_{r_iu_k}$ can be derived as

$$\frac{\frac{\partial \rho_{ik}}{\partial p_{sr_i}}|p^*_{sr_i}}{\frac{\partial \rho_{ik}}{\partial p_{r_iu_k}}|p^*_{r_iu_k}} = \frac{p^2_{r_iu_k}(\phi_2+\phi_3)}{p^2_{sr_i}\phi_1}. \tag{32}$$

From (31) and (32), we obtain

$$p^*_{r_iu_k} = \sqrt{\frac{(q\varsigma_s + v_k)\,\phi_1}{(q\varsigma_r + v_k)\,(\phi_2+\phi_3)}}\,p^*_{sr_i}. \tag{33}$$

Algorithm 3 is proposed to search for the optimal values, $p^*_{sr_i}$ and $p^*_{r_iu_k}$, according to (33), (15d) and (15e).

With $p^*_{sr_i}$ and $p^*_{r_iu_k}$, we take the derivative of $L(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik}, \varphi_k, \lambda_k, v_k, v, w_k)$ in (23) with respect to $\tilde{x}_{ik}$ to determine the optimal user selection, given by

$$\begin{aligned}&\frac{\partial L\left(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik}, \varphi_k, \lambda_k, v_k, v, w_k\right)}{\partial \tilde{x}_{ik}}|_{p^*_{sr_i}, p^*_{r_iu_k}}\\ &= \frac{\rho_{ik}\left(1+\lambda_k\right)\left(1+\tilde{x}_{ik}\rho_{ik}\right)^{-(\beta_{ik}+1)}}{2\ln 2E_\gamma\left(\left(1+\tilde{x}_{ik}\rho_{ik}\right)^{-\beta_{ik}}\right)}|_{p^*_{sr_i}, p^*_{r_iu_k}, \tilde{x}_{ik}=1}\\ &\quad + \varphi_k = D_{ik}.\end{aligned} \tag{34}$$

Based on (34), the optimal user selection policy is given by

$$\tilde{x}^*_{ik} = \begin{cases} 1, & \text{if } k = \max_j \ D_{ij} \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

Equation (35) means that, given the ith relay, the jth user with the maximum marginal benefit D_{ij} among all the M_i near-cell-edge users should be scheduled to receive data from the ith relay.

B. Solving for the Outer Layer Problem

The inner layer problem in (24) is maximized with the optimal power allocation, $p^*_{sr_i}$ and $p^*_{r_iu_k}$ for a fixed set of Lagrange multipliers λ_k, v_k, v and w_k. With the derived $p^*_{sr_i}$ and $p^*_{r_iu_k}$, we can solve the outer layer problem in (24) for specified values v_k and λ_k. However, the Lagrange multipliers related to the constraints should also be optimized. Due to the difficulty of obtaining the optimal Lagrange multipliers, we use a subgradient method to obtain the optimal v^*_k and λ^*_k as

Algorithm 4: Inner layer problem solution for obtaining the optimal power allocation and user selection for a given q_i.

```
Input: $h_{sr_i}, \hat{h}_{r_iu_k} (k = 1, 2, ..., M_i), N_0, \sigma^2_e, p^C_s, p^C_r, \varsigma_s, \varsigma_r,$
       $p_{\max}, A_{ik}, \theta_{ik}, M_i, q_i$ and $\epsilon_{in}$
Output: $p^*_{sr_i}, p^*_{r_iu_k}, x^*_{ik}$
Select the $i$th cluster
Initialization : $\lambda_k(t), v_k(t)$
$t = 1$
while $\{\lambda_k(t) - \lambda_k(t-1)\} \geq \epsilon_{in}$ &&
      $\{v_k(t) - v_k(t-1)\} \geq \epsilon_{in}$ do
    for $k = 1$ to $M_i$ do
        Search $p^*_{sr_i}$ and $p^*_{r_iu_k}$ according to Algorithm 3.
    end for
    Determine $\tilde{x}^*_{ik}$ based on (40)
    Update $v_k(t+1)$ based on (41)
    Update $\lambda_k(t+1)$ based on (42)
    $t = t + 1$
end while
return $p^*_{sr_i}, p^*_{r_iu_k}, x^*_{ik}$
```

$$v_k\,(t+1) = \left[v_k\,(t) - \alpha_1\,(t)\,(p_{\max} - p_{sr_i} - p_{r_iu_k})\right]^+ \tag{36}$$

and

$$\lambda_k\,(t+1) = \left[\lambda_k\,(t) - \alpha_2\,(t)\,(EC_{ik}\,(p_{sr_i}, p_{r_iu_k}) - A_{ik})\right]^+ \tag{37}$$

where t is the iteration index, $\alpha_1(t)$ and $\alpha_2(t)$ are step sizes for the tth iteration, and $[x]^+ = \max\{0, x\}$. When the step sizes are chosen properly, the convergence to the optimal solution is guaranteed [32]. For example, the step size can be chosen equal to $\frac{1}{\sqrt{t}}$ for the tth iteration.

The procedures for solving the inner layer problem to obtain the optimal power allocation and user selection with a given q_i is presented in Algorithm 4.

V. Simulation Results

In this section, simulation results are presented to demonstrate the system performance in terms of EC, EEE, and delay bound violation probability for the proposed resource allocation schemes. There are 6 relays with fixed locations, one in each cluster, residing between the BS and the cell boundary, and 10 users are uniformly distributed in every cluster. A typical log-normal shadowing urban cellular wireless environment has variance of shadowing and path loss exponent being 8 dB and 3 respectively [26]. The channel estimation errors with variance σ^2_e are also assumed. The radius of the cell is 1 km and reference distance from the BS is 100 m. The channel bandwidth for each cluster and the carrier frequency are 0.6 MHz and 2.4 GHz, and the noise power spectral density is 10^{-17} W/Hz. Unless specified differently, power consumption on circuit blocks in all the transmitters and receivers at the BS and relay, p^C_s and p^C_r, are 0.3 W and 0.1 W, respectively. The maximum power limit on the relay link is $P_{\max} = 20$ dBm per user, and the reciprocal

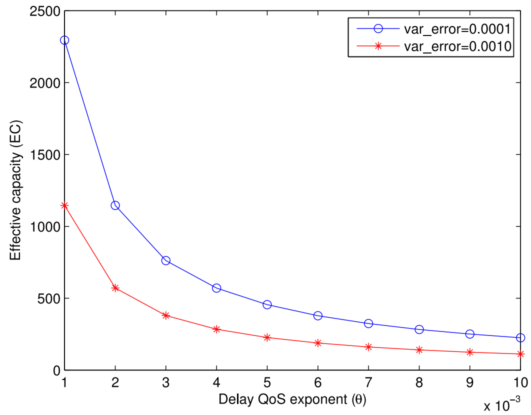


Fig. 2. EC varies with delay QoS exponent θ in P1 at $A_{ik} = 120$ kb/s.

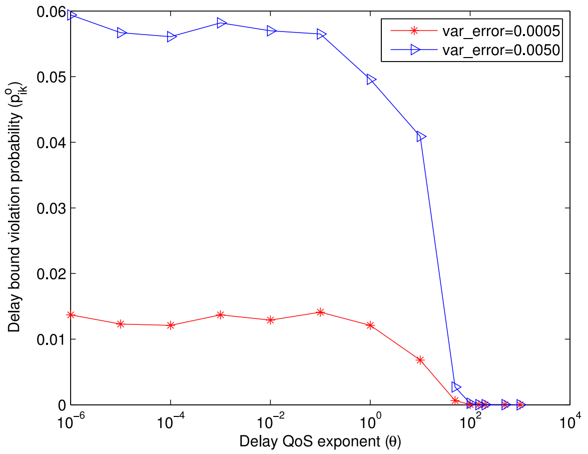


Fig. 3. Delay bound violation probability varies with delay QoS exponent θ in P1 with $D_{\max} = 50$ ms, $A_{ik} = 120$ kb/s.

of drain efficiency of power amplifiers in the BS and relays, ς_s and ς_r, are both equal to 4. The maximum delay bound $D_{ik}^{\max}$ is equal to 1 ms. All simulation results are averaged over 10000 independent channel realizations.

Fig. 2 shows the effective capacity with respect to the QoS exponent for different channel estimation errors in P1. It can be seen that EC decreases with an increase of QoS exponent θ. Moreover, for a given QoS exponent, the EC decreases with an increase of channel estimation errors. The reasons are as follows: When the QoS exponent approaches to zero, the EC approaches to the Shannon capacity, which is a theoretical upper bound. However, with an increase of the QoS exponent, for a more and more stringent delay requirement, the EC decreases gradually to meet the QoS requirement. When the QoS exponent becomes large enough, the EC decreases to a low value. For a given QoS exponent, with a decrease in channel estimation accuracy, i.e., an increase of channel estimation error variance, it is more difficult to obtain the optimal user selection and power allocation to achieve the maximum EC.

Fig. 3 shows the delay bound violation probability with respect to the QoS exponent for different channel estimation errors in P1. As can be seen in Fig. 3, when the QoS exponent is less than 0.1, indicating a loose delay requirement, the delay bound violation probability remains stable. However, when the QoS exponent becomes larger than 0.1, indicating a stringent delay requirement, the delay bound violation probability decrease sharply with the increase of QoS exponent. The reason is stated as follows: When the delay requirement is loose, the system capacity approaches to the Shannon capacity at the price of increasing delay bound violation probability. However, when the QoS exponent increases, the delay bound violation probability decreases to satisfy the more stringent QoS requirement. Moreover, with larger variance of channel estimation error, the delay bound violation probability also increases, because the power allocation on the relay link, upon the imperfect channel estimation information, cannot guarantee the selected user's EC be greater than the data arrival rate, resulting in data accumulation in the BS's buffer.

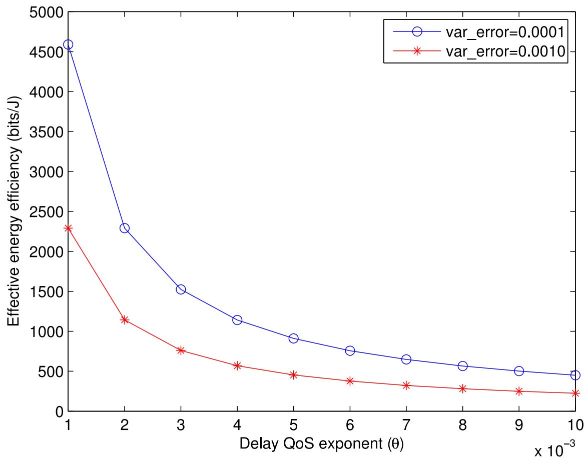


Fig. 4. EEE varies with delay QoS exponent in P1 at $A_{ik} = 120$ kb/s.

Fig. 4 shows the effective energy efficiency with respect to the QoS exponent for different channel estimation errors in P1. It can be seen that the EEE decreases with an increase of the QoS exponent. On one hand, with θ approaching to zero, the EC approaches to the Shannon capacity. On the other hand, with an increase of θ, the delay requirement is more stringent at the cost of decreased EC, resulting in the decrease of EEE. Moreover, reducing the variance of channel estimation errors can improve the system EEE. The reason for this is straightforward. A small variance of channel estimation error means more accurate channel estimations, which improves the system EEE.

Fig. 5 shows the effective energy efficiency with respect to the delay QoS exponent for different amount of circuit power consumption on the BS and relays in P1. For a given circuit power consumption, the system EEE decreases with an increase of delay QoS exponent. Moreover, for a given delay QoS exponent, the system EEE decreases with an increase of circuit power consumed. It is because the total energy consumption in each cluster increases monotonically with the power consumption at the BS and relay, therefore, EEE decreases with the increase of $p_s^C + p_r^C$. As expected, it decreases the EEE due to an increase of total power consumption.

Fig. 6 shows the system effective energy efficiency achieved when P2 is solved for the optimal resource allocation. The EEE

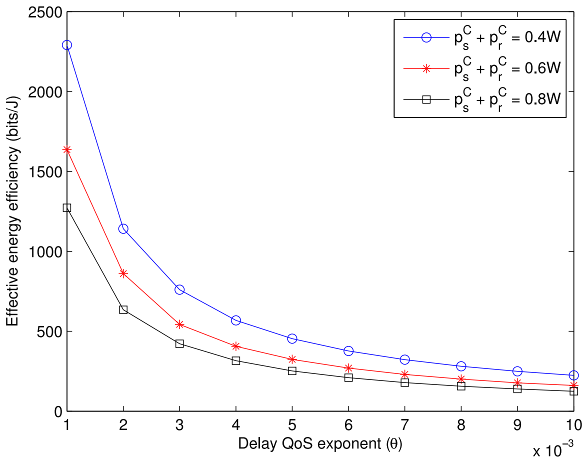


Fig. 5. EEE versus delay QoS exponent θ in P1 with $var\_error = 0.001$ with $D_{\max} = 1$ ms and $A_{ik} = 120$ kb/s.

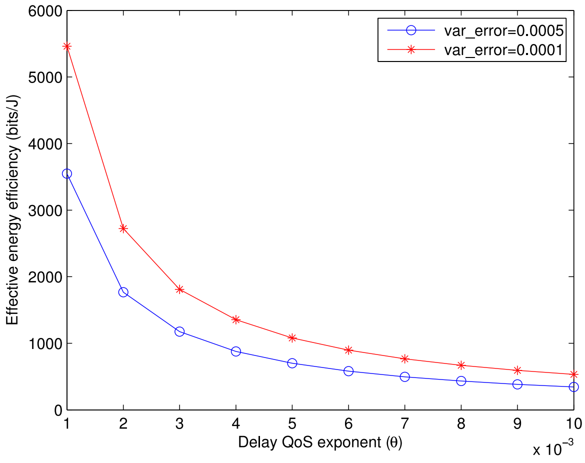


Fig. 6. EEE versus the delay QoS exponent (θ) in P2 with $D_{\max} = 3.5$ ms and $A_{ik} = 120$ kb/s.

gradually decreases with an increase of QoS exponent. Moreover, the EEE decreases with an increase of channel estimation errors. Since the QoS exponent is used to control the system's strictness for QoS requirement, a larger value of QoS exponent means a more stringent QoS requirement. Therefore, to guarantee a higher delay requirement, the EEE decreases.

Fig. 7 shows the delay bound violation probability varies with delay bound for different delay QoS exponent in P2. We can see that the delay bound violation probability decreases with an increase of delay bound for a stringent QoS requirements ($\theta = 5$, 10), which, however, stays stable for a loose QoS requirement ($\theta = 0.001$). As a matter of fact, for larger QoS exponents, an increase of user's delay bound results in a loose delay requirement and a decrease of delay bound violation probability. However, for a small QoS exponent, the system effective capacity approaches Shanon capacity, leading to a high and steady delay bound violation probability. Moreover, for a given delay bound, the delay bound violation probability decreases with an increase of QoS exponent (θ).

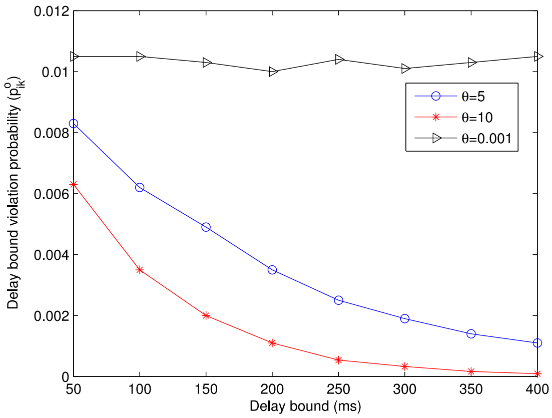


Fig. 7. Delay bound violation probability versus delay bound (ms) in P2 with $var\_error = 0.0005$ and $A_{ik} = 120$ kb/s.

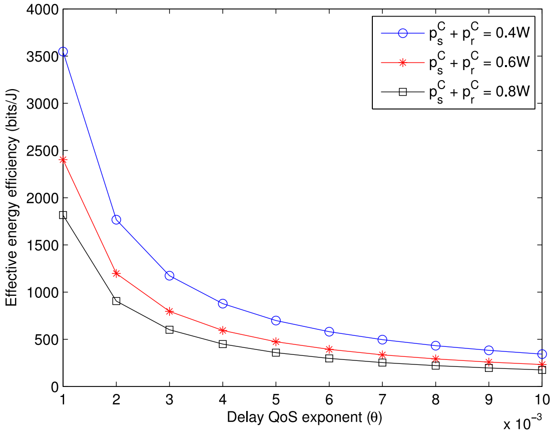


Fig. 8. EEE varies with delay QoS exponent in P2 at $D_{\max} = 3.5$ ms, $var\_error = 0.0005$ and $A_{ik} = 120$ kb/s.

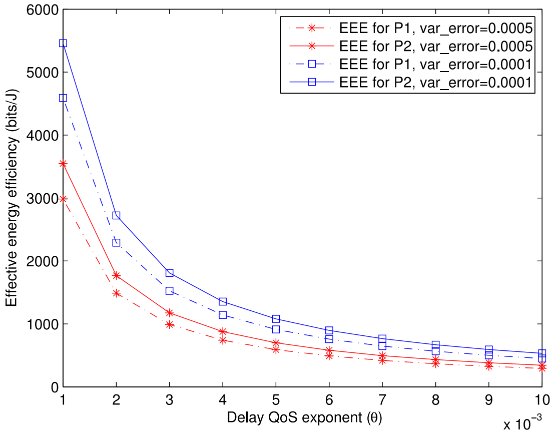


Fig. 9. Comparison of EEE between P1 and P2 with $D_{\max} = 3.5$ ms and $A_{ik} = 120$ kb/s.

Fig. 8 shows the effective energy efficiency versus the delay QoS exponent for different circuit power consumptions at the BS and relays. For a given delay QoS exponent, the EEE decreases with an increase of circuit power consumption on the BS and relays.

Fig. 9 illustrates a comparison of the effective energy efficiency achieved by maximizing the effective capacity and by maximizing the effective energy efficiency. For a given $var\_error = 0.0001$, the EEE by maximizing the EEE is nearly 1.19 times of that by maximizing the EC, when the QoS exponent θ equals 0.001. Moreover, with an increase of the QoS exponent, the difference decreases. As discussed, the EEE can be improved by reducing the channel estimation errors.

VI. Conclusion

In this paper, we investigate the downlink user selection and power allocation problems for a single cellular networks with multiple relays and multiple users by maximizing effective capacity and effective energy efficiency, respectively. For the proposed schemes, QoS requirement in terms of delay bound violation probability in effective capacity maximizing problem and effective energy efficiency maximizing problem can be both guaranteed at a given packet arrival rate and a given delay bound. Moreover, small variance of channel estimation error generates small delay bound violation probability. For a given packet arrival rate, more stringent delay QoS requirement is guaranteed with the sacrifice of more system capacity. Moreover, improving the accuracy of channel estimation helps to increase the system effective capacity and effective energy efficiency. When the circuit power consumptions on the BS and the relays are taken into account, system effective energy efficiency decreases with an increase of consumed circuit power for a given delay QoS requirement. As to the effective energy efficiency, the proposed scheme to maximize system effective energy efficiency achieves much better performance than that to maximize system effective capacity.

For future work, we will investigate the user selection and power allocation schemes for maximizing effective capacity and effective energy efficiency, respectively, when there are also direct transmission links between the BS and near-cell-edge users. Moreover, we will investigate the impact of relay's buffer size on the system effective capacity, effective energy efficiency, and delay bound violation probability.

Appendix
Proof of Convexity for P1$'$

For brevity, we let

$$f_1(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik}) = \frac{\tilde{x}_{ik} p_{sr_i} \phi_1 p_{r_iu_k} \phi_2}{p_{sr_i} \phi_1 p_{r_iu_k} \phi_3 + p_{sr_i} \phi_1 + p_{r_iu_k} (\phi_2 + \phi_3)}. \tag{38}$$

By calculating the Hessian matrix of $f_1(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik})$ with respect to the variables $p_{sr_i}, p_{r_iu_k}$ and $\tilde{x}_{ik}$, we obtain the leading principal minors as

$$D_1 = -\frac{2\tilde{x}_{ik} p^2_{r_iu_k} (p_{r_iu_k} \phi_3 + 1)(\phi_2 + \phi_3) \phi_1^2 \phi_2}{(p_{sr_i} p_{r_iu_k} \phi_1 \phi_3 + p_{sr_i} \phi_1 + p_{r_iu_k} \phi_2 + p_{r_iu_k} \phi_3)^3} \tag{39}$$

$$D_2 = \frac{4\tilde{x}^2_{ik} p^2_{sr_i} p^2_{r_iu_k} \phi_1^4 \phi_2^2 (\phi_2 + \phi_3) \phi_3}{(p_{sr_i} p_{r_iu_k} \phi_1 \phi_3 + p_{sr_i} \phi_1 + p_{r_iu_k} \phi_2 + p_{r_iu_k} \phi_3)^5} \tag{40}$$

and

$$D_3 = 0. \tag{41}$$

Since $0 < p_{sr_i} < p_{\max}$, $0 < p_{r_iu_k} < p_{\max}$, $0 < \tilde{x}_{ik} < 1$ and $\phi_i > 0\ (i = 1, 2, 3)$, we have $D_1 < 0$, $D_2 > 0$ and $D_3 \leq 0$, which means that $f_1(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik})$ is concave for all $(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik}) \in \mathbf{dom} f_1$.

Theorem 1: If $f(x, \gamma)$ is convex in x ($x \in \mathbf{R}^{\mathbf{n}}_{+}$) for each γ, where γ is a nonnegative random variable, then $E_\gamma[f(x, \gamma)]$ is also convex in x.

Proof: Since $f(x, \gamma)$ is convex in x and $E_\gamma[f(x, \gamma)] = \int_0^\infty f(x, r) p_\gamma(r) dr$ where $p_\gamma(r)$ is the probability density function for the random variable γ, $E_\gamma[f(x, \gamma)]$ can be regarded as a nonnegative weighted sum of an infinite number of convex functions, which is also convex [33].

Based on the fact that $f_1(p_{sr_i}, p_{r_iu_k}, \tilde{x}_{ik})$ is concave and Theorem 1, P1$'$ can be proved as a convex optimization problem according to the composition theorem [33].

References

[1] A Green Touch White Paper, "GreenTouch green meter research study: Reducing the net energy consumption in communications networks by up to 90% by 2020," GreenTouch, Wakefield, MA, USA, Jun. 2013.

[2] C. Han *et al.*, "Green Radio: Radio techniques to enalbe energy efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, Jun. 2011.

[3] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas, "Minimizing base station power consumption," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 297–306, Feb. 2014.

[4] S. Korotky, "Semi-empirical description and projections of internet traffic trends using a hyperbolic compound annual growth rate," *Bell Labs Tech. J.*, vol. 18, no. 3, pp. 5–21, Dec. 2013.

[5] C. X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[6] Y. Chen, X. Fang, and B. Huang, "Energy-efficient relay selection and resource allocation in nonregenerative relay OFDMA systems, *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3689–3699, Oct. 2014.

[7] G. Miao, N. Himayat, Y. Li, and D. Bormann, "Energy efficient design in wireless OFDMA," in *Proc.IEEE Int. Conf. Commun.*, May 2008, pp. 3307–3312.

[8] G. Miao, N. Himayat, G. Li, and S. Talwar, "Distributed interference-aware energy-efficient power optimization," *IEEE Trans. Wireless Commun.*, vol. 10, no. 4, pp. 1323–1333, Apr. 2011.

[9] X. Zhang, X. Tao, Y. Li, N. Ge, and J. Lu, "On relay selection and subcarrier assignment for multiuser cooperative OFDMA networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4704–4717, Nov. 2014.

[10] C. Isheden and G. P. Fettweis, "Energy-efficient link adaptation on a Rayleigh fading channel with receiver CSI," in *Proc. IEEE Int. Conf. Commun.* , Jun. 2011, pp. 1–5.

[11] K. T. K. Cheung, S. Yang, and L. Hanzo, "Achieving maximum energy-effciency in multi-relay OFDMA cellular networks: A fractional programming approach," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2746–2757, Jul. 2013.

[12] Y. Yu, A. Pang, P. Hsiuand, and Y. Fang, "Energy-adaptive downlink resource allocation in wireless cellular systems," *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1833–1846, Sep. 2015.

[13] Q. Ye, W. Zhuang, L. Li, and P. Vigneron, "Traffic load adaptive medium access control for fully-connected mobile and ad-hoc networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9358–9371, Nov. 2016.

[14] L. Musavian and Q. Ni, "Effective capacity maximization with statistical delay and effective energy efficiency requirements," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3824–3835, Jul. 2015.

[15] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[16] W. Yu, L. Musavian, and Q. Ni, "Weighted tradeoff between effective capacity and energy efficiency," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 238–243.

[17] M. Sinaie, A. Zappone, E. A. Jorswieck, and P. Azmi, "A novel power consumption model for effective energy efficiency in wireless networks," *IEEE Commun. Lett.*, vol. 5, no. 2, pp. 152–155, Dec. 2015.

[18] S. Efazati and P. Azmi, "Effective capacity maximization in multirelay networks with a novel cross-layer transmission framework and power-allocation scheme," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1691–1702, May 2014.

[19] C. Xiong, G. Y. Li, Y. Liu, Y. Chen, and S. Xu, "Energy-efficient design for downlink OFDMA with delay-sensitive traffic," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 3085–3095, Jun. 2013.

[20] A. Helmy, L. Musavian, and T. L. Ngoc, "Energy-efficient power adaptation over a frequency-selective fading channel with delay and power constraints," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4529–4541, Sep. 2013.

[21] W. Yu, L. Musavian, and Q. Ni, "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3339–3353, May 2016.

[22] Y. Li, L. Liu, H. Li, J. Zhang, and Y. Yi, "Resource allocation for delay-sensitive traffic over LTE-Advanced relay networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4291–4303, Apr. 2015.

[23] K. T. Phan, T. L. Ngoc, and L. B. Le, "Optimal resource allocation for buffer-aided relaying with statistical QoS constraint," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 959–972, Mar. 2016.

[24] K. P. Peppas, P. T. Mathiopoulos, and J. Yang, "On the effective capacity of amplify-and-forward multihop transmission over arbitrary and correlated fading channels," *IEEE Commun. Lett.*, vol. 5, no. 3, pp. 248–251, Jun. 2016.

[25] R. Devarajan, S. C. Jha, U. Phuyal, and V. K. Bhargava, "Energy-aware resource allocation for cooperative cellular network using multi-objective optimization approach," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1797–1807, May 2012.

[26] T. Rappaport, *Wireless Communications: Principles and Practice.* Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[27] F. Gao, T. Cui, and A. Nallanathan, "On channel estimation and optimal training design for amplify and forward relay networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1907–1916, May 2008.

[28] T. Yoo and A. Goldsmith, "Capacity and power allocation for fading MIMO channels with channel estimation error, " *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[29] Y. Zhao, R. Adve, and T. Lim, "Improving amplify-and-forward relay networks: Optimal power allocation versus selection," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3114–123, Aug. 2007.

[30] M. Seyfi, S. Muhaidat, and J. Liang, "Amplify-and-forward selection cooperation over Rayleigh fading channels with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 199–209, Jan. 2012.

[31] G. Farhadi and N. Beaulieu, "On the ergodic capacity of multi-hop wireless relaying systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2286–2291, May 2009.

[32] S. Cui, A. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[34] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, pp. 492–498, Mar. 1967.

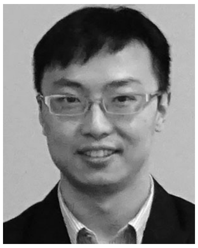
[35] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

Jiping Li received the B.Sc. degree from Hubei University, Wuhan, China, in 2001, the M.Sc. degree from the Ocean University of China, Qingdao, China, in 2006, and the Ph.D. degree from Central China Normal University, Wuhan, China, in 2012. He is currently an Associate Professor with the School of Computer and Information Science, Hubei Engineering University, Wuhan, China. His research interests include wireless resource allocation and security of wireless sensor networks.

Yaoming Ding received the B.Sc. and M.Sc. degrees in physic science from Central China Normal University, Wuhan, China, in 1986 and 2000, respectively, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2011. He is currently a Professor in physic science with Hubei Engineering University, Xiaogan, China. His research interests include optical communication, wireless resource allocation, and security of wireless sensor network.

Qiang Ye (S'16–M'17) received the B.S. degree in network engineering and the M.S. degree in communication and information system from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2016.

He has been a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, since 2016. His current research interests include resource virtualization and network slicing for 5G networks, SDN and NFV, virtual network embedding and end-to-end performance analysis, medium access control, and performance optimization in mobile ad hoc networks and Internet of Things.



Ning Zhang (M'15) received the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. He is currently an Assistant Professor with the Department of Computing Science, Texas A&M University-Corpus Christi, Corpus Christi, TX, USA. Before that, he was a Postdoctoral Research Fellow with the University of Waterloo and the University of Toronto. His current research interests include next generation wireless networks, software defined networking, vehicular networks, and physical layer security. He was the recipient of the Best Paper Award at IEEE Globecom 2014 and IEEE WCSP 2015. He is a Lead Guest Editor for the *Wireless Communications and Mobile Computing* and *International Journal of Distributed Sensor Networks*, and a Guest Editor for the *Mobile Information System*.

Weihua Zhuang (M'93–SM'01–F'08) has been, since 1993, with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks. She is the recipient of 2017 Technical Recognition Award from the IEEE Communications Society Ad Hoc and Sensor Networks Technical Committee, one of 2017 ten N2Women (Stars in Computer Networking and Communications), and the corecipient of several best paper awards from IEEE conferences. She was the Editor-in-Chief for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (2007–2013), the Technical Program Chair/Co-Chair of the IEEE VTC Fall 2017 and Fall 2016, and the Technical Program Symposia Chair of the IEEE Globecom 2011. She is a Fellow of the the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. She is an elected member in the Board of Governors and VP Publications of the IEEE Vehicular Technology Society. She was an IEEE Communications Society Distinguished Lecturer (2008–2011).