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 requirements 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].
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.
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.

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 \( h_{r,i,k} \) and \( h_{r,u,k} \), 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]

\[
h_{r,u,k} = \hat{h}_{r,u,k} + e_{r,u,k}
\]  

(1)

where \( \hat{h}_{r,u,k} \) is the channel gain based on MMSE estimation, and \( e_{r,u,k} \) is the estimation error which is uncorrelated with the estimated channel gain [28]. We assume that \( e_{r,u,k} \) behaves as additive white Gaussian noise (AWGN) with variance denoted by \( \sigma_e^2 \). 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

<table>
<thead>
<tr>
<th>Notations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h_{sr,i} )</td>
<td>channel gains from BS to the ith relay</td>
</tr>
<tr>
<td>( h_{r,u,k} )</td>
<td>channel gains form the ith relay to the kth user</td>
</tr>
<tr>
<td>( \hat{h}_{r,u,k} )</td>
<td>channel gain based on MMSE estimation</td>
</tr>
<tr>
<td>( e_{r,u,k} )</td>
<td>estimation error</td>
</tr>
<tr>
<td>( \sigma_e^2 )</td>
<td>variance of the estimation error</td>
</tr>
<tr>
<td>( p_{sr,i} )</td>
<td>transmit power at the BS</td>
</tr>
<tr>
<td>( p_{r,u,k} )</td>
<td>transmit power for the kth user at the ith relay</td>
</tr>
<tr>
<td>( n_{sr,i}, n_{r,u,k} )</td>
<td>receiving noise at the ith relay and the kth user</td>
</tr>
<tr>
<td>( N_0 )</td>
<td>variance of the random variables ( n_{sr,i}, n_{r,u,k} )</td>
</tr>
<tr>
<td>( \beta )</td>
<td>amplification factor for an AF relay</td>
</tr>
<tr>
<td>( \rho_{ik} )</td>
<td>effective receiving SNR of kth user via the ith relay</td>
</tr>
<tr>
<td>( r_{ik} )</td>
<td>Shannon capacity of the two-hop transmission link</td>
</tr>
<tr>
<td>( D_{ik}, D_{r,ik}^{max} )</td>
<td>end-to-end transmission delay, maximum delay bound</td>
</tr>
<tr>
<td>( p_{sr} )</td>
<td>delay bound violation probability</td>
</tr>
<tr>
<td>( \epsilon_{ik} )</td>
<td>probability of a nonempty queue</td>
</tr>
<tr>
<td>( \theta_{ik} )</td>
<td>QoS exponent</td>
</tr>
<tr>
<td>( A_{ik} )</td>
<td>the packet arrival rate</td>
</tr>
<tr>
<td>( EC_{ik} (\theta_{ik}) )</td>
<td>effective capacity with the QoS exponent ( \theta_{ik} )</td>
</tr>
<tr>
<td>( M_i )</td>
<td>number of users in the ith cluster</td>
</tr>
<tr>
<td>( \epsilon_{r}, \varsigma )</td>
<td>reciprocal of drain efficiency of power amplifiers</td>
</tr>
<tr>
<td>( p_C, p_R )</td>
<td>power consumption on circuit blocks at BS and relay</td>
</tr>
<tr>
<td>( P_T )</td>
<td>total power consumption for the ith cluster</td>
</tr>
<tr>
<td>( x_{ik} )</td>
<td>a binary variable indexing user selection</td>
</tr>
<tr>
<td>( \xi_{ik} )</td>
<td>a real variable within the interval ([0, 1])</td>
</tr>
<tr>
<td>( \eta_{ik} )</td>
<td>effective energy efficiency in the ith cluster</td>
</tr>
<tr>
<td>( p_{max} )</td>
<td>allowable source and relay transmit power limit</td>
</tr>
</tbody>
</table>

![Fig. 1. A relay-assisted downlink cellular network.](image-url)
given, respectively, by [29], [30]

\[ y_{sr_i} = \sqrt{P_{sr_i}} h_{sr_i} s + n_{sr_i}, \]

and

\[ y_{r_i u_k} = \beta y_{sr_i} h_{r_i u_k} + n_{r_i u_k} = \beta(\sqrt{P_{sr_i}} h_{sr_i} s + n_{sr_i}) h_{r_i u_k} + n_{r_i u_k}, \]

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

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

\[ y_{r_i u_k} = \beta \sqrt{P_{sr_i}} h_{sr_i} h_{r_i u_k} s + \beta \sqrt{P_{sr_i}} h_{sr_i} c_{r_i u_k} s + \beta h_{r_i u_k} n_{sr_i} + \beta c_{r_i u_k} n_{sr_i} + n_{r_i u_k}. \]

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 \( h_{r_i u_k} \), \( c_{r_i u_k} \), \( n_{sr_i} \), and \( n_{r_i u_k} \), the effective end-to-end receiving SNR for the two-hop transmission link from the BS to the \( k \)th user via the \( i \)th relay is given by

\[ \rho_{ik} = \frac{p_{sr_i} \phi_1 p_{r_i u_k} \phi_2}{1 + p_{sr_i} \phi_1 p_{r_i u_k} \phi_2 + p_{sr_i} \phi_1 + p_{r_i u_k} (\phi_2 + \phi_3)} \]

where \( \phi_1 = \frac{|h_{sr_i}|^2}{N_0}, \phi_2 = \frac{|h_{sr_i} c_{r_i u_k}|^2}{N_0} \) 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

\[ \rho_{ik} = \frac{p_{sr_i} \phi_1 p_{r_i u_k} \phi_2}{1 + p_{sr_i} \phi_1 + p_{r_i u_k} \phi_2}. \]

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

\[ \rho_{ik} \approx \frac{p_{sr_i} \phi_1 p_{r_i u_k} \phi_2}{p_{sr_i} \phi_1 p_{r_i u_k} \phi_3 + p_{sr_i} \phi_1 + p_{r_i u_k} (\phi_2 + \phi_3)}. \]

Therefore, the Shannon capacity of the two-hop transmission link from the BS to the \( k \)th user via the \( i \)th relay can be express as [31]

\[ r_{ik} = \frac{1}{2} \log_2(1 + \rho_{ik}), \]

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 \( k \)th user in the \( i \)th 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 \( D_{ik}^{\text{max}} \), can be computed as [16]

\[ p_{ik}^0 = \Pr \left( D_{ik} \geq D_{ik}^{\text{max}} \right) \approx e^{-\theta_{ik} A_{ik} D_{ik}^{\text{max}}} \]

where \( e^{-\theta_{ik}} \approx \lim_{p_{sr_i} \rightarrow 0} A_{ik} \) 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. \( \theta_{ik} \) 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 \( k \)th user in the \( i \)th cluster, and \( EC_{ik}(\theta_{ik}) \) is the effective capacity of the wireless channels from the BS to the \( k \)th user via the \( i \)th relay with the QoS exponent \( \theta_{ik} \). Equation (9) shows that if a source requires a delay-bound violation probability of at most \( p_{ik}^0 \), 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 \( k \)th user via the \( i \)th relay can be expressed as [16]

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

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 \( \theta_{ik} \) 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 \( i \)th cluster is expressed as

\[ P_i^t = M_i \left[ (s_k p_{sr_i} + c_k p_{r_i u_k}) + (p_s^C + p_r^C) \right] \]

where \( M_i \) is the number of users in the \( i \)th cluster, \( s_k \) and \( c_k \) 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 \( i \)th cluster, \( \eta_{EE_i} \), 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
where \( x_{ik} \) is a binary variable for user selection which equals to 1 if the \( k \)th 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 \( i \)th virtual cluster as \( P_1 \).

\[
P_1 : \quad \max_{p_{sr}, p_{ru}, x_{ik}} \sum_{k=1}^{M_i} EC_{ik} \]

\[
\begin{align*}
C1 : & \quad x_{ik} \in \{0, 1\}, \quad k = 1, 2, \ldots, M_i \\
C2 : & \quad \sum_{k=1}^{M_i} x_{ik} = 1, \quad \forall i \\
s.t. & \quad C3 : \quad EC_{ik} (p_{sr}, p_{ru}, x_{ik}) \geq A_{ik}, \quad k = 1, 2, \ldots, M_i \\
C4 : & \quad p_{sr}, + p_{ru} \leq p_{\text{max}}, \quad k = 1, 2, \ldots, M_i \\
C5 : & \quad p_{sr}, \geq 0, \quad p_{ru} \geq 0, \quad k = 1, 2, \ldots, M_i.
\end{align*}
\]

In \( P_1 \), 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 \( p_{ik}^0 \); \( C4 \) is to limit the transmit power consumption on both the BS and the relay, where \( p_{\text{max}} \) denotes allowable source and relay transmit power limit.

2) **Effective Energy Efficiency Maximization:** In \( P_1 \), 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 \( P_2 \), where the set of decision variables and constraints are the same as those in \( P_1 \).

\[
P_2 : \quad \max_{p_{sr}, p_{ru}, x_{ik}} \eta_{i}^{\text{EEE}} \]

\[
\begin{align*}
C1 : & \quad x_{ik} \in \{0, 1\}, \quad k = 1, 2, \ldots, M_i \\
C2 : & \quad \sum_{k=1}^{M_i} x_{ik} = 1, \quad \forall i \\
s.t. & \quad C3 : \quad EC_{ik} (p_{sr}, p_{ru}, x_{ik}) \geq A_{ik}, \quad k = 1, 2, \ldots, M_i \\
C4 : & \quad p_{sr}, + p_{ru} \leq p_{\text{max}}, \quad k = 1, 2, \ldots, M_i \\
C5 : & \quad p_{sr}, \geq 0, \quad p_{ru} \geq 0, \quad k = 1, 2, \ldots, M_i.
\end{align*}
\]

### III. Optimal Power Allocation and User Selection for Maximizing EC

In this section, we present how to solve \( P_1 \). Since \( P_1 \) 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 \( P_1 \) as a real number \( \tilde{x}_{ik} \) within the interval \([0, 1]\). Then, we formulate a transformed problem of \( P_1 \) as \( P_1' \).

\[
P_1' : \quad \min_{p_{sr}, p_{ru}, \tilde{x}_{ik}} \sum_{k=1}^{M_i} -EC_{ik} \]

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

Since \( P_1' \) can be proved as a convex optimization problem with respect to \( p_{sr}, p_{ru} \) 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 \( k \)th user is selected as the destination for the \( i \)th relay. Thus, the Lagrangian function for \( P_1' \) can be expressed as

\[
L (p_{sr}, p_{ru}, A_{ik}, \lambda_{1k}, \lambda_{2k}, \lambda_{3}, \lambda_{4k}) = -\sum_{k=1}^{M_i} EC_{ik} (p_{sr}, p_{ru}) + \lambda_{1k} (A_{ik} - EC_{ik} (p_{sr}, p_{ru})) + \sum_{k=1}^{M_i} \lambda_{2k} (p_{sr}, + p_{ru} - p_{\text{max}}) - \lambda_{3} p_{sr}, - \sum_{k=1}^{M_i} \lambda_{4k} p_{ru}
\]

where \( \lambda_{1k}, \lambda_{2k}, \lambda_{3}, \lambda_{4k} \) are Lagrangian multipliers. The optimal transmit power at the BS and the \( i \)th relay can be achieved by differentiating \( L(p_{sr}, p_{ru}, A_{ik}, \lambda_{1k}, \lambda_{2k}, \lambda_{3}, \lambda_{4k}) \) with respect to \( p_{sr}, \) and \( p_{ru} \). In the second step, based on the obtained optimal power allocation \( p_{sr}^* \) and \( p_{ru}^* \), the optimal user selection can be determined as

\[
x_{ik} = \begin{cases} 1, & \text{if } k = \text{arg} \max_{m} EC_{im} (p_{sr}, p_{ru}) \\ 0, & \text{otherwise.} \end{cases}
\]

The KKT conditions for \( P_1' \) are given by

\[
\begin{align}
\lambda_{1k} (A_{ik} - EC_{ik} (p_{sr}^*, p_{ru}^*)) &= 0 \quad (15a) \\
\lambda_{2k} (p_{sr}^* + p_{ru}^* - p_{\text{max}}) &= 0 \quad (15b) \\
\lambda_{3} p_{sr}^* &= 0, \quad \lambda_{4k} p_{ru}^* &= 0 \quad (15c) \\
A_{ik} &\leq EC_{ik} (p_{sr}^*, p_{ru}^*) \quad (15d) \\
p_{sr}^* + p_{ru}^* &\leq p_{\text{max}} \quad (15e) \\
p_{sr}^* \geq 0, \quad p_{ru}^* &\geq 0 \quad (15f) \\
\lambda_{1k} &\geq 0, \quad \lambda_{2k} \geq 0, \quad \lambda_{3} \geq 0, \quad \lambda_{4k} \geq 0 \quad (15g)
\end{align}
\]
\[
\frac{\partial L}{\partial p_{s,i}} = - \left( \frac{E_{\gamma} \left( e^{-\frac{1}{2} \theta_{i,k} \log (1 + p_{i,k}^*) \theta_{i,k} p_{i,k}^* \left| r_{s,i,k} \right|} \right)}{\theta_{i,k} E_{\gamma} \left( e^{-\frac{1}{2} \theta_{i,k} \log (1 + p_{i,k}^*)} \right)} + \lambda_{2k}^* - \lambda_{3k}^* = 0 \right)
\]

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 \(p_{s,r}\) and \(p_{r,u}\) should be both positive. Therefore, we have \(\lambda_{3k}^* = \lambda_{4k}^* = 0\). From (16), (17) and (15b), we have \(\lambda_{2k}^* > 0\), and \(p_{s,r}^* + p_{r,u,k}^* = p_{\text{max}}\). Then, we derive \(p_{s,r}^*\) as
\[
p_{s,r}^* = \sqrt{\frac{\phi_1}{\phi_2 + \phi_3}} p_{r,u,k}^*.
\]
Finally, we obtain the optimal solutions
\[
p_{s,r}^* = \left( \phi_2 + \phi_3 \right) - \sqrt{\phi_1 \left( \phi_2 + \phi_3 \right)} p_{\text{max}}\] (19)
and
\[
p_{r,u,k}^* = \sqrt{\phi_1 \left( \phi_2 + \phi_3 \right)} - \phi_1 p_{\text{max}}\] (20)

Detailed steps in finding optimal transmission power \(p_{s,r}^*\) and \(p_{r,u,k}^*\) at the BS and the \(i\)th relay, respectively, are given in Algorithm 1.

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

**Input:** \(h_{s,r}, \tilde{h}_{r,u,k} (k = 1, 2, \ldots, M_i), p_{\text{max}}, A_{ik}, \theta_{ik}, M_i\)

**Output:** \(p_{s,r}^*, p_{r,u,k}^*, x_{ik}^*\) and \(EC_{i,k}^*\)

1. Select the \(i\)th cluster of the cell
2. for \(k = 1\) to \(M_i\) do
3. Calculate \(\phi_1 = |h_{s,r}|^2 / N_0\), \(\phi_2 = |\tilde{h}_{r,u,k}|^2\), \(\phi_3 = \sigma^2 / N_0\)
4. Calculate \(p_{s,r}^*\) based on (19) and \(p_{r,u,k}^*\) based on (20)
5. end for
6. Calculate \(x_{ik}^*\) based on (14) and \(\sum_{k=1}^{M_i} EC_{i,k}^*\) based on (10).
7. return \(P^*, X^*\) and \(EC_{i}^{*}\)

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′.

\[
P^{*2} = \max_{p_{s,r}, p_{r,u,k}, \tilde{x}_{ik}} \tilde{h}_{i}^{\text{EEE}}
\]

\[
\begin{aligned}
C_1 : & \tilde{x}_{ik} \in [0, 1], \quad k = 1, 2, \ldots, M_i \\
C_2 : & \sum_{k=1}^{M_i} \tilde{x}_{ik} \geq 1, \quad \forall i \\
C_3 : & EC_{i,k} \left( p_{s,r}, p_{r,u,k}, x_{ik} \right) \geq A_{ik}, \quad k = 1, 2, \ldots, M_i \\
C_4 : & p_{s,r} + p_{r,u,k} \leq p_{\text{max}}, \quad k = 1, 2, \ldots, M_i \\
C_5 : & p_{s,r} \geq 0, \quad p_{r,u,k} \geq 0, \quad k = 1, 2, \ldots, M_i.
\end{aligned}
\]

We denote the set of feasible solutions of P2′ as \(S\), and the maximal value of the objective function of P2′ as \(q_i^*\), which can be expressed as
\[
q_i^* = \max_{S} \left\{ \sum_{k=1}^{M_i} \tilde{x}_{ik} \left( EC_{i,k} - q_i^* \left( \sum_{k=1}^{M_i} \left( s, p_{s,r}^* \right) \right) \right) \right\}.
\]

Based on the solution of the fractional programming, the quasi-concave optimization problem can be converted to the following parametric concave problem:
\[
F(q_i^*) = \max_{S} \left( \sum_{k=1}^{M_i} \tilde{x}_{ik} EC_{i,k} - q_i^* \left( \sum_{k=1}^{M_i} \left( s, p_{s,r}^* \right) \right) \right).
\]

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.
For a given $q_i$ in (22), we solve the problem by using Lagrangian function as
\[
L(p_{sr}, p_{r}, u_k, x_{ik}, \lambda_k, v_k, v, w_k) = \sum_{k=1}^{M_i} \left( -\frac{1}{\theta_{ik}} \ln \left( E_{\gamma} \left( (1 + x_{ik} \rho_{ik})^{-\beta_{ik}} \right) \right) \right) - q_i \left( (s_k p_{sr} + \zeta r_{r,u_k} + (p_{sr}^C + p_{r}^C)) \right) + \sum_{k=1}^{M_i} \lambda_k \left( -\frac{1}{\theta_{ik}} \ln \left( E_{\gamma} \left( (1 + x_{ik} \rho_{ik})^{-\beta_{ik}} \right) \right) \right) - A_{ik} + \sum_{k=1}^{M_i} u_k \left( p_{max} - p_{sr} - p_{r,u_k} \right) + \phi p_{sr} + \phi^2 \left( \sum_{k=1}^{M_i} x_{ik} - 1 \right) + \sum_{k=1}^{M_i} w_k p_{r,u_k} \right)
\]
where $\beta_{ik} = \frac{\theta_{ik}}{2 \ln 4}$, and $\lambda_k$, $v_k$, $v$ 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
\[
\min_{\lambda_k, v_k, v, w_k} \max_{S} L(p_{sr}, p_{r,u_k}, \lambda_k, v_k, v, w_k). \tag{24}
\]
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 $p^*_{sr}, p^*_{r,u_k}$ and $x^*_{ik}$ for a fixed set of Lagrange multipliers $\lambda_k$, $v_k$, $v$ 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 $i$th 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 $p_{sr}, p_{r,u_k}$ and $x_{ik}$, 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 $\lambda_k, v_k, v$ and $q_{i-1}$, the problem
\[
\max_{p_{sr}, p_{r,u_k}} L(p_{sr}, p_{r,u_k}, \lambda_k, v_k, v, w_k) \tag{25}
\]
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
\[
\frac{\partial L}{\partial p_{sr}} \bigg|_{p^*_{sr}} = \frac{(1 + \lambda_{2k})(1 + \rho_{ik})^{-\beta_{ik}} + f_1(\gamma)\rho'_{ik}|p^*_{sr}}{E_{\gamma} \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2} - q\varsigma_k - q_k + v = 0 \tag{26}
\]
and
\[
\frac{\partial L}{\partial p_{r,u_k}} \bigg|_{p^*_{r,u_k}} = \frac{(1 + \lambda_{2k})(1 + \rho_{ik})^{-\beta_{ik}} + f_1(\gamma)\rho'_{ik}|p^*_{r,u_k}}{E_{\gamma} \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2} - q\varsigma_k - q_k + w_k = 0. \tag{27}
\]
For a successful transmission from the BS to the $k$th near-cell-edge user via the $i$th relay, both $p_{sr}$ and $p_{r,u_k}$ have to be positive. Thus, with the following complementary slackness conditions
\[
v \cdot p_{r,u_k} = 0, \quad w_k \cdot p_{r,u_k} = 0 \tag{28}
\]
we have $v = 0$ and $w_k = 0$. Based on (26) and (27), we obtain
\[
\frac{\partial p_{sr}}{\partial p_{sr}} \bigg|_{p^*_{sr}} = \frac{(q\varsigma_k + v_{2k})E_{\gamma} \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2}{(1 + \lambda_{2k})(1 + \rho_{ik})^{-\beta_{ik}} + f_1(\gamma)} \tag{29}
\]
and
\[
\frac{\partial p_{r,u_k}}{\partial p_{r,u_k}} \bigg|_{p^*_{r,u_k}} = \frac{(q\varsigma_k + v_{2k})E_{\gamma} \left[(1 + \rho_{ik})^{-\beta_{ik}} \right] 2 \ln 2}{(1 + \lambda_{2k})(1 + \rho_{ik})^{-\beta_{ik}} + f_1(\gamma)} \tag{30}
\]
Then, (29)–(30) can be further simplified as
\[
\frac{\partial p_{sr}}{\partial p_{sr}} \bigg|_{p^*_{sr}} = \frac{q\varsigma_k + v_k}{q\varsigma_k + v_k}. \tag{31}
\]
Algorithm 3: Search for the optimal $p^*_{sr,i}$ and $p^*_{r,u,k}$ for given $q_{i-1}$ and $v_k$.

**Input:** $q_{i-1}, v_k, \phi_1, \phi_2, \phi_3, \varsigma_s, s_r, p_{max}, A_{ik}, \theta_{ik}, M_i$  
**Output:** $p^*_{sr,i}, p^*_{r,u,k}$

1: Select the $i$th cluster and let $m = \sqrt{\frac{(q_{i-1}, v_k + v_k)}{(q_{i-1}, v_k + v_k) + \phi_2 \phi_3}}$.
2: $p^*_{sr,i} = \frac{1}{m} p_{max}/(1 + m)$, $p^*_{r,u,k} = mp^*_{sr,i}$.
3: $t = 1$
4: while $\{\frac{1}{g_{ik}} \ln [E_i(1 + p_{ik})^{\lambda_k}] - A_{ik} < 0\}$ do
5: $p^*_{sr,i} = p^*_{sr,i} + (\frac{q_{sr,i}}{m + 1} - p^*_{sr,i})/(2\sqrt{t})$
6: $p^*_{r,u,k} = mp^*_{sr,i}$
7: $t = t + 1$
8: end while
9: return $p^*_{sr,i}, p^*_{r,u,k}$

From (8), the ratio of partial derivatives of $\rho_{ik}$ with respect to $p_{sr,i}$, and $p_{r,u,k}$ can be derived as

$$\frac{\partial \rho_{ik}}{\partial p_{sr,i}}|_{p^*_{r,u,k}} = p^2_{sr,i} (\phi_2 + \phi_3).$$

From (31) and (32), we obtain

$$p^*_{r,u,k} = \sqrt{\frac{(q_{sr,i} + v_k) \phi_1}{(q_{sr,i} + v_k) \phi_2 + \phi_3}} p^*_{sr,i}.$$  

Algorithm 3 is proposed to search for the optimal values, $p^*_{sr,i}$, and $p^*_{r,u,k}$, according to (33), (15d) and (15e).

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

$$\frac{\partial L(p_{sr,i}, p_{r,u,k}, \tilde{x}_{ik}, \varphi_k, \lambda_k, v_k, v, w_k)}{\partial \tilde{x}_{ik}}|_{p^*_{sr,i}, p^*_{r,u,k}} = \rho_k (1 + \lambda_k)(\frac{1}{\lambda_k(1 + \varphi_k)\rho_k})^{\beta_k}.\left(\frac{\beta_k + 1}{2\ln 2} + \frac{\varphi_k}{D_{ik}}\right).$$

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

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

Equation (35) means that, given the $i$th relay, the $j$th 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 $i$th 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 $p^*_{r,u,k}$ for a fixed set of Lagrange multipliers $\lambda_k, v_k, v, w_k$. With the derived $p^*_{sr,i}$ and $p^*_{r,u,k}$, we can solve the outer layer problem in (24) for specified values $v_k$ and $\lambda_k$. 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 sub-gradient method to obtain the optimal $v_k$ and $\lambda_k$ as

$$v_k(t + 1) = [v_k(t) - \alpha_1(t) (p_{max} - p_{sr,i} - p_{r,u,k})]^{+}$$

and

$$\lambda (t + 1) = [\lambda_k(t) - \alpha_2(t) (EC_{ik} (p_{sr,i}, p_{r,u,k}) - A_{ik})]^{+}$$

where $t$ is the iteration index, $\alpha_1(t)$ and $\alpha_2(t)$ are step sizes for the $t$th 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 $t$th 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.

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

**Input:** $q_{i-1}, v_k, \phi_1, \phi_2, \phi_3, \varsigma_s, s_r, p_{max}, A_{ik}, \theta_{ik}, M_i, q_i$ and $\epsilon_{in}$

**Output:** $p^*_{sr,i}, p^*_{r,u,k}, \tilde{x}_{ik}^*$

1: Select the $i$th cluster
2: Initialization: $\lambda_k(t), v_k(t)$
3: $t = 1$
4: while $\{\lambda_k(t) - \lambda_k(t - 1) || \geq \epsilon_{in} \wedge \{v_k(t) - v_k(t - 1) || \geq \epsilon_{in} \}$ do
5: for $k = 1$ to $M_i$ do
6: Search $p^*_{sr,i}$ and $p^*_{r,u,k}$ according to Algorithm 3.
7: end for
8: Determine $\tilde{x}_{ik}^*$ based on (40)
9: Update $v_k(t + 1)$ based on (41)
10: Update $\lambda_k(t + 1)$ based on (42)
11: $t = t + 1$
12: end while
13: return $p^*_{sr,i}, p^*_{r,u,k}, \tilde{x}_{ik}^*$

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 $\sigma^2_z$ 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, $p^C_r$ 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...
of drain efficiency of power amplifiers in the BS and relays, $\varsigma_s$ and $\varsigma_r$, are both equal to 4. The maximum delay bound $D_{\text{max}}$ 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 $\theta$. 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 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 consumption 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 $p_C^s + p_C^r$. As expected, it decreases the EEE due to an increase of total power consumption.

Fig. 4 shows the effective energy efficiency achieved when P2 is solved for the optimal resource allocation. The EEE 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.
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 ($\theta$).
by maximizing the effective energy efficiency. For a given $\text{var}_{\text{error}} = 0.0001$, the EEE by maximizing the EEE is nearly 1.19 times of that by maximizing the EC, when the QoS exponent $\theta$ 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

$$f_1(p_{sr}, p_{r, u}, \tilde{x}_{ik}) = \frac{\tilde{x}_{ik} p_{sr} \phi_1 p_{r, u} \phi_2}{p_{sr} \phi_1 p_{r, u} \phi_3 + p_{sr} \phi_1 + p_{r, u} (\phi_2 + \phi_3)}.$$  \hfill (38)

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

$$D_1 = -\frac{2\tilde{x}_{ik}^2 p_{r, u}^2 (p_{sr} \phi_3 + 1) (\phi_2 + \phi_3) \phi_1 \phi_2}{(p_{sr} \phi_1 p_{r, u} \phi_3 + p_{sr} \phi_1 + p_{r, u} \phi_2 + p_{r, u} \phi_3)^2}$$  \hfill (39)

and

$$D_3 = 0.$$  \hfill (41)

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

Theorem 1: If $f(x, \gamma)$ is convex in $x (x \in \mathbb{R}^n)$ for each $\gamma$, where $\gamma$ 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 $E_{\gamma} [f(x, \gamma)] = \int_0^\infty f(x, r) p_r(r) dr$ where $p_r(r)$ is the probability density function for the random variable $\gamma$, $E_{\gamma} [f(x, \gamma)]$ 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 $f_1(p_{sr}, p_{r, u}, \tilde{x}_{ik})$ is concave and Theorem 1, P1’ can be proved as a convex optimization problem according to the composition theorem [33].

REFERENCES


